Assessing the quality of research
Paul Glasziou, Jan Vandenbroucke, Iain Chalmers

Inflexible use of evidence hierarchies confuses practitioners and irritates researchers. So how can we improve the way we assess research?

The widespread use of hierarchies of evidence that grade research studies according to their quality has helped to raise awareness that some forms of evidence are more trustworthy than others. This is clearly desirable. However, the simplifications involved in creating and applying hierarchies have also led to misconceptions and abuses. In particular, criteria designed to guide inferences about the main effects of treatment have been uncritically applied to questions about aetiology, diagnosis, prognosis, or adverse effects. So should we assess evidence the way Michelin guides assess hotels and restaurants? We believe five issues should be considered in any revision or alternative approach to helping practitioners to find reliable answers to important clinical questions.

Different types of question require different types of evidence

Ever since two American social scientists introduced the concept in the early 1960s, hierarchies have been used almost exclusively to determine the effects of interventions. This initial focus was appropriate but has also engendered confusion. Although interventions are central to clinical decision making, practice relies on answers to a wide variety of types of clinical questions, not just the effect of interventions. Other hierarchies might be necessary to answer questions about aetiology, diagnosis, disease frequency, prognosis, and adverse effects. Thus, although a systematic review of randomised trials would be appropriate for answering questions about the main effects of a treatment, it would be ludicrous to attempt to use it to ascertain the relative accuracy of computerised versus human reading of cervical smears, the natural course of prion diseases in humans, the effect of carriage of a mutation on the risk of venous thrombosis, or the rate of vaginal adenocarcinoma in the daughters of pregnant women given diethylstilboesterol.

To answer their everyday questions, practitioners need to understand the “indications and contraindications” for different types of research evidence. Randomised trials can give good estimates of treatment effects but poor estimates of overall prognosis; comprehensive non-randomised inception cohort studies with prolonged follow up, however, might provide the reverse.

Systematic reviews of research are always preferred

With rare exceptions, no study, whatever the type, should be interpreted in isolation. Systematic reviews are required of the best available type of study for answering the clinical question posed. A systematic review does not necessarily involve quantitative pooling in a meta-analysis.

Although case reports are a less than perfect source of evidence, they are important in alerting us to potential rare harms or benefits of an effective treatment. Standardised reporting is certainly needed, but too few people know about a study showing that more than half of suspected adverse drug reactions were confirmed by subsequent, more detailed research. For reliable evidence on rare harms, therefore, we need a systematic review of case reports rather than a haphazard selection of them. Qualitative studies can also be incorporated in reviews—for example, the systematic compilation of the reasons for non-compliance with hip protectors derived from qualitative research.
Level alone should not be used to grade evidence

The first substantial use of a hierarchy of evidence to grade health research was by the Canadian Task Force on the Preventive Health Examination. 12 Although such systems are preferable to ignoring research evidence or failing to provide justification for selecting particular research reports to support recommendations, they have three big disadvantages. Firstly, the definitions of the levels vary within hierarchies so that level 2 will mean different things to different readers. Secondly, novel or hybrid research designs are not accommodated in these hierarchies—for example, reanalysis of individual data from several studies or case crossover studies within cohorts. Thirdly, and perhaps most importantly, hierarchies can lead to anomalous rankings. For example, a statement about one intervention may be graded level 1 on the basis of a systematic review of a few, small, poor quality randomised trials, whereas a statement about an alternative intervention may be graded level 2 on the basis of one large, well conducted, multicentre, randomised trial.

This ranking problem arises because of the objective of collapsing the multiple dimensions of quality (design, conduct, size, relevance, etc) into a single grade. For example, randomisation is a key methodological feature in research into interventions, 13 but reducing the quality of evidence to a single level reflecting proper randomisation ignores other important dimensions of randomised clinical trials. These might include:

- Other design elements, such as the validity of measurements and blinding of outcome assessments
- Quality of the conduct of the study, such as loss to follow up and success of blinding
- Absolute and relative size of any effects seen
- Confidence intervals around the point estimates of effects.

None of the current hierarchies of evidence includes all these dimensions, and recent methodological research suggests that it may be difficult for them to do so. 14 Moreover, some dimensions are more important for some clinical problems and outcomes than for others, which necessitates a tailored approach to appraising evidence. 15 Thus, for important recommendations, it may be preferable to present a brief summary of the central evidence (such as “double-blind randomised controlled trials with a high degree of follow up over three years showed that . . .”), coupled with a brief appraisal of why particular quality dimensions are important. This broader approach to the assessment of evidence applies not only to randomised trials but also to observational studies. In the final recommendations, there will also be a role for other types of scientific evidence—for example, on aetiological and pathophysiological mechanisms—because concordance between theoretical models and the results of empirical investigations will increase confidence in the causal inferences. 16 17

What to do when systematic reviews are not available

Although hierarchies can be misleading as a grading system, they can help practitioners find the best relevant evidence among a plethora of studies of diverse quality. For example, to answer a therapeutic question, the hierarchy would suggest first looking for a systematic review of randomised controlled trials. However, only a fraction of the hundreds of thousands of reports of randomised trials have been considered for possible inclusion in systematic reviews. 18 So when there is no existing review, a busy clinician might next try to identify the best of several randomised trials. If the search fails to identify any randomised trials, non-randomised cohort studies might be informative. For non-therapeutic questions, however, search strategies should accommodate the need for observational designs that answer questions about aetiology, prognosis, or adverse effects. 19 Whatever evidence is found, this should be clearly described rather than simply assigned to a level. Such considerations have led the authors of the BMJ’s Clinical Evidence to use a hierarchy for finding evidence but to forgo grading evidence into levels. Instead, they make explicit the type of evidence on which their conclusions are based.

Balanced assessments should draw on a variety of types of research

For interventions, the best available evidence for each outcome of potential importance to patients is needed. 20 Often this will require systematic reviews of several different types of study. As an example, consider a woman interested in oral contraceptives. Evidence is available from controlled trials showing their contraceptive effectiveness. Although contraception is the main intended beneficial effect, some women will also be interested in the effects of oral contraceptives on acne or dysmenorrhoea. These may have been assessed in short term randomised controlled trials comparing different contraceptives. Any beneficial intended effect needs to be weighed against possible harms, such as increases in thromboembolism and breast cancer. The best evidence for such potential harms is likely to come from non-randomised cohort studies or case-control studies. For example, fears about negative consequences on fertility after long term use of oral contraceptives were allayed by such non-randomised studies. The figure gives an example of how all this information might be amalgamated into a balance sheet. 21 22

Sometimes, rare, dramatic adverse effects detected with case reports or case control studies prompt further investigation and follow up of existing randomised cohorts to detect related but less severe adverse effects. For example, the case reports and case-control studies showing that intrauterine exposure to diethylstilboestrol could cause vaginal adenocarcinoma led to further investigation and follow up of the mothers and children (male as well as female) who had participated in the relevant randomised trials. These investigations showed several less serious but more frequent adverse effects of diethylstilboestrol that would have otherwise been difficult to detect. 9

Conclusions

Given the flaws in evidence hierarchies that we have described, how should we proceed? We suggest that there are two broad options: firstly, to extend, improve, and standardise current evidence hierarchies 23; and, secondly, to abolish the notion of evidence hierarchies and levels of evidence, and concentrate instead on teaching practitioners general principles of research so
that they can use these principles to appraise the quality and relevance of particular studies.\text{\textsuperscript{5}}

We have been unable to reach a consensus on which of these approaches is likely to serve the current needs of practitioners more effectively. Practitioners who seek immediate answers cannot embark on a systematic review every time a new question arises in their practice. Clinical guidelines are increasingly prepared professionally—for example, by organisations of general practitioners and of specialist physicians or the NHS National Institute for Clinical Excellence—and this work draws on the results of systematic reviews of research evidence. Such organisations might find it useful to reconsider their approach to evidence and broaden the type of problems that they examine, especially when they need to balance risks and benefits. Most importantly, however, the practitioners who use their products should understand the approach used and be able to judge easily whether a review or a guideline has been prepared reliably.

Evidence hierarchies with the randomised trial at the apex have been pivotal in the ascendancy of numerical reasoning in medicine over the past quarter century.\text{\textsuperscript{11}} Now that this principle is widely appreciated, however, we believe that it is time to broaden the scope by which evidence is assessed, so that the principles of other types of research, addressing questions on aetiology, diagnosis, prognosis, and unexpected effects of treatment, will become equally widely understood. Indeed, maybe we do have something to learn from Michelin guides, they have separate grading systems for hotels and restaurants, provide the details of the several quality dimensions behind each star rating, and add a qualitative commentary (www.viamichelin.com).

<table>
<thead>
<tr>
<th>Benefits</th>
<th>Long term outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contraception</td>
<td>Q Effective (2 controlled trials\textsuperscript{17,18})</td>
</tr>
<tr>
<td>Dysmenorrhoea</td>
<td>Q Possible reduction in pain and work absence (systematic review of 5 poor-quality RCTs\textsuperscript{18})</td>
</tr>
<tr>
<td>No applicable</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Harms</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer</td>
<td>Increased risk: relative risk 1.24 (95% CI 1.15 to 1.33) for current users (individual patient data analysis of 54 observational studies\textsuperscript{20})</td>
</tr>
<tr>
<td>No increased risk detected 10 years after cessation (systematic review of 46 observational studies\textsuperscript{21})</td>
<td></td>
</tr>
</tbody>
</table>

| Venous thromboembolism | Increased risk: 2.5-fold to 5-fold increase in relative risk (systematic review of 5 non-randomised studies\textsuperscript{5}) |
| Return to background risk after cessation\textsuperscript{22} |

<table>
<thead>
<tr>
<th>Minimal or uncertain effects</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight gain</td>
<td>Q No weight gain (3 placebo-controlled RCTs of 4.3 months\textsuperscript{23})</td>
</tr>
<tr>
<td>Heavy menstrual bleeding</td>
<td>Q Insufficient evidence (one, 3 arm RCT with 43 participants\textsuperscript{18})</td>
</tr>
<tr>
<td>Unknown</td>
<td></td>
</tr>
</tbody>
</table>

Example of possible evidence table for short and long term effects of oral contraceptives. (Absolute effects will vary with age and other risk factors such as smoking and blood pressure. RCT = randomised controlled trial)

Summary points

- Different types of research are needed to answer different types of clinical questions
- Irrespective of the type of research, systematic reviews are necessary
- Adequate grading of quality of evidence goes beyond the categorisation of research design
- Risk-benefit assessments should draw on a variety of types of research
- Clinicians need efficient search strategies for identifying reliable clinical research

We thank Andy Oxman and Mike Rawlins for helpful suggestions.

Contributors and sources: As a general practitioner, PG uses the his own and others’ evidence assessments, and as a teacher of evidence based medicine helps others find and appraise research. JV is an internist and epidemiologist by training; he has extensively collaborated in clinical research, which made him strongly aware of the diverse types of evidence that clinicians use and need. IC’s interest in these issues arose from witnessing the harm done to patients from eminence based medicine.

Competing interests: None declared.

2 Sackett DL, Rosenberg JE. Choosing the best research design for each question. BMJ 1997;315:1636.
19 Mallett S, Clarke M. How many Cochrane reviews are needed to cover existing evidence on the effects of health care interventions. ACP J Club 2003;139:A11.

\text{\textsuperscript{1}} Campbell DT, Stanley JC. Experimental and quasi-experimental designs for research. Chicago, IL: Rand McNally College, 1969.
\text{\textsuperscript{2}} Sackett DL, Rosenberg JE. Choosing the best research design for each question. BMJ 1997;315:1636.
\text{\textsuperscript{5}} Vanderbroucke JP. Observational research and evidence-based medicine: What should we teach young physicians? J Clin Epidemiol 1998;51:1467-72.
\text{\textsuperscript{8}} Aronson JK. Anecdotes as evidence. BMJ 2003;320:1346.
\text{\textsuperscript{15}} Mallett S, Clarke M. How many Cochrane reviews are needed to cover existing evidence on the effects of health care interventions. ACP J Club 2003;139:A11.
\text{\textsuperscript{17}} Glasziou PP, Irving LI, for the GRADE Working Group. Letters, numbers, symbols and words how to communicate grades of evidence and recommendations. CMAJ 2003;169:677-80 (Accepted 20 October 2003)
Reliable assessment of the effects of treatment on mortality and major morbidity, I: clinical trials

Rory Collins, Stephen MacMahon

This two-part review is intended principally for practising clinicians who want to know why some types of evidence about the effects of treatment on survival, and on other major aspects of chronic disease outcome, are much more reliable than others. Although there are a few striking examples of treatments for serious disease which really do work extremely well, most claims for big improvements turn out to be evanescent. Unrealistic expectations about the chances of discovering large treatment effects could misleadingly suggest that evidence from small randomised trials or from non-randomised studies will suffice. By contrast, the reliable assessment of any more moderate effects of treatment on major outcomes—which are usually all that can realistically be expected from most treatments for most common serious conditions—requires studies that guarantee both strict control of bias (which, in general, requires proper randomisation and appropriate analysis, with no unduly data-dependent emphasis on specific parts of the overall evidence) and strict control of random error (which, in general, requires large numbers of deaths or of some other relevant outcome). Past failures to produce such evidence, and to interpret it appropriately, have already led to many premature deaths and much unnecessary suffering.

Some treatments for the chronic diseases of middle age have been found to produce large effects on death and disability. For example, it is obvious that prompt treatment of diabetic coma or cardiac arrest saves lives. But, given the heterogeneity of any particular condition (as indicated by the different survival durations of apparently similar patients) and the variety of different mechanisms that can lead to death or disability (only one of which may be appreciably influenced by any one treatment), hopes of large effects of treatment on major outcomes have often been unrealistically high. Some such expectations might derive from extrapolation of the effects of treatment on “surrogate” outcomes. For example, cardiac arrhythmias are associated with a poor prognosis, and antiarrhythmic drugs can markedly reduce their frequency. However, various antiarrhythmic regimens have been found to increase, rather than decrease, mortality. Many other treatments have large effects on one part of a disease process—for example, zidovudine on viral titre in early HIV infection, and radiotherapy on local recurrence in breast cancer—but uncertainty remains as to whether their routine use produces worthwhile improvements in survival. In general, if such uncertainty exists about a treatment, any effects on mortality or major morbidity are likely to be either negligibly small or of only moderate size. As will be discussed, support for this conclusion comes from the modest effects typically suggested by the aggregated results (ie, meta-analyses or systematic overviews) of all relevant clinical trials of any particular therapy for a chronic disease; and, in certain special cases, by the modest strength of the relation in observational studies between disease risk and a risk factor that treatment can modify (eg, blood pressure or cholesterol).

In many circumstances, even moderate improvements in survival or in major morbidity would still be regarded as worthwhile by patients and their doctors (provided, of course, that any benefits are not substantially offset by some serious adverse effects). Clearly, however, if such treatment effects are to be reliably detected or reliably refuted, then any errors in their assessment need to be much smaller than the difference between a moderate but worthwhile effect, and an effect that is too small to be of any material importance. Systematic errors (ie, biases) in the assessment of treatment can be produced by differences in factors other than the treatment under investigation (panel 1). Observational studies, in which outcome is compared between individuals who received the treatment of interest and those who did not, can be subject to large systematic errors. Instead, the guaranteed avoidance of material biases typically requires the proper randomised allocation of treatment and appropriate statistical analysis, with no unduly data-dependent emphasis on specific subsets of the overall evidence (panel 2). Random errors in the assessment of treatment effects relate to the impact of the play of chance on outcome among those exposed or not exposed to the treatment of interest (panel 1). These errors are determined by the number of deaths or other relevant outcomes in the study, and their size can be quantified (eg, in terms of a confidence interval that indicates the range of effects statistically compatible with the observed result). The only way to guarantee small random errors is to study...
Avoidance of moderate systematic errors
- Proper randomisation (non-randomised methods may cause moderate or large biases)
- Analysis by allocated treatment (including all randomised patients: intention-to-treat analysis)
- Chief emphasis on overall results (without undue data-dependent emphasis on particular subgroups)
- Meta-analyses of all relevant studies (without undue data-dependent emphasis on particular studies)

Avoidance of moderate random errors
- Large numbers of major outcomes in any new studies (with streamlined study methods to facilitate recruitment)
- Meta-analyses of all relevant studies (yielding the largest possible numbers of deaths and other major outcomes)

Panel 2: Requirements for reliable assessment of moderate treatment effects: simultaneous avoidance of moderate systematic errors and moderate random errors

Avoidance of moderate systematic errors
- Proper randomisation
- Analysis by allocated treatment
- Chief emphasis on overall results
- Meta-analyses of all relevant studies

Avoidance of moderate random errors
- Large numbers of major outcomes in any new studies
- Meta-analyses of all relevant studies

Large numbers of outcomes by doing large individual studies and large meta-analyses (panel 2). It is not much use, however, having very small random errors if there could be moderate biases, so even the large size of some observational studies cannot guarantee reliable assessment of moderate treatment effects. Clinical trials and observational studies have provided much of the available evidence about the effects on death and major non-fatal outcomes (such as heart attacks, strokes, cancers) of different treatments for disease. But not all such epidemiological evidence is reliable, and the consequences of this may be substantial: for example, ineffective or dangerous treatments might continue to be used, or effective and safe treatments might not be used appropriately widely. The first part of this review is concerned with the reliable demonstration of any moderate effects of treatment on mortality and major morbidity, which requires the simultaneous avoidance of moderate biases and moderate random errors. This requirement determines the need for appropriately large, properly randomised, trials. As will be discussed, non-randomised observational studies, and unduly small randomised trials or meta-analyses, are all much inferior as sources of evidence about such moderate, though potentially important, effects of treatment. In the second part of this review, the ways in which observational studies can be useful for the assessment of treatment effects are discussed; in particular, for the detection of large effects on rare outcomes, and for helping to generalise the results of randomised trials to different circumstances.

CLINICAL TRIALS: Minimising both systematic and random errors

Proper randomisation
The fundamental reason for random allocation of treatment in clinical trials is to maximise the likelihood that each type of patient will have been allocated in similar proportions to the different treatment strategies being investigated. In a properly randomised trial, the decision to enter a patient is made irreversibly in ignorance of which trial treatments that patient will be allocated. Foreknowledge of the next treatment allocation could affect the decision to enter the patient, and those allocated one treatment might then differ systematically from those allocated another. For example, in a study comparing amniotomy (rupture of membranes) versus oxytocin for induction of labour, the halving in poliomyelitis cases observed in the large non-randomised comparison between those children who had been vaccinated and those who had not been vaccinated was confirmed by the large randomised trial of vaccine versus placebo (table 2). But, since non-random methods introduce the potential for moderate biases, non-randomised studies cannot be guaranteed to provide appropriately unbiased assessments when the real effects of treatment are of moderate size. So, for example, the mortality reduction observed in the aggregate of all available randomised trials of oral anticoagulants for acute myocardial infarction was found to be only about a third as large as the highly significant 30–40% mortality reduction observed in the non-randomised concurrently-controlled studies (which mainly used alternate allocation). Hence, the biases inherent in non-randomised studies can be at least as big as any moderate effects of treatment on mortality and major morbidity that might exist.

Intention-to-treat analysis
Even in a properly randomised trial, bias can be inadvertently introduced by the post-randomisation exclusion of certain patients (such as those who are non-compliant with study treatment), especially if the prognosis of those excluded from one treatment group differs from that of those excluded from another. This point is illustrated by the Coronary Drug Project randomised trial of cholesterol-lowering therapy: patients who took at least 80% of their

Table 1: Imbalance in patients’ characteristics between treatment groups due to foreknowledge of treatment allocation: trial of amniotomy or oxytocin for induction of labour

<table>
<thead>
<tr>
<th>Cervical “ripeness”</th>
<th>Amniotomy: odd dates of birth (n=142)</th>
<th>Oxytocin: even dates of birth (n=113)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>Intermediate</td>
<td>58</td>
<td>56</td>
</tr>
<tr>
<td>Most</td>
<td>45</td>
<td>29</td>
</tr>
</tbody>
</table>

Comparison of the distribution between treatment groups of cervical ripeness before treatment allocation: χ²=16·1 (p<0·0005).

Table 2: Confirmation by randomised trial of observed effect in non-randomised trial: Salk vaccine for poliomyelitis

<table>
<thead>
<tr>
<th>Type of study</th>
<th>Poliomyelitis cases/total (rate per 100 000)</th>
<th>Odds ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vaccine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-randomised</td>
<td>60/233 902 (26)</td>
<td>381/725 173 (64)</td>
</tr>
<tr>
<td>Randomised**</td>
<td>57/200 745 (28)</td>
<td>142/201 229 (71)</td>
</tr>
</tbody>
</table>

*Excludes 8484 vaccine-allocated and 8577 placebo-allocated non-compliant children with data on outcome not fully available.

Table 2: Confirmation by randomised trial of observed effect in non-randomised trial: Salk vaccine for poliomyelitis
allocated clofibrate had substantially lower 5-year mortality than those who did not (15·0% vs 24·6%, respectively; p=0·0001), but there was an even more striking difference in outcome between good and poor compliers in the placebo group (15·1% vs 28·3%, respectively; p<0·00001). The primary statistical analysis of any trial should, therefore, compare outcome among all those originally allocated one treatment (even though some of them may not have actually received it) with outcome among all those allocated the other treatment—that is, an intention-to-treat analysis of the impact of a general policy of using the treatment. This is not to say that additional analyses may not also be of value: for example, in describing the frequency of some very specific side-effect, it may be preferable to describe its incidence only among those who actually received the treatment because strictly randomised comparisons might not be needed to assess extreme relative risks.\footnote{1}

Since there is bound to be some non-compliance with the allocated treatments in clinical trials, intention-to-treat analyses will tend to underestimate the effects produced by full compliance with the study treatments. But, rather than using potentially biased “on treatment” comparisons among only those who were compliant, more appropriate allowance can be made by applying an approximate estimate of the level of compliance to the estimate of the treatment effect provided by the intention-to-treat comparison.\footnote{21} For example, in a meta-analysis of the randomised trials of prolonged use of aspirin and other antiplatelet agents among patients with occlusive vascular disease, the average compliance 1 year after treatment allocation seemed to have been no more than 80%.\footnote{22} Application of this estimate of compliance to the proportional reduction of about 30% in non-fatal heart attacks and strokes estimated from intention-to-treat analyses of these trials suggests that full compliance with antiplatelet therapy produces reductions in risk of about 35–40%.

**Problems produced by data-dependent emphasis**

Apparent differences between therapeutic effects in different subgroups of study participants can often be produced just by the play of chance and, in particular subgroups, chance can mimic or obscure moderate treatment effects. For example, in the large Second International Study of Infarct Survival (ISIS-2) randomised trial of the emergency treatment of heart attacks, the 1-month survival advantage produced by aspirin was particularly clear (804 vascular deaths among 8857 patients allocated aspirin vs 1016 among 8600 allocated placebo-control; proportional reduction of 23% [SD 4]; p<0·000001).\footnote{23} To illustrate the unreliability of subgroup analyses, these overall results were subdivided by the patients’ astrological birth signs into 12 subgroups. In some subgroups the results for aspirin were about average, but in others they were, by chance, slightly better or slightly worse than average. Taking the subgroups with the least promising results, which happened to be Libra or Gemini, no fewer deaths were observed with aspirin than with placebo (table 3). Clearly, it would be unwise to conclude from such an analysis that patients born under the astrological birth signs of Libra or Gemini are unlikely to benefit from aspirin. Yet, similar conclusions based on “exploratory” data-derived subgroup analyses that are no more reliable than these are often reported and may be accepted, with inappropriate emphasis on practice. For example, despite the highly significant survival advantage observed overall in the large Gruppo Italiano per lo Studio della Streptochinasi nell’infarto miocardico (GISSI) randomised trial, it was suggested that fibrinolytic therapy might not save lives among patients who had had a previous heart attack (based on 157 deaths among such patients allocated streptokinase vs 147 among those allocated control).\footnote{24} By contrast, subsequent trials have shown unequivocally that the benefits of fibrinolytic therapy are similar among those with and without a history of prior infarction.\footnote{25} In another example of the impact of unduly selective emphasis on small subgroups in particular trials, the use of aspirin after transient ischaemic attacks was, until very recently, approved in the USA for men but not for women.\footnote{26} This has turned out to have been a lethal error, resulting in many women being denied a life-saving treatment that produces about the same benefits for women as for men.\footnote{27}

Similarly, when several studies have all addressed much the same therapeutic question, choice of only a few of them for emphasis could be a source of serious bias, since chance fluctuations for or against treatment might affect this choice. To avoid such bias, it is often appropriate to base inference chiefly on a meta-analysis of all of the results from all randomised trials that have addressed the particular question (or, at least, on an unbiased subset of the relevant trials).\footnote{28} Such meta-analyses will also minimise random errors in the assessment of treatment effects because far more patients (and, most importantly, more events) are typically included in a meta-analysis than in any individual trial that contributes to it. The separate trials might well be heterogeneous, but this merely argues for careful interpretation of the results of any meta-analysis (rather than arguing against any such analyses)\footnote{29} since, without meta-analyses, moderate biases and random errors often cannot both be avoided reliably. For example, meta-analysis of the relevant randomised trials showed clearly that prolonged antiplatelet therapy after myocardial infarction reduces the risk of major vascular events (ie, death, recurrent infarction, or stroke) by about a quarter (figure 1).\footnote{30} These findings have led to the appropriately widespread use of such treatment (in particular, low-dose aspirin), and the prevention of tens of thousands of deaths and disabling events each year worldwide. By contrast, selective emphasis on the trial with the least promising result\footnote{31} could lead to the dangerously misleading conclusion that antiplatelet therapy is not beneficial for such patients.\footnote{32} Similarly, the inference drawn from a subgroup of one trial that the beneficial effects of angiotensin-converting-enzyme inhibitors on mortality and hospital admission for heart failure are lost in the presence of aspirin\footnote{33} is not supported by a meta-analysis of all such trials in patients with ventricular dysfunction.\footnote{34}

**Subgroups defined by post-randomisation characteristics**

In general, any prognostic features that are to be used in analyses of treatment effects in randomised trials should be irreversibly recorded before the treatment is allocated. For, if the recorded value of some feature is affected by the trial treatment allocation, then comparisons within subgroups that are defined by that factor might be biased. As an example, consider a study of mastectomy with axillary clearance versus lumpectomy alone for women with breast...

---

**Table 3: Unreliability of “data-dependent” subgroup analyses:**

<table>
<thead>
<tr>
<th>Astrological birth sign</th>
<th>Vascular death by 1 month</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aspirin</td>
<td>Placebo</td>
</tr>
<tr>
<td>Libra or Gemini</td>
<td>150 (11·3%)</td>
<td>147 (10·2%)</td>
</tr>
<tr>
<td>All other signs</td>
<td>654 (9·0%)</td>
<td>869 (12·1%)</td>
</tr>
<tr>
<td>Any birth sign</td>
<td>804 (9·4%)</td>
<td>1016 (11·8%)</td>
</tr>
</tbody>
</table>
cancer. An unusually careful search of the axilla among those allocated axillary clearance could result in the discovery of tiny deposits of cancer cells that would otherwise have been overlooked. Hence, some of the women in the axillary clearance group for whom the staging was less careful.31 Similarly, in randomised trials of treatment for heart attacks were reported.34 Each of those mega-trials, by contrast with failure of previous much smaller trials GISSI-1 and ISIS-2 trials of fibrinolytic therapy among 12 000 and 17 000 patients with acute myocardial infarction,23,24 along with results of small trials contributing to a previous meta-analysis.32 Conventions as in figure 1. that might exist (especially if effects in different subgroups are to be assessed reliably).3 For example, between the late 1950s and the early 1980s, about two dozen randomised trials of intravenous fibrinolytic therapy for the emergency treatment of heart attacks were reported.35 Each of those trials was too small—none involved even 1000 patients—to provide reliable evidence about any moderate effects of this treatment on mortality (figure 2), although several were large enough to show the large relative effects on bleeding. As a result, fibrinolytic therapy was generally regarded as both ineffective and dangerous, and so not appropriate for routine coronary care. By contrast, during the mid-1980s, the GISSI-131 and ISIS-232 “mega-trials” each involved more than 10 000 patients (and, most relevantly, more than 1000 deaths), and provided such definite evidence about the beneficial effects of fibrinolytic therapy that worldwide treatment patterns changed rapidly. Consequently, at least half a million patients per year are now given fibrinolytic treatment, avoiding at least 10 000

<table>
<thead>
<tr>
<th>Trial</th>
<th>Vascular events/patients</th>
<th>Odds ratio (95% CI)</th>
<th>Trial author (date published)</th>
<th>Deaths/patients</th>
<th>Odds ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Antithrombotic group</td>
<td>Control group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiff-I</td>
<td>397/615</td>
<td>676/624</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiff-II</td>
<td>129/847</td>
<td>186/878</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PARIS-I</td>
<td>262/1620</td>
<td>4x(82/406)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PARIS-II</td>
<td>179/1563</td>
<td>235/1565</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMIS-I</td>
<td>47/651</td>
<td>102/771</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ART-I</td>
<td>89/122</td>
<td>130/816</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARIS-I</td>
<td>48/566</td>
<td>55/362</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Micristin</td>
<td>65/672</td>
<td>106/668</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rome</td>
<td>3/40</td>
<td>4/40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>1331/9877</td>
<td>1693/9914</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test for heterogeneity: $X^2_{10} = 12.3; p<0.1$

Figure 1: Clear demonstration of worthwhile benefits in meta-analysis of available trial data, by contrast with failure of individual trials to provide convincing evidence Vascular events (ie, death, myocardial infarction, or stroke) in collaborative meta-analysis of trials of prolonged antithrombotic therapy after myocardial infarction.21 Numbers in the control group of one trial with a deliberately uneven allocation have been adjusted so that the overall numbers allocated antithrombotic therapy and control are similar, but all statistical calculations are based on actual numbers studied. Black squares=point estimates (with area proportional to number of events) and horizontal lines=95% CI for observed effects in individual trials (with arrow head when CI extends beyond odds ratio axis). Diamond=point estimate and CI for overall effect.

Avoidance of moderate random errors

Problems with false-negative results
It is still not sufficiently widely appreciated just how large clinical trials need to be to detect reliably the sort of moderate, but important, differences in major outcomes that might exist (especially if effects in different subgroups are to be assessed reliably).3 For example, between the late 1950s and the early 1980s, about two dozen randomised trials of intravenous fibrinolytic therapy for the emergency treatment of heart attacks were reported.35 Each of those trials was too small—none involved even 1000 patients—to provide reliable evidence about any moderate effects of this treatment on mortality (figure 2), although several were large enough to show the large relative effects on bleeding. As a result, fibrinolytic therapy was generally regarded as both ineffective and dangerous, and so not appropriate for routine coronary care. By contrast, during the mid-1980s, the GISSI-131 and ISIS-232 “mega-trials” each involved more than 10 000 patients (and, most relevantly, more than 1000 deaths), and provided such definite evidence about the beneficial effects of fibrinolytic therapy that worldwide treatment patterns changed rapidly. Consequently, at least half a million patients per year are now given fibrinolytic treatment, avoiding at least 10 000

<table>
<thead>
<tr>
<th>Trial</th>
<th>Deaths/patients</th>
<th>Odds ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fletcher (1959)</td>
<td>1/12</td>
<td>4/11</td>
</tr>
<tr>
<td>Dewar (1963)</td>
<td>4/21</td>
<td>7/21</td>
</tr>
<tr>
<td>Lippchutz (1965)</td>
<td>6/43</td>
<td>7/41</td>
</tr>
<tr>
<td>1st European (1969)</td>
<td>20/83</td>
<td>15/84</td>
</tr>
<tr>
<td>Heikenheimi (1971)</td>
<td>22/219</td>
<td>17/207</td>
</tr>
<tr>
<td>Italian (1971)</td>
<td>19/164</td>
<td>18/157</td>
</tr>
<tr>
<td>2nd European (1971)</td>
<td>69/373</td>
<td>94/357</td>
</tr>
<tr>
<td>Australian (1973/77)</td>
<td>51/376</td>
<td>63/371</td>
</tr>
<tr>
<td>2nd Frankfurt (1973)</td>
<td>13/102</td>
<td>29/104</td>
</tr>
<tr>
<td>Gormsen (1973)</td>
<td>2/14</td>
<td>3/14</td>
</tr>
<tr>
<td>NHLBI SWAT (1974)</td>
<td>7/53</td>
<td>3/54</td>
</tr>
<tr>
<td>Frank (1975)</td>
<td>6/55</td>
<td>6/53</td>
</tr>
<tr>
<td>Valen (1975)</td>
<td>11/49</td>
<td>9/42</td>
</tr>
<tr>
<td>Brocher (1975)</td>
<td>2/60</td>
<td>8/60</td>
</tr>
<tr>
<td>European Collaborative (1975)</td>
<td>41/172</td>
<td>34/169</td>
</tr>
<tr>
<td>UK Collaborative (1976)</td>
<td>48/302</td>
<td>52/293</td>
</tr>
<tr>
<td>Klein (1976)</td>
<td>4/14</td>
<td>1/9</td>
</tr>
<tr>
<td>Austrian (1977)</td>
<td>37/352</td>
<td>65/376</td>
</tr>
<tr>
<td>Wlchtz (1977)</td>
<td>5/32</td>
<td>5/26</td>
</tr>
<tr>
<td>Lasers (1977)</td>
<td>3/13</td>
<td>3/11</td>
</tr>
<tr>
<td>N German Collaborative (1977)</td>
<td>63/249</td>
<td>51/234</td>
</tr>
<tr>
<td>3rd European (1979)</td>
<td>25/156</td>
<td>50/159</td>
</tr>
<tr>
<td>Schreiber (1984)</td>
<td>1/19</td>
<td>4/19</td>
</tr>
</tbody>
</table>

Overall: small trials

|                             |                       |                     |
| 463/2961                    | 553/2896              | 19-15%              |
|                             | 22% (SD 6) reduction   | (p<0.0005)          |
| Italian (1971)              | 628/5860              | 758/5852            |
| (10-7%)                     | (13-0%)               |                     |
|                             | 13% (SO 5)            | (p<0.0005)          |
| ISIS-2 mega-trial (1988)     | 795/6592              | 1033/8195           |
| (9-3%)                      | (12-0%)               |                     |
|                             | 25% (SO 4) reduction   | (p<0.000001)        |

Figure 2: Clear demonstration of worthwhile benefits in mega-trials, by contrast with failure of previous much smaller trials GISSI-1 and ISIS-2 trials of fibrinolytic therapy among 12 000 and 17 000 patients with acute myocardial infarction,23,24 along with results of small trials contributing to a previous meta-analysis.32 Conventions as in figure 1.
premature deaths annually. But, if GISSI-1 and ISIS-2 had been only a tenth as large (which would still have been larger than any of the previous trials), the observed reduction in mortality of about a quarter would not have been conventionally significant, and would therefore have had much less influence on medical practice. Indeed, the inadequate size of the earlier trials—which delayed the convincing demonstration of the benefits of fibrinolytic therapy for more than two decades—can now be seen to have been the cause of some hundreds of thousands of unnecessary deaths.

Problems with false-positive results

Small-scale evidence about the effects of treatment on major outcomes (whether from a single randomised trial or from a meta-analysis of trials) is often unreliable, and will frequently be found in retrospect to have been misleading. For example, a review of the small randomised trials of antiplatelet therapy in pregnancy suggested that such treatment reduced the incidence of pre-eclampsia by about three-quarters, and produced a much better outcome for the fetus (with less intrauterine growth retardation and fewer perinatal deaths). By contrast, the effects in several subsequent, much larger, randomised trials were much less promising, indicating reductions of only about a sixth in pre-eclampsia and no apparent improvement in fetal outcome. Small-scale evidence from randomised trials can be misleading not just about the size but even about the direction of the effects of treatment on major outcomes. For example, it was concluded from a small randomised trial among patients with heart failure that the inotropic agent vesnarinone more than halved the risk of death (13 vesnarinone vs 33 placebo deaths, p=0·002). By contrast, when the same regimen was studied in much larger numbers of the same type of patient, mortality was significantly increased (292 vesnarinone vs 242 placebo deaths, p=0·02). Further examples of treatments for which extreme observations from initial small trials have not been confirmed by much larger randomised trials include calcium supplementation for the prevention of pre-eclampsia, of intravenous nitrates or magnesium for the emergency treatment of heart attacks, of heparins or calcium antagonists for the emergency treatment of strokes, and of vitamin E for the prevention of coronary disease. Frequent disappointment with small randomised trials can be traced to a number of factors. Sometimes, the results published in the initial small trials have been somewhat misleading because the patients were so different from one another that it is only when the effects of treatment on outcome are compared among sufficiently large groups of patients divided at random that the proportions of patients with good and bad prognoses allocated the different treatments can be relied on to be sufficiently similar. Moreover, the identification of those particular types of patient most likely to benefit from a treatment will often require even larger-scale evidence from randomised trials, and even more careful interpretation, than is required to show an overall treatment effect reliably. There are three main remedies for this unavoidable conflict between the reliable subgroup-specific conclusions that doctors and their patients want, and the unreliable findings that subgroup analyses of clinical trials might offer (panel 3).

Basing inference on overall effects on particular outcomes

The first approach is to emphasise chiefly the overall results of a trial—or, better still, of a meta-analysis of all such trials—for particular outcomes as a guide to the qualitative results in various specific subgroups of patients, and to give less weight to the actual results in each separate subgroup. This is clearly the right way to interpret the astrophysical subgroups in table 3, but it is also likely in many other circumstances to provide the best assessment of whether a treatment is effective in particular subgroups. For example, on the basis of adjusted analyses of large observational databases, it has been claimed that 1-month mortality is increased by fibrinolytic therapy in patients aged 75 or older who present within 12 h with electrocardiographic changes indicative of acute myocardial infarction. By contrast, a meta-analysis of the major randomised trials of fibrinolytic therapy has provided especially strong evidence of overall benefit, with no significant difference between the mortality reductions seen among such patients at different ages: 27 (SD 3) fewer deaths per 1000 patients younger than 75 compared with 34 (SD 16) fewer deaths per 1000 older patients. Hence, when a treatment has been shown unequivocally to

Panel 3: Estimating the effects of treatment in particular types of patient

- Base inference on the overall effects observed on particular outcomes (without unduly selective emphasis on the results in each separate subgroup of patients)
- Give greater emphasis to results in prespecified, rather than retrospectively data-derived, subgroups (provided they involve sufficiently large numbers of outcomes)
- Consider subgroup analyses of mortality in the context of analyses of other relevant major outcomes (which might be more statistically stable)
be beneficial overall, really good evidence should be required of lack of benefit in some particular subgroup (rather than merely lack of a clearly significant effect in that subgroup taken on its own) before it is considered safe to conclude that the treatment is not of value for such patients.

Because the effects of treatment on different outcomes may differ in terms of size or direction, estimates from trials of the separate effects on each outcome are likely to be more widely generalisable than would be an estimate of the combined effect on these outcomes. For example, the average 5–6 mm Hg reduction in diastolic blood pressure achieved in previous trials of antihypertensive therapy produced proportional risk reductions of about 40% for stroke and of about 15% for coronary heart disease, and each of these proportional effects seemed to be similar among different types of patient.6 Hence, the relative frequency of strokes and of coronary events in different circumstances will influence both the proportional and absolute effects of blood-pressure lowering on the overall risk of vascular disease. Similarly, for endarterectomy in patients with symptomatic carotid-artery stenosis, the net effect on stroke risk is dependent on the balance in a particular population between the beneficial effects of surgery on ipsilateral stroke and the adverse effects of surgery on other strokes. Consequently, estimates from trials of the separate effects on each of these types of stroke would be expected to be more informative about the net effect on the risk of stroke in different situations64 than would the overall effects on total stroke observed in any single population.64

Prespecification of analyses within particular subgroups

The second approach to determining effects in particular types of patient is to prespecify a limited number of subgroup analyses, provided there are good a priori reasons for anticipating that the effect of treatment might be different in different circumstances. Generally, such prespecified analyses should then be taken more seriously than other subgroup analyses, as long as they are based on sufficiently large numbers of events. For example, the benefits of fibrinolytic therapy for heart attacks were expected to be greater the earlier patients were treated, so some studies prespecified that the analyses should be subdivided by time from onset of symptoms to treatment. None of the individual studies of fibrinolytic therapy could show this clearly on its own, but a meta-analysis of the major trials included large enough numbers of patients to show that the benefit was indeed greatest for those treated earliest after the onset of acute myocardial infarction (although the mortality reduction was still substantial for those treated several hours after symptom onset).29

Interpretation of mortality analyses in the context of morbidity analyses

Finally, in considering the likely effects of treatment on the survival of particular patients, it might be useful to take account not only of the mortality data in specific subgroups but also of the data on some other relevant major outcomes (eg, recurrence-free survival in cancer trials, or non-fatal as well as fatal myocardial infarction in heart disease trials). For, if the overall results for such outcomes are similar but much more highly significant than for mortality (due chiefly to the larger number of events, but perhaps also because effects on non-fatal outcomes emerge more rapidly), subgroup analyses of these major outcomes will be more stable. Hence, they may provide a better guide to the existence of any large differences between subgroups in the effects of treatment (particularly if such subgroup analyses were specified before results were available). For example, in the early 1990s, a collaborative meta-analysis of all relevant randomised trials of the oestrogen-receptor-blocking drug tamoxifen in women with early breast cancer showed clearly that tamoxifen reduces the risks of breast cancer recurrence and of death from breast cancer among postmenopausal women.65 Far fewer data were available at the time for premenopausal women and, although there was a definite improvement in recurrence-free survival, there was no clear improvement in survival among such women considered on their own. As a consequence, tamoxifen was not used routinely for these younger women,66 yet it has recently been shown that prolonged treatment with tamoxifen produces substantial survival advantages not only for postmenopausal but also for premenopausal women.67 In retrospect, therefore, the decision by many clinicians not to place sufficient emphasis on the overall findings for survival, supported by the age-specific benefits for recurrence, was mistaken.

**SUMMARY: The need for large-scale randomised evidence**

In a world in which moderate effects of treatment on mortality or major morbidity are generally more plausible than large effects, claims of striking effects from small-scale randomised trials, and from other sources (including observational studies1), will often prove evanescent. The assumption that both a moderate difference or no difference may be plausible, and that an extreme difference is much less so, has surprisingly strong consequences for the interpretation of evidence from trials. In particular, it implies that even highly significant (eg, 2p=0·001) differences that are based on only relatively small numbers of events in selected studies may provide untrustworthy evidence of the existence of any real difference66—as with the initial results for aspirin in pre-eclampsia,35 vesnarinone in heart failure,59 magnesium in heart attacks,60 and heparin in stroke.44 For this reason, recent claims of large effects based on small randomised trials (eg, the healing of leg ulcers with oral aspirin;69 or the prevention of coronary outcomes with antioxidants,70 of dementia with anti-hypertensive therapy,71 or of either pre-eclampsia72 or vascular complications in endstage renal disease73 with antioxidant vitamins) should probably be treated with far greater caution—both by journal editors and by their readers—than is often, at present, customary. Moreover, when there is not good evidence of any effect on major outcomes, estimates of the “number needed to treat” to prevent such outcomes are of little or no value, and it is particularly inappropriate to fail to provide a clear indication of the range of uncertainties around such estimates74 (as, for example, with the claim that lowering blood pressure could prevent 19 cases of dementia per 1000 patients treated for 5 years,75 when the results were also compatible with the prevention of no cases of dementia).

As will be discussed in the second part of this review,1 observational studies may provide useful evidence about any large effects of treatment that do exist (such as rare, but serious, hazards), and about the risks of death and disability in particular types of patient that may help to generalise from clinical trials to clinical practice. But, only sufficiently large-scale evidence from randomised trials can reliably assess moderate effects of treatment on mortality and major morbidity—and past failures to produce such evidence, and to interpret it appropriately, has already led to many premature deaths and much unnecessary suffering.
References


A list of further reading can be found at www.thelancet.com
Reliable assessment of the effects of treatment on mortality and major morbidity, II: observational studies

Stephen MacMahon, Rory Collins

Observational studies and randomised trials can contribute complementary evidence about the effects of treatment on mortality and on major non-fatal outcomes. In particular, observational studies have an important role in the identification of large adverse effects of treatment on infrequent outcomes (ie, rare, but serious, side-effects) that are not likely to be related to the indications for (or contraindications to) the treatment of interest. Such studies can also provide useful information about the risks of death and disability in particular circumstances that can help to generalise from clinical trials to clinical practice. But, due to their inherent potential for moderate or large biases, observational studies have little role in the direct assessment of any moderate effects of treatment on major outcomes that might exist. Instead, sufficiently large-scale evidence from randomised trials is needed to assess such treatment effects appropriately reliably. Wider appreciation of the different strengths and weaknesses of these two types of epidemiological study should increase the likelihood that the most reliable evidence available informs decisions about the treatments doctors use—and patients receive—for the management of a wide range of life-threatening conditions.

Epidemiological studies of the effects of treatments on mortality and major non-fatal outcomes can take the form of either clinical trials or observational studies. The first part of this review1 dealt with clinical trials—in particular, those in which the treatment is assigned to patients at random. As discussed, randomisation minimises systematic errors (ie, biases) in the estimates of treatment effects, allowing any moderate effects that exist to be detected unbiasedly in studies of appropriately large size.1 By contrast, observational studies—such as cohort studies and case-control studies—involves comparisons of outcome among patients who have been exposed to the treatment of interest, typically as part of their medical care, with outcome among others who were not exposed (or comparisons between those with different amounts of exposure). The reasons why certain patients received a particular treatment while others did not are often difficult to account for fully, and, largely as a consequence, observational studies are more prone to bias than are randomised trials. The primary objective of the second part of this review is to distinguish between situations in which biases in observational studies could lead to misleading conclusions and those in which such studies could provide useful evidence about the effects of treatment.

OBSERVATIONAL STUDIES: Non-randomised assessment of treatment

Assessment of adverse effects of treatment

Observational studies can have an important role in the identification of large adverse effects of treatments, particularly on infrequent outcomes that are not likely to be related to the indications for, or contraindications to, the treatment of interest (panel 1). Perhaps one of the best illustrations of this is the detection of increased risks of abnormal fetal limb development after maternal use of thalidomide.2 A decade later, observational studies also detected the many-fold increased risk of vaginal clear-cell adenocarcinoma among the daughters of women who used diethylstilboestrol.3 Other more recent examples include the demonstration of a 20-fold increased risk of cardiac-valve abnormalities among patients taking the appetite-suppressant drugs fenfluramine, dexfenfluramine, and phentermine (table 1), and even larger increases in the risk of Stevens-Johnson syndrome and toxic epidermal necrolysis with antiepileptic therapy.4 In each of these examples, the outcome was rare among unexposed individuals and the excess risk was large among exposed individuals, making it unlikely that systematic errors could reasonably account for the entire association.

On the other hand, since the disease of interest is rare in such circumstances, individual studies may well involve too few cases to detect, or quantify reliably, even large increases in risk. Hence, to minimise random error, combined analyses of the aggregated results (ie, meta-analyses) of all relevant observational studies are being done with increasing frequency. For example, a meta-analysis found that more than 10 years of oestrogen replacement therapy unopposed by progestagen was associated with almost a ten-fold increase in the risk of endometrial cancer among postmenopausal women.6 Such large effects are unlikely to be entirely the consequence of bias, but it is not so easy to exclude the possibility that biases might largely or wholly explain more modest increases in risk: for

Panel 1: Situations in which an observational study is more likely to provide reliable evidence about adverse effects of treatment

- The outcome of interest is rare among individuals not exposed to the treatment
- The excess risk among individuals exposed to the treatment is large (eg, a several-fold increase in risk)
- There are no obvious sources of bias likely to account for most, or all, of the observed association
example, the 40% increased incidence of malignant melanoma seen among users of hormone replacement therapy in another meta-analysis of observational studies. For, although these meta-analyses may help avoid the biases produced by unduly selective emphasis on particular parts of the available evidence (as with meta-analyses of randomised trials), the combination of observational evidence that is subject to other systematic errors might merely compound those biases—that is, produce more precise, but still biased, estimates of the effects of treatment.

Assessment of beneficial effects of treatment

Reliable evidence about the effects of treatment on mortality and major morbidity can also emerge from observational studies when outcome among untreated patients is typically poor and a large proportion of patients derive benefit from the treatment. For example, the beneficial effects of penicillin on survival in patients with sepsis, and of antihypertensive treatment on death and stroke in patients with malignant hypertension, were demonstrated in simple case series. Another example is provided by oral rehydration therapy, which seemed to reduce mortality from about 30% to less than 5% when introduced during a cholera epidemic among Bangladeshi refugees. But, as discussed in the first part of this review, outcome for many other common serious conditions is less predictable, and the most plausible expectation of benefit is that a treatment produces only a good evidence of benefit from the treatment.18

Table 1: Detection of large adverse effects of treatment in an observational study: cardiac-valve regurgitation with appetite-suppressant drugs

<table>
<thead>
<tr>
<th>Valve abnormalities</th>
<th>Any appetite suppressant (n=233)</th>
<th>Unexposed controls (n=233)†</th>
<th>Odds ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>53 (23%)</td>
<td>3 (1%)</td>
<td>22.6 (7.1-114.2)</td>
</tr>
</tbody>
</table>

*163 on fenfluramine and phentermine, 31 on dexfenfluramine and phentermine, and 39 on dexfenfluramine alone.† Matched for sex, age, height, and body mass index.

Table 2: Major sources of bias in observational studies of treatment

- **Confounding:** A factor (such as pre-existing disease severity) is associated with the use (or avoidance) of the treatment and, independently, influences the risk of the outcome of interest.
- **Recall bias:** The reliability of recall of treatment exposure differs between those who develop an adverse outcome and those who do not.
- **Detection bias:** The reliability of detection of adverse outcomes differs between those exposed to the treatment of interest and those not exposed.

Factor is associated with the exposure of interest—but is not a direct consequence of it—and, independently, influences the risk of the outcome of interest (panel 2). Observational studies of the effects of exposure to treatment are particularly prone to confounding by indication (or by contraindication), with the development of a medical condition leading both to the use of the treatment (or its avoidance) and to the outcome of interest. This type of bias can produce misleading estimates not just of the size but also of the direction of treatment effects, depending on the nature of the associations between the confounding factors and the outcome.

A recent example of misleading evidence about the size of a treatment effect is provided by a large observational study in which patients who received β-blockers after myocardial infarction were about half as likely to die as those who did not receive such treatment (table 2). By contrast, large-scale evidence from randomised trials has clearly shown that long-term β-blocker use in patients with a history of myocardial infarction reduces the risk of death by only about a quarter (as have trials in higher-risk patients with congestive heart failure). The patients who received β-blockers in this observational study were significantly younger, and had a lower-risk medical history, than those who did not. Statistical adjustments were made for these, and other, potential confounding factors that had been recorded, but such adjustments may well be incomplete due to insufficient correction for factors that were recorded (because of random errors in their measurement) and to lack of correction for other relevant factors. Hence, it seems likely that the overestimation in this observational study of the survival advantage produced by β-blocker therapy reflects some residual bias (due, perhaps, to a selective tendency for these drugs to be used less frequently in higher-risk patients).

Table 2: Different sizes of apparent effect in an observational study and in randomised trials

<table>
<thead>
<tr>
<th>Deaths/patients</th>
<th>Risk ratio † (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>β-blocker*</td>
<td>0.57 (0.47-0.69)</td>
</tr>
<tr>
<td>No β-blocker</td>
<td></td>
</tr>
</tbody>
</table>

† Treatment recorded at baseline in the observational study and assigned at random in the trials. †Multivariate adjusted relative risk in the observational study, and stratified odds ratio in the meta-analysis of randomised trials. †Exact numbers in each treatment group of the observational study were not reported.
heart disease (as well as those of stroke). Similarly, antihypertensive treatment reduces the risks of coronary controlled trials have clearly demonstrated that stroke among patients who had been taking 650–1300 (NASCET) indicated that the risk of perioperative outcome among participants in the North American example, retrospective observational analyses of evidence about the effects of different drug doses. For these examples, due to a tendency for the treatments to bias remains the most likely explanation (probably, in the observational study and stratified odds ratio in the meta-analysis of randomised trials. An example of misleading evidence about the direction of a treatment effect is provided by an observational study in which there was almost a two-fold greater risk of coronary events among patients receiving antihypertensive therapy than among those not receiving such treatment (table 3). By contrast, randomised controlled trials have clearly demonstrated that antihypertensive treatment reduces the risks of coronary heart disease (as well as those of stroke). Similarly, whereas a large observational study found nearly a doubling in the risk of major coronary events among those regularly taking aspirin,73 randomised controlled trials have shown unequivocally that antiplatelet therapy reduces the risks of heart attacks by about a quarter. These misleading findings from observational studies persisted after statistical adjustment for a variety of confounding factors and after restriction of analyses to individuals without a recorded history of cardiovascular disease. Once again it seems that uncontrolled residual bias remains the most likely explanation (probably, in these examples, due to a tendency for the treatments to be used more frequently in higher-risk patients). Fortunately, for both antihypertensive and antiplatelet therapy, the evidence from the randomised trials has chiefly influenced practice patterns, resulting in the appropriately widespread use of these treatments and the consequent prevention of many hundreds of thousands of premature deaths each year. By contrast, reliance on the evidence from the observational studies might have led to the inappropriate abandonment of these treatments (or, at the very least, to restriction of their use) and to much unnecessary suffering. Observational studies can also provide misleading evidence about the effects of different drug doses. For example, retrospective observational analyses of outcome among participants in the North American Symptomatic Carotid Artery Endarterectomy Trial (NASCET) indicated that the risk of perioperative stroke among patients who had been taking 650–1300 mg aspirin daily was less than half that among patients who had taken lower doses (table 4). Subsequently, however, a randomised trial designed to test this hypothesis in patients undergoing carotid endarterectomy found a non-significantly higher stroke incidence with 650–1300 mg/day aspirin than with lower doses (as well as a marginally significant higher risk of the composite of stroke, myocardial infarction or death). In this instance, reliance on the evidence from the observational study alone could have led to the inappropriate abandonment of lower-dose regimens which cause fewer side-effects and are better tolerated long-term.

### Table 3: Different directions of apparent effect in an observational study and in randomised trials: antihypertensive therapy and coronary heart disease

<table>
<thead>
<tr>
<th>CHD events/patients</th>
<th>Antihypertensive therapy*</th>
<th>No antihypertensive therapy*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observational study</strong></td>
<td>50/839 (6%)</td>
<td>420/20475 (2%)</td>
</tr>
<tr>
<td><strong>Randomised trials</strong></td>
<td>934/23847 (4%)</td>
<td>1104/23806 (5%)</td>
</tr>
<tr>
<td><strong>Risk ratio† (95% CI)</strong></td>
<td>1.8 (1.3–2.6)</td>
<td>0.84 (0.77–0.92)</td>
</tr>
</tbody>
</table>

*Treatment recorded at baseline in the observational study and assigned at random in the trials. †Multivariate adjusted relative risk in the observational study and stratified odds ratio in the meta-analysis of randomised trials.

### Table 4: Discordance between apparent effects of different drug doses in an observational study and a randomised trial: higher-dose versus lower-dose aspirin and stroke after carotid endarterectomy

<table>
<thead>
<tr>
<th>Stroke/patients</th>
<th>Lower-dose aspirin (≤650 mg daily)*</th>
<th>Higher-dose aspirin (650–1300 mg daily)*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observational study</strong></td>
<td>96/1391 (7%)</td>
<td>15/835 (2%)</td>
</tr>
<tr>
<td><strong>Randomised trial</strong></td>
<td>64/1417 (5%)</td>
<td>86/1432 (6%)</td>
</tr>
<tr>
<td><strong>Risk ratio† (95% CI)</strong></td>
<td>3.9 (2.1–7.0)</td>
<td>0.74 (0.53–1.03)</td>
</tr>
</tbody>
</table>

*Treatment recorded at baseline in the observational study and assigned at random in the randomised trial. †Univariate relative risk in the observational study and odds ratio in the randomised trial.

The disease of interest and controls that do not. Although it is unlikely that recall bias could account for the many-fold increases in risk seen, for example, with limb abnormalities and thalidomide use, it might well be responsible for more moderate differences in apparent risk. For example, an early case-control study of childhood cancer obtained data on maternal X-ray exposure through interviews with mothers, and observed that the risk of death from malignancy among the children of women who reported being exposed to abdominal X-rays was almost twice as great as that among the children of women who reported no such exposure. To determine whether this association might, at least in part, reflect more complete recall of exposure by the mothers of affected children, a second study was done in which exposure was determined from prenatal medical records. That study also found an increased risk of cancer among offspring of exposed women, but the relative risk was only half as large as in the first study. It has been suggested that such bias might be kept to a minimum by making comparisons between exposures reported by mothers of children with some particular birth defect and those reported by mothers of children with other anomalies. That strategy would not, however, exclude entirely the possibility of differential recall between the mothers of children with different types of birth defect. Moreover, it might obscure a real effect of the treatment if exposure was associated with more than one type of congenital anomaly.

### Bias due to differential detection of outcomes

Individuals receiving any treatment will tend to be seen by doctors or other health professionals more frequently than will others, and this may result in the earlier detection of a variety of outcomes. For example, although a highly significant increase of a quarter in the risk of breast cancer was seen among women taking hormonal contraceptives, this finding could largely reflect the earlier detection of less advanced breast cancer among such women. For, much of the observed excess risk was due to an excess of localised tumours, without any clear increase in the risk of tumours that had spread beyond the breast. Another possible example of such detection bias is provided by studies of first-trimester exposure to the antifungal drug itraconazole. Congenital malformations were seen in 13% of children of exposed women in a retrospective study compared with only 3% in a prospective study, perhaps reflecting the greater likelihood of including women who have affected babies in a retrospective study.

### Efforts to control biases in observational studies

The effects of biases in observational studies are frequently underestimated in the interpretation of

---

For personal use only. Reproduce with permission from The Lancet Publishing Group.
associations found between treatment and outcome. Even when statistical adjustment for measured confounding factors fails to reduce the size of such associations materially, this provides little reassurance that residual bias is not still a major cause of any observed associations. These difficulties are illustrated by an observational study of antihypertensive treatment in which a 60% higher risk of heart attacks was seen among patients receiving a calcium antagonist compared with those receiving other agents. In that study, calcium antagonists seem to have been preferentially prescribed to higher-risk patients (such as those with pre-existing coronary heart disease or other risk factors for cardiovascular disease), but the association between use of calcium antagonists and subsequent myocardial infarction remained conventionally significant after adjustment for measured confounders and after excluding those with a history of cardiovascular disease. Residual bias remains a plausible explanation for at least part of the observed excess risk, however, since the data collected on prognostic factors are unlikely to describe all of the factors that contributed to the tendency to prescribe calcium antagonists to higher-risk patients. This may explain why the largest risk reported in that observational study is not consistent with the much smaller difference in heart attack incidence (relative risk 1.12 [95% CI 1.00–1.26]) between calcium antagonists and diuretics or β-blockers in a prospectively-planned meta-analysis of all relevant randomised trials.49

Various statistical methods have been proposed to deal with the problem of residual biases in observational studies of treatment. For example, instrumental variable estimation involves grouping patients according to their likelihood of receiving the treatment of interest, by use of observable factors (ie, instrumental variables) that affect treatment use, but— it is hoped— do not directly affect patients’ outcomes.50 Although this method has been described as mimicking randomisation, it depends entirely on the untestable assumption that the observed instrumental variables are not correlated with unobserved factors that directly affect outcome. Moreover, since the range of variation between groups of patients in the likelihood of receiving some particular treatment might be narrow (eg, one such assessment of coronary-artery catheterisation was based on its use in 20% vs 26% of patients51), any difference in outcome due to this differential use of the treatment would probably be very small (and, hence, difficult to assess even in a properly randomised controlled trial).

Another method that has been proposed involves case-crossover (or case-series) analysis,52,53 in which outcomes are compared between periods before and after treatment exposure within the same individuals. But, although this may avoid biases resulting from differences between exposed and non-exposed patients, variations in the underlying disease state within individuals could still determine both the necessity for treatment and the likelihood of the outcome of interest occurring. For example, a case-crossover study reported a 60% higher risk of road-traffic accidents during periods of exposure to benzodiazepines.54 At least in part, this could have been due to exacerbation of certain conditions that led both to an increased use of benzodiazepines and, independently, to an increased risk of accidents. Hence, these and other non-randomised methods55 do not provide assurance that all sources of known and unknown bias are adequately controlled, and so cannot exclude the possibility that moderate biases have obscured or inflated any moderate effects, or have falsely indicated a treatment effect when none existed.

**Potential for small random errors in observational studies**

One advantage of observational studies is that it is often easier to study much larger numbers of patients—and, consequently, much larger numbers of deaths and other relevant outcomes— than it is in randomised trials. Observational studies can, therefore, provide estimates of treatment effects that are subject to relatively small random errors, allowing the reliable detection of some extreme though rare adverse effects of treatments. But, as discussed earlier, small random errors in large observational studies can also lead to the detection of more moderate differences in risk that are merely the result of bias, rather than the effect of treatment (ie, more precise, but biased, estimates). For example, in meta-analyses of observational studies of hormone replacement therapy, women who had taken such treatment were seen to have significantly lower risks of coronary heart disease with oestrogen alone (relative risk 0.70 [95% CI 0.65–0.75]) or with oestrogen plus progesterin (0.66 [0.53–0.84]), lower risks of colorectal cancer (0.8 [0.7–0.9]), higher risks of breast cancer (relative risk increasing by 2.3% [1.1–3.6] with each year of use),56 and higher risks of malignant melanoma (relative risk 1.4 [1.2–1.7]). But, there is evidence that women who take hormone replacement therapy may have better pretreatment coronary risk-factor profiles57 and better access to preventive health care58 than those who do not, and several of the risk factors for coronary heart disease that differ between users and non-users of hormone replacement therapy (such as physical inactivity and obesity) are also risk factors for colon cancer.59 On the other hand, women who take hormone replacement therapy (like those who take oral contraceptives) may be more likely to have breast cancer and melanoma diagnosed at an earlier stage because of greater contact with doctors. As a consequence, the balance of any true benefits and risks of hormone replacement therapy cannot be determined reliably from observational studies. In this regard, it is of interest that a relatively small randomised placebo-controlled trial of hormone replacement therapy for the secondary prevention of coronary heart disease60 failed to confirm the one-third reduction in risk suggested by the observational studies.61 Indeed, during about 4 years of follow-up in that study, there seemed to be an early excess of vascular events followed by a later shortfall among those assigned active treatment (and a similar early trend has recently been reported from the larger Women’s Health Initiative randomised trial62). These findings do not, however, preclude the possible emergence of worthwhile effects with more prolonged use of hormone replacement therapy in the large, long-term trials that are currently in progress.63

**Evidence from observational studies in the context of results from randomised trials**

A more prominent role for observational studies in the assessment of treatment effects has been argued in two reviews56,57 on the basis of examples in which there were considered to be no apparent differences between the results of observational studies and those of randomised trials. But, several of the examples included in those reviews involved estimates of treatment effects that were subject to large random errors. For example, separate meta-analyses of the observational studies and of the
randomised trials comparing laparoscopic and open appendectomy were interpreted as having shown similar reductions in infection rates with laparoscopic procedures, even though the 95% CI for the risk reduction in each type of study ranged from about 10% to about 70%. It has also been noted that other examples in those reviews did not even involve observational studies of treatments, but instead misleadingly compared the effects found in randomised trials of treatment that alter risk factors (such as lowering blood cholesterol or blood pressure) with estimates from observational studies of the associations between risk-factor levels and disease risk. Most trials of treatment that alter risk factors (such as misleadingly compared the effects found in randomised trials of treatment that alter risk factors (such as lowering blood cholesterol or blood pressure) with estimates from observational studies of the associations between risk-factor levels and disease risk. Most pertinently, any similarity of the treatment effects estimated from observational studies and from randomised trials in any one particular circumstance provides little reassurance that observational studies will provide unbiased estimates of the effects of treatment in some other circumstance.

Furthermore, it makes little sense to continue to base inference on observational studies when their results have been reliably refuted by large-scale randomised trials. For example, it had previously been suggested by observational analyses that digoxin might increase mortality substantially. By contrast, the large randomised Digitalis Investigation Group (DIG) trial showed unequivocally that the addition of digoxin to current therapy reduces the risk of hospital admission for heart failure (relative risk 0·72 [95% CI 0·66–0·79]) by about as much as do angiotensin-converting-enzyme inhibitors, with no apparent adverse effect on total mortality (1181 deaths in patients allocated digoxin and 1194 in patients allocated placebo; 0·99 [0·91–1·07]). In the light of this evidence, the prominent reporting of new claims from observational analyses that digoxin doubles the risk of death in just the sort of patients studied in the DIG trial is not appropriate.

Without clear confirmatory evidence from large-scale randomised trials or their meta-analyses, reports of modest treatment effects from observational studies should not be interpreted as providing good evidence of either adverse or protective effects of these agents (and, contrary to other suggestions, the absence of evidence from randomised trials does not in itself provide sufficient justification for relying on observational data). In this regard, it is salutary to note the example provided by early reports from observational studies of moderately increased risks of breast cancer among hypertensive patients treated with reserpine. Those reports led to avoidance of one of the few effective antihypertensive agents available at that time, and only much later was this association with breast cancer shown to be the likely result of bias. A more recent report from an observational study has suggested a moderately high risk of cancer among hypertensive patients treated with a calcium antagonist. But, once again, only limited data are currently available from randomised trials to assess the reliability of this observation.

**Use of observational studies to estimate potential effects of treatment**

**Prediction of the relative effects of treatment**

When a treatment alters an established risk factor for disease, observational studies of the association between that risk factor and the disease may provide some indication of the potential effects of the treatment on disease risk. For example, in observational studies, a prolonged 5 mm Hg lower diastolic blood pressure is associated with a one-third lower risk of stroke among middle-aged individuals, and randomised trials of blood-pressure lowering show that much, or all, of this predicted long-term effect is achieved within 5 years (with similar relative treatment effects in a variety of subgroups of patients; figure). However, although such estimates from observational studies of the potential effects of treatments may be valuable, they could overestimate the actual effects of treatment if disease risk is only partly reversed (at least in the short term). For example, observational studies have shown that a prolonged 5 mm Hg lower diastolic blood pressure is associated with about a one-fifth lower risk of coronary heart disease among middle-aged individuals. By contrast, randomised trials of treatments that reduce blood pressure or cholesterol suggest that only about a half or two-thirds of the long-term effects predicted from the observational studies are produced within about 5 years of altering these risk factors.

Treatments might also have independent effects on disease risk that offset or augment the benefits of altering a particular risk factor. For example, the reduction in stroke risk of about a third in a recent randomised trial of an angiotensin-converting-enzyme inhibitor is about twice as great as would be predicted from observational studies for the achieved 3 mm Hg reduction in systolic blood pressure. Moreover, differences in outcome associated with some putative risk factors may not be causal. For example, many observational studies have found that greater consumption, and higher blood concentrations, of β-carotene are associated with lower incidence of cancer. The first reviews of these findings stressed that such associations might merely reflect some type of confounding, and emphasised the need for large-scale randomised trials of the effects of long-term β-carotene supplementation on cancer incidence. Even after more than 10 years of treatment in such trials, no clear evidence of benefit has emerged, suggesting that the inverse associations in observational studies were not causal.
and Russia,71 or among patients with pre-existing vascular disease rates that are very high (as in parts of eastern Europe much greater in populations with stroke and coronary populations. Hence, the absolute benefits of such treatment outcomes found in observational studies of specific patients (whether in individual trials or meta-analyses), sufficiently by the study of appropriately large numbers of conditions. By contrast, random allocation of treatment expected from most treatments for most common serious outcomes, which are generally all that can realistically be provided there is a causal relation with disease. But, due to the potential biases, be importantly misleading. Instead, greater efforts need to be made to remove or overcome any obstacles that inappropriately prevent the provision of reliable evidence from randomised trials of adequate size (as has been achieved for treatments of numerous vascular and neoplastic conditions).

It has also been argued that observational studies could provide more generalisable evidence about the effects of treatment because they involve populations of patients, or clinicians, that are more representative of particular practice settings than those involved in clinical trials.57,58 But, the inclusion of more representative participants does not prevent observational studies from producing biased estimates of any moderate treatment effects that might exist. Moreover, as has been discussed, careful consideration of the effects seen among the different types of patient included in randomised trials can often allow the results of clinical trials to be generalised widely. In particular, applying estimates derived from appropriately large randomised trials (or meta-analyses of trials) of the relative effects of a treatment on specific outcomes to the absolute risks of those outcomes observed in representative patient populations may well provide a broadly reliable guide to the balance of the absolute benefits and absolute risks conferred by the treatment in routine clinical practice.

In conclusion, observational studies and randomised trials provide complementary evidence about the effects of treatment on mortality and major morbidity. Wider appreciation of the different strengths and weaknesses of these two types of epidemiological study should increase the likelihood that the most reliable evidence available informs decisions about the treatments doctors use—and patients receive—for the management of a wide range of life-threatening conditions.

We thank Colin Baigent, Iain Chalmers, Richard Doll, Richard Peto, Anthony Rodgers, and Peter Sandercock for comments and suggestions on previous drafts of this review. This work was supported by an Eccles Award from the National Health and Medical Research Council of Australia, a program grant from The Medical Foundation of the University of Sydney, a British Heart Foundation personal chair, and direct support from the UK Medical Research Council and Imperial Cancer Research Fund.

References
The uses of error: The complexity of general practice

General practitioners encounter a large number and variety of health problems every day. In coping with these problems the general practitioner has to modify his or her role accordingly, from mastery inactivity to active diagnostic and therapeutic pursuit. This case is an example of complex role-modification leading to a tragic, though probably understandable, failure to diagnose breast cancer.

A 37-year-old patient used to visit my practice regularly for various minor problems. I was rather fond of her and admired her for her cheerful outlook of life: she was a young mother of a single teenage son, and it was no secret that she was less than happy with her elderly husband. She had consulted me for a lump in her right breast, and I managed to arrange for a consultation with a surgeon next day.

However, during that consultation as well as the next, we discussed her new perspective on life and the impossibility of integrating it with her previous one. It was only after a number of such highly emotional consultations that I dared to request permission to re-examine her breast. The examination disclosed a persistent lump in the right breast, and I managed to arrange for a consultation with a surgeon next day.

Three memories of that period prevail. It was a considerable shock to receive a warm welcome when I visited her in hospital after the operation, because I had been expecting remorseless bleeding. The outcome was as striking as it was predictable. A carcinoma was diagnosed, an operation performed, 8 months later metastases were discovered, and just over a year later she died.

Two memories of that period prevail. It was a considerable shock to receive a warm welcome when I visited her in hospital after the operation, because I had been expecting remorseless reproach of my delayed diagnosis. The other memory is of the complicated, yet highly rewarding, period of terminal home care including a still fundamentally puzzled ex-husband and a most supportive girlfriend.

This case exemplifies the complexity of general practice. A lump in the breast is a common presenting complaint and important life events are often discussed between patients and their family doctor. This example highlights the difficulty of simultaneously pursuing a clinical diagnosis while providing a sympathetic ear for a patient’s often unrelated difficulties, a duty that the medical profession must carry out on both accounts.

Professor Chris van Weel
Department of General Practice, University of Nijmegen, Netherlands

References

Bad reporting does not mean bad methods for randomised controlled trials: observational study of randomised controlled trials performed by the Radiation Therapy Oncology Group

Heloisa P Soares, Stephanie Daniels, Ambuj Kumar, Mike Clarke, Charles Scott, Suzanne Swann, Benjamin Djulbegovic

Abstract

Objective To determine whether poor reporting of methods in randomised controlled trials reflects on poor methods.

Design Observational study.

Setting Reports of randomised controlled trials conducted by the Radiation Therapy Oncology Group since its establishment in 1968.

Participants The Radiation Therapy Oncology Group.

Outcome measures Content of reports compared with the design features described in the protocols for all randomised controlled trials.

Results The methodological quality of 56 randomised controlled trials was better than reported. Adequate allocation concealment was achieved in all trials but reported in only 42% of papers. An intention to treat analysis was done in 83% of trials but reported in only 69% of papers. The sample size calculation was performed in 76% of the studies, but reported in only 16% of papers. End points were clearly defined and α and β errors were prespecified in 76% and 74% of the trials, respectively, but only reported in 10% of the papers. The one exception was the description of drop outs, where the frequency of reporting was similar to that contained in the original statistical files of the Radiation Therapy Oncology Group.

Conclusions The reporting of methodological aspects of randomised controlled trials does not necessarily reflect the conduct of the trial. Reviewing research protocols and contacting trialists for more information may improve quality assessment.

Introduction

Evaluation of the quality of published clinical research is central to informed decision making. Information on trial quality is particularly important during peer review or when results from individual studies are evaluated in systematic reviews or meta-analyses. The quality of research should always be considered when a report is used in decision making in health care. Poorly conducted and reported research seriously compromises the integrity of the research process, especially if biased results receive false credibility.

Many efforts have been made to improve the quality of studies and their related publications. The best example was the publication of the Consolidated Standards of Reporting of Trials (CONSORT) statement to improve the quality of trial reports. Such efforts to improve the quality of clinical research, however, imply that if certain design or methodological features are not reported then they were not done. Ideally, assessment of the quality of clinical research should not only address reporting but also the original design and intended plan for its conduct and analysis as specified in the trial’s research protocol. The importance of linking the final report of clinical trials with their original research protocols led some authors to argue that no randomised controlled trial should be conducted without publication of its research protocol. The reasons behind this are that critical comments may be encouraged leading to improvements in trial design, publication can be coupled with trial registration, the original protocol can be compared with what was subsequently done, and investigators can more easily appreciate what research is being conducted in their areas of interest. More importantly, publication of research protocols is one of the best ways to minimise bias by explicitly stating a priori hypotheses and methods without the prior knowledge of results. Many randomised controlled trials are preceded by the preparation of a written protocol, which describes the conduct of the trial more comprehensively than is possible in many journal articles, and making these protocols available would provide much useful additional information. We aimed to test the assumption that poor reporting reflects poor methods by comparing research protocols with the information published in the final reports of a set of randomised controlled trials.

Methods

We studied randomised controlled trials conducted by the Radiation Therapy Oncology Group. This is a national clinical cooperative group with a focus on the development of studies to improve survival and the quality of life of patients with cancer. It was established in 1968 and is publicly funded by the National Cancer Institute in the United States. The group consists of both clinical and laboratory investigators from over 260 institutions across the United States and Canada, and its membership includes nearly 90% of all comprehensive and clinical cancer centres designated by the National Cancer Institute. Before activation, the group’s research protocols must pass through a rigorous peer review process and be reviewed and approved through its own committee system and the National Cancer Institute. Development of a protocol consists of six phases (box).

Our analysis included data related to primary outcomes from all terminated phase III trials conducted by the Radiation Therapy Oncology Group since its establishment in 1968. We extracted data on methodological domains that have been acknowledged as vital.
for minimising bias in the conduct and analysis of randomised controlled trials. The effect of chance is usually minimised by appropriate planning of the trial’s size, through a statistical power analysis using estimates for the expected differences between the interventions and prespecified type I (α) and type II (β) error levels. To investigate systematic bias, we extracted data on the quality of the randomisation process (selection bias) and drop outs (attrition bias). Since the primary outcome was survival in most of the studies, we did not consider quality related to observer bias, such as the use of placebo or independent reading of outcomes (there were only three placebo controlled trials). We extracted data from all papers and protocols. The accuracy of this data was verified by the group’s statistical centre.

Results

Overall, there were 59 terminated phase III randomised controlled trials, three of which had not been published. We found 58 published papers for the remaining 56 protocols for use in our study. The figure summarises the results according to information from the papers, protocols, and the Radiation Therapy Oncology Group’s statistical office. This shows that the reporting of methods in the publications does not necessarily reflect the methodological quality of the associated protocols. For example, a priori sample size calculations were performed in 44 (76%) trials, but this information was given in only nine of the 58 published papers (16%). Although all trials had adequate allocation concealment (through central randomisation), this was reported in only 24 (41%) of the papers. From our initial data extraction, we found that 40 (69%) of these trials used an intention to treat analysis. This number was increased to 48 (83%) after verification by the Radiation Therapy Oncology Group. End points were clearly defined, and α and β errors were prespecified in 44 (76%) and 43 (74%) trials, respectively, but only reported in six (10%) of the papers. Interestingly, reporting of drop outs was meticulous; we found no difference in frequency (91%) between data presented in the papers and those in the original files.

Discussion

Poor reporting of randomised controlled trials may not indicate poor quality of the trials themselves. We are aware of two other studies that reported empirical assessments of this relation. One study evaluated the quality of 65 randomised controlled trials of breast cancer treatment. Data were extracted from publications related to these trials and the results compared with the information provided by the principal investigators. The study concluded that faulty reporting reflected faulty methods. Another study, however, concluded that even well designed and conducted trials may be badly reported. This conclusion was drawn indirectly from an assessment of three key indicators of quality: adequate allocation concealment, appropriate blinding, and use of intention to treat analysis. Unlike our study, neither of these studies reported a comparison of the quality of reporting with the methods specified in the original research protocols.

In general, the Radiation Therapy Oncology Group and, we predict, other cooperative oncology groups sponsored by the National Cancer Institute, have conducted research of good quality. Our study is the first formal investigation of this kind, and we believe, the first examination of the methodological quality of randomised controlled trials performed by a cooperative oncology group.

The relation between poor reporting and poor methods was raised in 1980 in a report on patient registration, randomisation, and the importance of avoiding bias in cooperative oncology trials. This may have helped the cooperative oncology groups to be especially aware of methodological issues relating to trials and before the start of modern research on methodological quality. Consequently, for cooperative oncology groups such as the Radiation Therapy Oncology Group, even if the published description of the methods of a randomised controlled trial is poor, the quality of the trial should not be assumed to be poor. It is important to note, however, that our findings are based on a select sample of trials, which may not be representative of randomised controlled trials. Further studies to confirm the generalisability of our findings are needed and would be useful.
Another important point relates to any assumptions that trials published before the 1996 CONSORT statement are more likely to be of poorer quality than those published after it. The CONSORT statement contains several methodological elements that should be followed to eliminate biased results. The intention of this statement was to improve the conduct, integrity, and reporting of randomised controlled trials. Our results show that studies conducted by the Radiation Therapy Oncology Group were of high quality even before publication of the CONSORT statement. It was the reports of the randomised controlled trials that showed deficiencies in their description of the methods used in the trials, not the trials themselves. Our findings indicate that although researchers in the Radiation Therapy Oncology Group were cognisant of key features in the design and conduct of good quality trials, they were less aware of the need to report these to a standard that would meet contemporary (CONSORT) requirements.

It is still appropriate to expect that the CONSORT statement will contribute to the conduct of higher quality randomised controlled trials in the future, since it incorporates and highlights many of the elements needed to perform a trial adequately. We agree with the call for all journals to adopt the policy of only publishing the report of a randomised controlled trial if it follows the CONSORT requirements. This is supported by empirical data that are now emerging about the usefulness of the CONSORT statement. For example, one study compared the quality of reports of trials before and after the CONSORT statement and found that the statement was associated with an improvement in the quality of reports. Further improvements in the quality of the conduct and reporting of clinical research would arise with the publication of research protocols.

Contributors: HPS and BD conceptualised the study, were involved in all aspects of the study, and wrote the first draft of the paper. SD and AR contributed to the study design, collection of data, analysis and interpretation of the data, and writing the report. MC contributed to the study design, interpretation of the data, and writing the report. CS and SS contributed to the collection of data and writing the report. BD will act as guarantor for the paper.

Funding: This research was supported by the Research Program on Research Integrity, an Office of Research Integrity/National Institute of Health collaboration, grant No 1R01NS144417-01.

Commentary: The quality of randomised controlled trials may be better than assumed

Auro del Giglio, Luciano Jose Costa

As readers of published articles, it is reassuring to know that the quality of published data is probably better than expected from the reporting of the methods. Soares and colleagues have addressed the discrepancies between the proposed methods in original research protocols and those reported in the final article for all 56 randomised controlled trials conducted by the Radiation Therapy Oncology Group since its creation in 1968. Good quality experimental designs were more often adhered to during the conduct of the studies than suggested by the final reports.

As readers of systematic reviews, however, we may have underestimated the quality of experimental methods, especially if reviewers had not checked the original protocols. This may have led to the exclusion of some potentially good quality papers from systematic reviews.
Another important point relates to any assumptions that trials published before the 1996 CONSORT statement are more likely to be of poorer quality than those published after it. The CONSORT statement contains several methodological elements that should be followed to eliminate biased results. The intention of this statement was to improve the conduct, integrity, and reporting of randomised controlled trials. Our results show that studies conducted by the Radiation Therapy Oncology Group were of high quality even before publication of the CONSORT statement. It was the reports of the randomised controlled trials that showed deficiencies in their description of the methods used in the trials, not the trials themselves. Our findings indicate that although researchers in the Radiation Therapy Oncology Group were cognisant of key features in the design and conduct of good quality trials, they were less aware of the need to report these to a standard that would meet contemporary (CONSORT) requirements.

It is still appropriate to expect that the CONSORT statement will contribute to the conduct of higher quality randomised controlled trials in the future, since it incorporates and highlights many of the elements needed to perform a trial adequately. We agree with the call for all journals to adopt the policy of publishing the report of a randomised controlled trial if it follows the CONSORT requirements. This is supported by empirical data that are now emerging about the usefulness of the CONSORT statement. For example, one study compared the quality of reports of trials before and after the CONSORT statement and found that the statement was associated with an improvement in the quality of reports. Further improvements in the quality of the conduct and reporting of clinical research would arise with the publication of research protocols.

Contributors: HPS and BD conceptualised the study, were involved in all aspects of the study, and wrote the first draft of the paper. SD and AR contributed to the study design, collection of data, analysis and interpretation of the data, and writing the report. MC contributed to the study design, interpretation of the data, and writing the report. CS and SS contributed to the collection of data and writing the report. BD will act as guarantor for the paper.

Funding: This research was supported by the Research Program on Research Integrity, an Office of Research Integrity/National Institute of Health collaboration, grant No 1R01NS/NR44417-01.

Commentary: The quality of randomised controlled trials may be better than assumed

Auro del Giglio, Luciano Jose Costa

As readers of published articles, it is reassuring to know that the quality of published data is probably better than expected from the reporting of the methods. Soares and colleagues have addressed the discrepancies between the proposed methods in original research protocols and those reported in the final article for all 56 randomised controlled trials conducted by the Radiation Therapy Oncology Group since its creation in 1968. Good quality experimental designs were more often adhered to during the conduct of the studies than suggested by the final reports.

As readers of systematic reviews, however, we may have underestimated the quality of experimental methods, especially if reviewers had not checked the original protocols. This may have led to the exclusion of some potentially good quality papers from systematic reviews.
Drug points

Rash and acute nephritic syndrome due to candesartan
Adam Morton, James Muir, Davin Lim

A 73 year old man presented with a two day history of a pruritic rash and oedema affecting both lower legs. He had had hypertension for three years, for which he had been taking candesartan, an angiotensin II receptor antagonist, for the past two years. He was not taking any other medications or preparations. He had no infective, gastrointestinal, or respiratory symptoms and no arthralgias.

Examination showed non-palpable, non-tender purpura affecting both legs, with pitting oedema, and urticarial lesions on the left knee and anterior chest. Testing of urine was positive for protein and blood. Urine microscopy showed >100 erythrocytes per 10³/l; the urine protein:creatinine ratio was 0.15 (normal range <0.04), serum creatinine was 90 μmol/l (70-120 μmol/l), C reactive protein was 117 mg/l (0-6 mg/l), and the erythrocyte sedimentation rate was 32 mm/hour (1-20 mm/hour). A provisional diagnosis of Schönlein-Henoch purpura due to candesartan was made, and the medication was stopped. Serum immunoglobulin A concentration, however, was normal. The results of an autoantibody screen and testing for antineutrophil cytoplasmic antibodies were negative, complements were normal, and cryoglobulins were not detected. A skin biopsy showed lymphocytic vasculitis involving vessels in the papillary and mid-dermis, as well as spongiosis and moderate orthokeratosis. We believed the features to be consistent with a drug reaction. The rash and microscopic haematuria resolved completely within a week of stopping candesartan and C reactive protein concentration became normal three weeks after presentation, though proteinuria took 10 weeks to resolve.

Inbesartan, another angiotensin II receptor antagonist, was the highest volume drug prescribed for hypertension in Australia on the pharmaceutical benefits scheme for the year ending December 2002. A major reason for the popularity of this class of drugs is their side effect profile, shown to be similar to that of placebo in double blind studies. Major side effects published in the literature, however, include hepatotoxicity, pancreatitis, angio-oedema, acute deterioration in renal function, and dysgeusia. Two cases of Schönlein-Henoch purpura have been described as being associated with therapy with losartan.1,2 Both had a similar presentation to the case described above, with purpuric rash, pedal oedema, microscopic haematuria, proteinuria, raised C reactive protein concentration, and rapid resolution when the drug was stopped. Both cases, however, were associated with raised serum IgA concentration and deposits in the dermal vessel walls on histology.

From 1999 until November 2002, seven cases of rash were reported to the Adverse Drug Reaction Advisory Committee in Australia, in which candesartan was the sole possible agent responsible. No cases of nephritis have been reported to the committee to date. Similarly, the manufacturer of candesartan (AstraZeneca) has not received any reported cases of nephritis with candesartan. In conclusion, the angiotensin II receptor antagonists are a very well tolerated group of antihypertensive drugs, but they should be considered to be a potential cause of rash or acute nephritic syndrome in any patients presenting with those symptoms, regardless of how long they have been taking the drug.

Funding: None.

Competing interests: None declared.


Endpiece

Students at Guy’s and St Thomas’s, 1819

I must be permitted to caution you against blindly adopting any system of opinion and practice which may be taught in the schools to which you may be . . . attached; this could be to degrade you to the ranks of empiricism. Think for yourselves.

Aesculapius. The hospital pupil’s guide at St Thomas’s and Guy’s Hospitals. Lond Med Repository 1819;11:128-30

Jeremy Hugh Baron, honorary professorial lecturer, Mount Sinai School of Medicine, New York

Contributors: AdG was the main author of the commentary and wrote the final manuscript. He will act as guarantor. LJC made contributions to the manuscript and contributed to the final manuscript.

Funding: None.

Competing interests: None declared.

Sources of Variation and Bias in Studies of Diagnostic Accuracy
A Systematic Review
Penny Whiting, MSc; Anne W.S. Rutjes, MSc; Johannes B. Reitsma, MD, PhD; Afina S. Glas, MD, PhD; Patrick M.M. Bossuyt, PhD; and Jos Kleijnen, MD, PhD

Background: Studies of diagnostic accuracy are subject to different sources of bias and variation than studies that evaluate the effectiveness of an intervention. Little is known about the effects of these sources of bias and variation.

Purpose: To summarize the evidence on factors that can lead to bias or variation in the results of diagnostic accuracy studies.

Data Sources: MEDLINE, EMBASE, and BIOSIS, and the methodologic databases of the Centre for Reviews and Dissemination and the Cochrane Collaboration. Methodologic experts in diagnostic tests were contacted.

Study Selection: Studies that investigated the effects of bias and variation on measures of test performance were eligible for inclusion, which was assessed by one reviewer and checked by a second reviewer. Discrepancies were resolved through discussion.

Data Extraction: Data extraction was conducted by one reviewer and checked by a second reviewer.

Data Synthesis: The best-documented effects of bias and variation were found for demographic features, disease prevalence and severity, partial verification bias, clinical review bias, and observer and instrument variation. For other sources, such as distorted selection of participants, absent or inappropriate reference standard, differential verification bias, and review bias, the amount of evidence was limited. Evidence was lacking for other features, including incorporation bias, treatment paradox, arbitrary choice of threshold value, and dropouts.

Conclusions: Many issues in the design and conduct of diagnostic accuracy studies can lead to bias or variation; however, the empirical evidence about the size and effect of these issues is limited.
conducted a systematic review of all studies in which the main focus was examine the effects of one or more sources of bias or variation on estimates of test performance.

Methods

Literature Searches

We searched MEDLINE, EMBASE, BIOSIS and the methodologic databases of the Centre for Reviews and Dissemination and the Cochrane Collaboration from database inception to 2001. Search terms included sensitivity*, mass-screening, diagnostic-test, laboratory-diagnosis, false positive*, false negative*, specificity*, screening, accuracy, predictive value*, reference value*, likelihood ratio*, sroc, and receiver op-

Table 1. Description of Sources of Bias and Variation

<table>
<thead>
<tr>
<th>Source Population</th>
<th>Bias or Variation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic features</td>
<td>Variation</td>
<td>Tests may perform differently in various samples. Therefore, demographic features may lead to variations in estimates of test performance.</td>
</tr>
<tr>
<td>Disease severity</td>
<td>Variation</td>
<td>Differences in disease severity among studies may lead to differences in estimates of test performance.</td>
</tr>
<tr>
<td>Disease prevalence</td>
<td>Variation</td>
<td>The prevalence of the target condition varies according to setting and may affect estimates of test performance. Context bias, the tendency of interpreters to consider test results to be positive more frequently in settings with higher disease prevalence, may also affect estimates of test performance.</td>
</tr>
<tr>
<td>Distorted selection of participants</td>
<td>Variation</td>
<td>The selection process determines the composition of the study sample. If the selection process does not aim to include a patient spectrum similar to the population in which the test will be used in practice, the results of the study may have limited applicability.</td>
</tr>
<tr>
<td>Test protocol: materials and methods</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test execution</td>
<td>Variation</td>
<td>A sufficient description of the execution of index and reference standards is important because variation in measures of diagnostic accuracy can be the result of differences in test execution.</td>
</tr>
<tr>
<td>Test technology</td>
<td>Variation</td>
<td>When the characteristics of a diagnostic test change over time as a result of technological improvement or the experience of the operator of the test, estimates of test performance may be affected.</td>
</tr>
<tr>
<td>Treatment paradox and disease progression bias</td>
<td>Bias</td>
<td>Disease progression bias occurs when the index test is performed an unusually long time before the reference standard, so the disease is at a more advanced stage when the reference standard is performed. Treatment paradox occurs when treatment is started on the basis of the knowledge of the results of the index test, and the reference standard is applied after treatment has started.</td>
</tr>
<tr>
<td>Reference standard and verification procedure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inappropriate reference standard</td>
<td>Bias</td>
<td>Errors of imperfect reference standard or standards bias the measurement of diagnostic accuracy of the index test.</td>
</tr>
<tr>
<td>Differential verification bias</td>
<td>Bias</td>
<td>Part of the index test results is verified by a different reference standard.</td>
</tr>
<tr>
<td>Partial verification bias</td>
<td>Bias</td>
<td>Only a selected sample of patients who underwent the index test is verified by the reference standard.</td>
</tr>
<tr>
<td>Interpretation (reading process)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Review bias</td>
<td>Bias</td>
<td>Interpretation of the index test or reference standard is influenced by knowledge of the results of the other test. Diagnostic review bias occurs when the results of the index test are known when the reference standard is interpreted. Test review bias occurs when results of the reference standard are known while the index test is interpreted.</td>
</tr>
<tr>
<td>Clinical review bias</td>
<td>Bias</td>
<td>The availability of information on clinical data, such as age, sex, and symptoms, during interpretation of test results may affect estimates of test performance.</td>
</tr>
<tr>
<td>Incorporation bias</td>
<td>Bias</td>
<td>The result of the index test is used to establish the final diagnosis.</td>
</tr>
<tr>
<td>Observer variability</td>
<td>Variation</td>
<td>The reproducibility of test results is one of the determinants of diagnostic accuracy of an index test. Because of variation in laboratory procedures or observers, a test may not consistently yield the same result when repeated. In 2 or more observations of the same diagnostic study, intraobserver variability occurs when the same person obtains different results, and interobserver variability occurs when 2 or more people disagree.</td>
</tr>
<tr>
<td>Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Handling of indeterminate results</td>
<td>Bias</td>
<td>A diagnostic test can produce an uninterpretable result with varying frequency depending on the test. These problems are often not reported in test efficacy studies; the uninterpretable results are simply removed from the analysis. This may lead to biased assessment of the test characteristics.</td>
</tr>
<tr>
<td>Arbitrary choice of threshold value</td>
<td>Variation</td>
<td>The selection of the threshold value for the index test that maximizes the sensitivity and specificity of the test may lead to overoptimistic measures of test performance. The performance of this cutoff in an independent set of patients may not be the same as in the original study.</td>
</tr>
</tbody>
</table>

erat* characteristic*. We also identified papers that had cited the key papers. Complete details of the search strategy are provided elsewhere (3). We contacted methodologic experts and groups conducting work in this field. Reference lists of retrieved articles were screened for additional studies.

Inclusion Criteria

All studies with the main objective of addressing bias or variation in the results of diagnostic accuracy studies were eligible for inclusion. Studies of any design, including reviews, and any topic area were eligible. Studies had to investigate the effects of bias or variation on measures of test performance, such as sensitivity, specificity, predictive
values, likelihood ratios, and diagnostic odds ratios, and indicate how a particular feature may distort these measures. Inclusion was assessed by one reviewer and checked by a second reviewer; discrepancies were resolved through discussion.

Data Extraction

One reviewer extracted data and a second reviewer checked data on the following parameters: study design, objective, sources of bias or variation investigated, and the results for each source. Discrepancies were resolved by consensus or consultation with a third reviewer.

Data Synthesis

We divided the different sources of bias and variation into groups (Table 1). Table 1 provides a brief description of each source of bias and variation; more detailed descriptions are available elsewhere (3). Results were stratified according to the source of bias or variation investigated. We classified studies that used actual data from one or more clinical studies to demonstrate the effect of a particular study feature as experimental studies, diagnostic accuracy studies, or systematic reviews. Experimental studies were defined as studies specifically designed to test a hypothesis about the effect of a particular feature. Discrepancies were resolved by consensus or consultation with a third reviewer.

Role of the Funding Source

The funding source was not involved in the design, conduct, or reporting of the study or in the decision to submit the manuscript for publication.

Data Synthesis

The literature searches identified a total of 8663 references. Of these, 569 studies were considered potentially relevant and were assessed for inclusion; 55, published from 1963 to 2000, met inclusion criteria. Nine studies were systematic reviews, 16 studies used an experimental design, 22 studies were diagnostic accuracy studies, and 8 studies used modeling to investigate the theoretical effects of bias or variation.

Population

Demographic Features

Ten studies assessed the effects of demographic features on test performance (Table 2) (4, 5, 7, 9, 11, 14, 15, 20, 22, 24). Eight studies were diagnostic accuracy studies, and 2 were systematic reviews. All but one study (22) found an association between the features investigated and overall accuracy. The study that did not find an association investigated whether estimates of exercise testing performance differed between men and women; after correction for the effects of verification bias, no significant differences were found (22).

In general, the studies found associations between the demographic factors investigated and sensitivity; the reported effect on specificity was less strong. Four studies found that various factors, including sex, were associated with sensitivity but showed no association with specificity (4, 5, 11, 20). The index tests investigated in these studies were exercise testing (5, 11, 20) to diagnose heart disease and body mass index to test for obesity (4). Two additional studies of exercise testing also reported an association with sensitivity, but the effects on specificity differed. One found that factors that lead to increased sensitivity also lead to a decrease in specificity (14); the second reported higher sensitivity and specificity in men than in women (16). A study of the diagnostic accuracy of an alcohol screening questionnaire found that overall accuracy was increased in certain ethnic groups (24). Sex was the most commonly investigated variable. Three studies found no association between test performance and sex. 9 found significant effects on sensitivity, and 4 found significant effects on specificity. Other variables shown to have significant effects on test performance were age, race, and smoking status.

Disease Severity

Six studies looked at the effects of disease severity on test performance (Table 2) (5, 11, 14, 19, 23, 25). Three studies were diagnostic accuracy studies, 2 were reviews, and one used modeling to investigate the effects of differences in disease severity. The modeling study also included an example from a diagnostic accuracy study of tests for the diagnosis of ovarian cancer (25). Three studies investigated tests for heart disease (5, 11, 14), one examined ventilation-perfusion lung scans for diagnosing pulmonary embolism (23), and one investigated 2 different laboratory tests (one for cancer and the other for bacterial infections) (19). All 6 studies found increased sensitivity with more severe disease; 5 found no effect on specificity (5, 11, 14, 19, 23), and one did not comment on the effects on specificity (25).

Disease Prevalence

Six studies looked at the effects of increased disease prevalence on test performance (Table 2) (8, 10, 13, 17, 21, 26). One study used an experimental design (8); the other studies were all diagnostic accuracy studies. The tests investigated in these studies covered a wide range of topics: dipstick for diagnosing urinary tract infection (10), magnetic resonance imaging and evoked potentials for diagnosing multiple sclerosis (17), exercise testing for diagnosing coronary artery disease (21), lung scans for diagnosing pulmonary embolism (8), clinical indications for diagnosing pneumonia (13), and ultrasonography for diagnosing epididymitis (26). Only 5 of the studies reported on the effects of disease prevalence on sensitivity; all found an in-
crease in sensitivity with increased disease prevalence (8, 10, 13, 17, 26). These studies also investigated the effects of increased disease prevalence on specificity and found mixed results; 2 found that specificity decreased (10, 13), 2 found no effect (8, 17), and one reported increased specificity (26). The remaining study looked only at the effects of disease prevalence on specificity, which was found to decrease (21).

**Distorted Selection of Participants**

Four studies examined the effects of distorted selection of participants on test performance (Table 2) (5, 12, 18, 27). A diagnostic accuracy study of exercise testing for heart disease found that overall accuracy was overestimated if reasons for exclusion commonly used by researchers were applied (18). The other 3 studies were reviews. The first, a review of the clinical and radiologic diagnosis of caries, found that in vivo studies gave higher estimates of test performance than in vitro studies (27). A review of exercise testing for heart disease found that avoiding a limited challenge group (that is, including patients with other confounding diseases or patients taking medications thought to produce false-positive results) did not have significant

<table>
<thead>
<tr>
<th>Study, Year (Reference)</th>
<th>Design</th>
<th>Index Test</th>
<th>Study Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curtin et al., 1997 (4)</td>
<td>Diagnostic accuracy</td>
<td>Body mass index</td>
<td>226 white persons</td>
</tr>
<tr>
<td>Detrano et al., 1988 (5)</td>
<td>Review</td>
<td>Exercise thallium scintigraphy</td>
<td>56 primary studies</td>
</tr>
<tr>
<td>Detrano et al., 1988 (6)</td>
<td>Review</td>
<td>Exercise electrocardiography</td>
<td>60 primary studies</td>
</tr>
<tr>
<td>Egglin and Feinstein, 1996 (8)</td>
<td>Experimental</td>
<td>Pulmonary arteriography</td>
<td>24 arteriograms</td>
</tr>
<tr>
<td>Hlatky et al., 1984 (9)</td>
<td>Diagnostic accuracy</td>
<td>Exercise electrocardiography</td>
<td>2269 patients</td>
</tr>
<tr>
<td>Lachs et al., 1992 (10)</td>
<td>Diagnostic accuracy</td>
<td>Dipsticks</td>
<td>366 consecutive patients</td>
</tr>
<tr>
<td>Levy et al., 1990 (11)</td>
<td>Diagnostic accuracy</td>
<td>Electrocardiography</td>
<td>4684 patients with suspected left ventricular hypotrophy</td>
</tr>
<tr>
<td>Lijmer et al., 1999 (12)</td>
<td>Review</td>
<td>Various tests</td>
<td>184 primary studies of 218 tests</td>
</tr>
<tr>
<td>Melbye and Straume, 1993 (13)</td>
<td>Diagnostic accuracy</td>
<td>Clinical cues</td>
<td>581 patients with suspected pneumonia</td>
</tr>
<tr>
<td>Moons et al., 1997 (14)</td>
<td>Diagnostic accuracy</td>
<td>Exercise test</td>
<td>295 consecutive patients with heart pain</td>
</tr>
<tr>
<td>Morise and Diamond, 1994 and 1995 (15, 16)</td>
<td>Diagnostic accuracy</td>
<td>Exercise electrocardiography</td>
<td>4467 patients with suspected coronary disease</td>
</tr>
<tr>
<td>O’Connor et al., 1996 (17)</td>
<td>Diagnostic accuracy</td>
<td>Magnetic resonance imaging and evoked potentials</td>
<td>303 patients with suspected multiple sclerosis</td>
</tr>
<tr>
<td>Philbrick et al., 1982 (18)</td>
<td>Diagnostic accuracy</td>
<td>Graded exercise test</td>
<td>208 consecutive patients evaluated for coronary arterial disease</td>
</tr>
<tr>
<td>Ransohoff and Feinstein, 1978 (19)</td>
<td>Review</td>
<td>Carcinoembryonic antigen and nitroblue tetrazolium tests</td>
<td>17 studies of carcinoembryonic antigen and 16 of nitroblue tetrazolium</td>
</tr>
<tr>
<td>Roger et al., 1997 (20)</td>
<td>Diagnostic accuracy</td>
<td>Exercise echocardiography</td>
<td>3679 consecutive patients</td>
</tr>
<tr>
<td>Rozanski et al., 1983 (21)</td>
<td>Diagnostic accuracy</td>
<td>Exercise radionuclide ventriculography</td>
<td>77 angiographically normal patients</td>
</tr>
<tr>
<td>Santana-Boado et al., 1998 (22)</td>
<td>Diagnostic accuracy</td>
<td>Single-photon emission computed tomography</td>
<td>702 consecutive patients evaluated for coronary disease</td>
</tr>
<tr>
<td>Stein et al., 1993 (23)</td>
<td>Diagnostic accuracy</td>
<td>Ventilation/perfusion scan</td>
<td>1050 patients</td>
</tr>
<tr>
<td>Steinbauer et al., 1998 (24)</td>
<td>Diagnostic accuracy</td>
<td>Screening tests for alcohol abuse</td>
<td>1333 adult family practice patients</td>
</tr>
<tr>
<td>van der Schouw et al., 1995 (26)</td>
<td>Diagnostic accuracy</td>
<td>Ultrasonography</td>
<td>483 consecutive patients; 372 included</td>
</tr>
<tr>
<td>Van Rijkom et al., 1995 (27)</td>
<td>Review</td>
<td>Tests for approximal caries</td>
<td>39 sets of sensitivity and specificity data</td>
</tr>
</tbody>
</table>

* NA = not applicable; † = increased; ‡ = decreased.
The final study, which reviewed many different tests, found that case–control studies overestimate overall accuracy; it also found that nonconsecutive patient enrollment and a retrospective study design did not affect the diagnostic odds ratio (12). This review also looked at the effects of failure to provide an appropriate description of the patient sample and found that this was associated with increased overall accuracy.

### Table 2—Continued

<table>
<thead>
<tr>
<th>Source of Bias or Variation</th>
<th>Factors Investigated</th>
<th>Effect on Sensitivity</th>
<th>Effect on Specificity</th>
<th>Effect on Overall Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic features</td>
<td>Increased weight; sex (female)</td>
<td>↑</td>
<td>None</td>
<td>NA</td>
</tr>
<tr>
<td>Demographic features</td>
<td>Sex, age, and medication use</td>
<td>Associated</td>
<td>None</td>
<td>NA</td>
</tr>
<tr>
<td>Distorted selection of participants</td>
<td>Inclusion of patients with previous myocardial infarction</td>
<td>↑</td>
<td>None</td>
<td>NA</td>
</tr>
<tr>
<td>Disease severity</td>
<td>Various patient-related characteristics (all are not associated)</td>
<td>Associated</td>
<td>Associated</td>
<td>NA</td>
</tr>
<tr>
<td>Disease prevalence</td>
<td>Context of interpretation: effect of increased disease prevalence</td>
<td>↑</td>
<td>None</td>
<td>NA</td>
</tr>
<tr>
<td>Demographic features</td>
<td>Exercise heart rate, number of diseased arteries, type of angina, age, and sex</td>
<td>Associated</td>
<td>Associated</td>
<td>NA</td>
</tr>
<tr>
<td>Disease prevalence</td>
<td>High pretest probability of disease</td>
<td>↑</td>
<td>↓</td>
<td>NA</td>
</tr>
<tr>
<td>Demographic features</td>
<td>Sex (male), increased age, decreased body mass index, not smoking</td>
<td>↑</td>
<td>None</td>
<td>NA</td>
</tr>
<tr>
<td>Disease severity</td>
<td>Increased severity of left ventricular hypertrophy</td>
<td>↑</td>
<td>None</td>
<td>NA</td>
</tr>
<tr>
<td>Distorted selection of participants</td>
<td>Diagnostic case–control studies</td>
<td>NA</td>
<td>NA</td>
<td>↑</td>
</tr>
<tr>
<td>Disease prevalence</td>
<td>Increased prevalence</td>
<td>↑</td>
<td>↓</td>
<td>NA</td>
</tr>
<tr>
<td>Demographic features</td>
<td>Sex, workload, diabetes, smoking, cholesterol level (all are not associated)</td>
<td>↑</td>
<td>↓</td>
<td>NA</td>
</tr>
<tr>
<td>Disease severity</td>
<td>Number of diseased vessels</td>
<td>↑</td>
<td>None</td>
<td>NA</td>
</tr>
<tr>
<td>Demographic features</td>
<td>Men</td>
<td>↑</td>
<td>↑</td>
<td>NA</td>
</tr>
<tr>
<td>Disease prevalence</td>
<td>Increased prevalence</td>
<td>↑</td>
<td>None</td>
<td>NA</td>
</tr>
<tr>
<td>Distorted selection of participants</td>
<td>Exclusion of patients with other clinical conditions</td>
<td>NA</td>
<td>NA</td>
<td>↑</td>
</tr>
<tr>
<td>Disease severity</td>
<td>Extensive disease</td>
<td>↑</td>
<td>None</td>
<td>NA</td>
</tr>
<tr>
<td>Demographic features</td>
<td>Sex (male)</td>
<td>↑</td>
<td>None</td>
<td>NA</td>
</tr>
<tr>
<td>Disease prevalence</td>
<td>Increased prevalence</td>
<td>↑</td>
<td>↓</td>
<td>NA</td>
</tr>
<tr>
<td>Demographic features</td>
<td>Sex</td>
<td>None</td>
<td>None</td>
<td>NA</td>
</tr>
<tr>
<td>Disease severity</td>
<td>Previous pulmonary disease</td>
<td>↑</td>
<td>None</td>
<td>NA</td>
</tr>
<tr>
<td>Demographic features</td>
<td>Race and sex</td>
<td>NA</td>
<td>NA</td>
<td>Associated</td>
</tr>
<tr>
<td>Disease severity</td>
<td>Clear cases of malignant disease</td>
<td>↑</td>
<td>Not reported</td>
<td>NA</td>
</tr>
<tr>
<td>Disease prevalence</td>
<td>Increased prevalence (inclusion criteria widened)</td>
<td>↑</td>
<td>↑</td>
<td>NA</td>
</tr>
<tr>
<td>Distorted selection of participants</td>
<td>In vivo studies compared with in vitro studies</td>
<td>NA</td>
<td>NA</td>
<td>↑</td>
</tr>
</tbody>
</table>

Effects on overall accuracy (5). The final study, which reviewed many different tests, found that case–control studies overestimate overall accuracy; it also found that nonconsecutive patient enrollment and a retrospective study design did not affect the diagnostic odds ratio (12). This review also looked at the effects of failure to provide an appropriate description of the patient sample and found that this was associated with increased overall accuracy.

**Test Protocol: Materials and Methods**

**Test Execution**

We found only 2 studies, both reviews, that specifically looked at the effects of differences in test execution (Table 3) (6, 12). The first, a review of several different tests, found that failure to describe the index test and reference standard execution leads to an overestimation of overall accuracy (12). The other found no effect of differences in protocol on overall accuracy in exercise testing (6).

**Test Technology**

Two studies looked at the effects of a change in the technology of the index test on test performance (Table 3) (6, 28). A systematic review of exercise scintigraphy studies found that automation of the test procedure improved sensitivity but decreased specificity (6). The other study, a
Bias and Variation in Diagnostic Studies

Evidence of bias (6).

Performance was found. This study, a review of exercise scintigraphy for the diagnosis of heart disease found that studies that used a specific test, found no effect on test performance (28).

**Treatment Paradox and Disease Progression Bias**

No studies that provided evidence of the effect of treatment paradox were identified. Only one study that looked at the effects of disease progression bias on test performance was found. This study, a review of exercise scintigraphy for the diagnosis of heart disease, found no evidence of bias (6).

**Reference Standard and Verification Procedure**

**Inappropriate Reference Standard**

Eight studies looked at reference standard error bias (Table 4) (6, 7, 27, 29, 31, 34, 41, 43). Four were systematic reviews, and the other 4 used modeling to investigate the theoretical effects of an imperfect reference standard. The reviews looked at reference standard error bias from slightly different perspectives, but all found evidence of bias. A review of patients who received a diagnosis of caries found that weaker validation methods may overestimate overall accuracy (27). A review of a hormone test for the diagnosis of heart disease, found that comparison with a more accurate test leads to increased sensitivity but did not report on the effect on accuracy (7).

The studies that used modeling to investigate the effects of an imperfect reference standard also found evidence of bias. One study suggested that with imperfect reference standards, specificity is most accurately estimated at low disease prevalence and sensitivity at high disease prevalence; it also suggested that considerable errors in estimates exist, even when the reference standard has close to perfect performance (31). Two studies found that inaccurate reference standards lead to underestimation of index test accuracy when the index test errors are statistically independent of the reference standard and overestimation when the index test errors are statistically dependent on the reference standard (41, 43). The final study found that overall accuracy is underestimated when the test being evaluated is more accurate than the reference standard (34, 43).

**Differential Verification Bias**

Only 2 studies looked at differential verification bias (Table 4) (12, 30). One was a review of several different tests (12), and the other was a diagnostic accuracy study of the clinical diagnosis of Alzheimer disease (30). Both found that differential verification bias leads to higher (inflated) measures of overall accuracy.

**Partial Verification Bias**

Twenty studies investigated the effects of partial verification bias (Table 4) (5, 7, 12, 16, 18–22, 28, 30, 32, 35–40, 42, 44). Two studies used models to investigate the theoretical effects of verification bias and found that partial verification bias increased sensitivity and decreased specificity (35, 36). A third study also used modeling to investigate the effects of verification bias; in addition, it provided an example from a diagnostic accuracy study. This study reported an association between overall accuracy and the presence of partial verification bias (44).

All of the remaining studies used actual data to investigate the effects of partial verification bias and were either diagnostic accuracy studies or reviews. Most of these studies examined some form of exercise testing for the diagnosis of heart disease (5, 6, 16, 18, 20, 21, 28, 32, 38). Other tests that were investigated included noninvasive tests for arterial disease (37), clinical diagnosis for Alzheimer disease (30), clinical findings for diagnosing hemorrhage in patients who had strokes (40), nuchal translucency for diagnosing Down syndrome (39), the carcinoembryonic antigen and nitro-blue tests (19), and serum ferritin levels for diagnosing hereditary hemochromatosis (42). Seven studies

---

**Table 3. Test Protocol: Materials and Methods**

<table>
<thead>
<tr>
<th>Study Details</th>
<th>Design</th>
<th>Index Test</th>
<th>Study Sample</th>
<th>Source of Bias or Variation</th>
<th>Factors Investigated</th>
<th>Effect on Sensitivity</th>
<th>Effect on Specificity</th>
<th>Effect on Overall Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detrano et al., 1988 (6)</td>
<td>Review</td>
<td>Exercise electrocardiography</td>
<td>60 primary studies</td>
<td>Test execution; Test technology; Disease progression bias</td>
<td>Exercise protocol; Automation of test; Maximum interval between scintigraphy and angiography</td>
<td>None</td>
<td>None</td>
<td>NA</td>
</tr>
<tr>
<td>Froelicher et al., 1998 (28)</td>
<td>Diagnostic accuracy</td>
<td>Electrocardiography and angiographic calipers</td>
<td>814 consecutive patients with angina pectoris</td>
<td>Test technology</td>
<td>Computerized readings</td>
<td>None</td>
<td>None</td>
<td>NA</td>
</tr>
<tr>
<td>Lijmer et al., 1999 (12)</td>
<td>Review</td>
<td>Various tests</td>
<td>184 primary studies of 218 tests</td>
<td>Test execution</td>
<td>Failure to describe index test execution; Failure to describe reference standard execution</td>
<td>NA</td>
<td>NA</td>
<td>↑</td>
</tr>
</tbody>
</table>

* NA = not applicable; ↑ = increased; ↓ = decreased.
found that sensitivity was increased and specificity decreased in the presence of partial verification bias (16, 18, 20, 28, 32, 38, 40); one study found that both sensitivity and specificity were increased (39), and 2 studies found that sensitivity was increased but did not report the effects on specificity (19, 42). One study found that specificity was increased in the presence of verification bias (5) and another study reported that verification bias decreased specificity (21). Neither of these studies reported on the effects on sensitivity. Two studies did not report on the effects of partial verification bias on sensitivity and specificity. One of these found that partial verification bias increased overall accuracy (37), and the second reported that there was “scope for verification bias” but provided no additional information (30).

Two more studies found no evidence of bias. One was a systematic review of studies of the diagnostic accuracy of exercise electrocardiography (45), and the other was a review of systematic reviews of several different tests (12). The latter study used the relative diagnostic odds ratio as the summary statistic. If partial verification bias tends to increase sensitivity and decrease specificity, as is suggested by some of the studies, then no effect on the diagnostic odds ratio would be expected. This may explain why this review did not find any evidence of partial verification bias.

**Interpretation (Reading Process)**

**Review Bias**

Four studies investigated review bias (6, 12, 19, 45), 3 (6, 19, 45) examined diagnostic and test review bias, and one looked only at diagnostic review bias (Table 5) (12). A review of exercise testing found no effect of either diagnostic or test review bias on sensitivity and specificity (7). A separate review of exercise testing reported that both diagnostic and test review bias led to an increase in sensitivity but had no effect on specificity (5). A study of carcinoembryonic antigen and nitro-blue tests found that failure to avoid review bias may overestimate sensitivity and specificity (19). A review of several different tests looked only at diagnostic review bias and found that it increased overall accuracy (12).

**Clinical Review Bias**

Nine studies looked at the effects of clinical review bias (Table 5) (28, 46, 52, 53, 55–57, 59, 61). Most of these studies examined radiography (46, 52, 56, 57, 61), mammography (55), and myelography and spinal computer tomography (53). Eight studies used an experimental design, and one was a diagnostic accuracy study (28). One found no difference in overall accuracy between tests interpreted with and without clinical history (56). The other studies all found evidence of bias; however, the direction of bias differed among studies. In general, studies found that providing clinical information improved overall accuracy. Six studies reported that sensitivity was increased when clinical information was available (28, 46, 52, 53, 57, 61).

The effects of providing clinical information on specificity varied among these studies: Two reported that specificity decreased (52, 53), 2 found no effect on specificity (46, 61), and the other 2 did not report on the effects on specificity (28, 57). The remaining 2 studies did not report on the effects of providing clinical history on sensitivity and specificity, but both found that overall accuracy was improved when clinical information was provided (55, 59).

**Incorporation Bias**

No studies that looked at the effects of incorporation bias were identified.

**Observer Variability**

Eight studies looked at observer variation; no studies addressed instrument variation (Table 5) (47–51, 54, 58, 60). All studies used an experimental design. Most studies were evaluations of imaging techniques: radiologic detection of fractures (47), mammography (48, 54), and myocardial imaging (51). Other techniques that were evaluated were fine-needle aspiration biopsy (49), histologic examination (50), cytologic examination (60), and bronchial brush specimens (58). All 8 studies found evidence of interobserver variability, and 2 found evidence of intraobserver variability (48, 50); one of these studies reported that interobserver variability was greater than intraobserver variability (48). Two studies found that more experienced reviewers, or experts, provided greater sensitivity (49, 60), whereas another found that experience was not related to interobserver variability (58).

**Analysis**

**Handling of Indeterminate Results**

Two studies looked at the effects of uninterpretable test results (Table 6) (7, 18). One of these studies stated that a large proportion of results would be excluded if unsatisfactory test results were excluded but provided no evidence on how this may lead to biased estimates of test performance (18). The other study found that the treatment of equivocal or nondiagnostic test results was not associated with overall accuracy (7).

**Arbitrary Choice of Threshold Value**

No studies that provided evidence of the effect of the choice of threshold value were identified.

**Discussion**

The searches identified a relatively small number of studies that looked specifically at the effects of bias and variation on estimates of diagnostic test performance. These studies were concentrated in 7 areas of bias and variation: demographic features (10 studies), disease prevalence (6 studies), disease severity (6 studies), inappropriate reference standard (8 studies), partial verification bias (20 studies), clinical review bias (9 studies), and observer variation (8 studies). Other sources of bias commonly believed...
to affect studies of diagnostic test performance, such as incorporation bias, treatment paradox, arbitrary choice of threshold value, and dropouts, were not considered in any studies.

**Population**

The evidence shows that differences in populations affect estimates of diagnostic performance. However, the extent and direction of the effect of variations in a population can vary, even among studies of the same index test.

Demographic features have shown strong associations with test performance and generally showed a greater effect on estimates of sensitivity than on specificity. Studies that observed effects on specificity generally found that factors that increased sensitivity also decreased estimates of specificity. There was also evidence that both disease severity and prevalence may affect estimates of test performance. Sensitivity tended to be increased in populations with more

<table>
<thead>
<tr>
<th>Study Details</th>
<th>Design</th>
<th>Index Test</th>
<th>Study Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arana et al., 1990 (29)</td>
<td>Review</td>
<td>Thyrotropin-releasing hormone stimulation</td>
<td>10 studies</td>
</tr>
<tr>
<td>Bowler et al., 1998 (30)</td>
<td>Diagnostic accuracy</td>
<td>Necropsy</td>
<td>307 patients</td>
</tr>
<tr>
<td>Boyko et al., 1988 (31)</td>
<td>Modeling</td>
<td>NA</td>
<td>Formulas used to model theoretical effects</td>
</tr>
<tr>
<td>Cecil et al., 1996 (32)</td>
<td>Diagnostic accuracy</td>
<td>Stress single-photon emission computed tomography thallium testing</td>
<td>4354 records selected from computerized database</td>
</tr>
<tr>
<td>De Neef, 1987 (34)</td>
<td>Modeling</td>
<td>New rapid antigen detection tests</td>
<td>Models used to vary reference standard accuracy</td>
</tr>
<tr>
<td>Detrano et al., 1988 (5, 6)</td>
<td>Review</td>
<td>Exercise thallium scintigraphy</td>
<td>56 primary studies</td>
</tr>
<tr>
<td>Detrano et al., 1989 (7)</td>
<td>Review</td>
<td>Exercise electrocardiography</td>
<td>60 primary studies</td>
</tr>
<tr>
<td>Diamond, 1991 (35)</td>
<td>Modeling</td>
<td>NA</td>
<td>Series of computer simulations using the Begg–Greenes method†</td>
</tr>
<tr>
<td>Diamond, 1992 (36)</td>
<td>Modeling</td>
<td>NA</td>
<td>Series of computer simulations using the Begg–Greenes method†</td>
</tr>
<tr>
<td>Froelicher et al., 1998 (28)</td>
<td>Diagnostic accuracy</td>
<td>Electrocardiography and angiographic calipers</td>
<td>814 consecutive patients with angina</td>
</tr>
<tr>
<td>Lijmer et al., 1999 (12)</td>
<td>Review</td>
<td>Various tests</td>
<td>184 primary studies of 218 tests</td>
</tr>
<tr>
<td>Lijmer et al., 1996 (37)</td>
<td>Diagnostic accuracy</td>
<td>Noninvasive tests</td>
<td>464 consecutive patients with suspected disease</td>
</tr>
<tr>
<td>Miller et al., 1998 (38)</td>
<td>Diagnostic accuracy</td>
<td>Stress imaging</td>
<td>15,945 low-risk patients</td>
</tr>
<tr>
<td>Mol et al., 1999 (39)</td>
<td>Review</td>
<td>Nuchal translucency measurement</td>
<td>25 studies</td>
</tr>
<tr>
<td>Morise and Diamond, 1994 and 1995 (15, 16)</td>
<td>Diagnostic accuracy</td>
<td>Exercise electrocardiography</td>
<td>4467 patients with suspected coronary disease</td>
</tr>
<tr>
<td>Panzer et al., 1987 (40)</td>
<td>Diagnostic accuracy</td>
<td>Clinical findings</td>
<td>374 patients with stroke and focal defects</td>
</tr>
<tr>
<td>Phelps and Hutson, 1995 (41)</td>
<td>Modeling</td>
<td>NA</td>
<td>Monte Carlo studies</td>
</tr>
<tr>
<td>Phiblack et al., 1982 (18)</td>
<td>Diagnostic accuracy</td>
<td>Graded exercise test</td>
<td>208 consecutive patients</td>
</tr>
<tr>
<td>Ransohoff and Muir, 1982 (42)</td>
<td>Review</td>
<td>Serum ferritin levels</td>
<td>2 studies</td>
</tr>
<tr>
<td>Ransohoff et al., 1978 (19)</td>
<td>Review</td>
<td>Carcinoembryonic antigen and nitroblue tetrazolium tests</td>
<td>17 studies of carcinoembryonic antigen and 16 of nitroblue tetrazolium</td>
</tr>
<tr>
<td>Roger et al., 1997 (20)</td>
<td>Diagnostic accuracy</td>
<td>Exercise echocardiography</td>
<td>3679 consecutive patients</td>
</tr>
<tr>
<td>Rozanski et al., 1983 (21)</td>
<td>Diagnostic accuracy</td>
<td>Exercise ventriculography</td>
<td>77 angiographically normal patients</td>
</tr>
<tr>
<td>Santana-Boado et al., 1998 (22)</td>
<td>Diagnostic accuracy</td>
<td>Single-photon emission computed tomography</td>
<td>702 consecutive low-risk patients</td>
</tr>
<tr>
<td>Thibodeau, 1981 (43)</td>
<td>Modeling</td>
<td>NA</td>
<td>Various statistical models</td>
</tr>
<tr>
<td>van Rijkom and Verdonschot, 1995 (27)</td>
<td>Review</td>
<td>Tests for approximal caries</td>
<td>39 sets of sensitivity and specificity data</td>
</tr>
<tr>
<td>Zhou, 1994 (44)</td>
<td>Modeling and diagnostic accuracy</td>
<td>NA</td>
<td>429 patients</td>
</tr>
</tbody>
</table>

* DSM-III = Diagnostic and Statistical Manual of the American Psychological Association, 3rd edition; NA = not applicable; RDC = Research Diagnostic Criteria; † = increased; ‡ = decreased. † From Begg C and Greenes R (33).
severe disease or increased disease prevalence. Disease severity had little effect on estimates of specificity, and the effect of disease prevalence on specificity varied. The way in which participants are selected for inclusion in studies of diagnostic accuracy has also been shown to affect test performance. However, the studies that investigated this variable looked at very different aspects of patient selection; thus, it is difficult to draw overall conclusions.

**Test Protocol**

Very few studies investigated the effects of biases and sources of variation associated with test protocol, and those that did reported mixed results. Because of the lack of evidence on the effects of test protocol, it is difficult to draw conclusions regarding the effect of this variable on estimates of test performance. The magnitude of the effect of these biases and sources of variation is probably linked to the test and condition being investigated. For example, the effect of differences in test execution is probably much greater for a test that requires some degree of expertise to perform than for a test that is very straightforward to perform. Similarly, treatment paradox and disease progression bias are more likely to have significant effects on studies of
tests for acute diseases that may be easily treated (for example, infections) and that may change more rapidly than chronic conditions that do not respond well to treatment and that may remain in the same stage for longer periods.

Reference Standard
The evidence was strong for the effect of biases associated with verification procedure on test performance. All studies that looked at the effects of using an inappropriate reference standard found that test performance was affected; however, the direction of the effect differed among studies. Theoretically, if the reference standard is not 100% accurate, the index test may correctly classify results that have been incorrectly classified by the reference standard. This would be expected to lead to an underestimation of test performance. It is also possible that an imperfect reference standard may classify results of the index test as being correct when they are actually incorrect. This would be expected to lead to overestimation of test performance. Thus, an inaccurate reference standard could affect test performance in either way.

Many studies looked at the effects of verification bias, especially partial verification bias. Most reported that verification influenced estimates of test performance. In theory, if all of the patients with negative test results are not verified by the reference standard and are subsequently omitted from the 2 × 2 table, estimates of sensitivity would be inflated because patients with false-negative test results will go undetected. This is supported by the evidence; all studies that observed a significant effect on sensitivity found that sensitivity was increased in the presence of verification bias. However, as with many other biases, the effects on specificity were less clear.

Table 5. Interpretation (Reading Process)*

<table>
<thead>
<tr>
<th>Study Details</th>
<th>Design</th>
<th>Index Test</th>
<th>Study Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arana et al., 1990 (29)</td>
<td>Review</td>
<td>Thyrotropin-releasing hormone stimulation</td>
<td>10 studies</td>
</tr>
<tr>
<td>Berbaum et al., 1988 (46)</td>
<td>Experimental</td>
<td>Radiography</td>
<td>40 radiographs examined with and without clinical information</td>
</tr>
<tr>
<td>Berbaum et al., 1989 (47)</td>
<td>Experimental</td>
<td>Radiography</td>
<td>40 radiographs examined by a group of radiologists and a group of orthopedic surgeons</td>
</tr>
<tr>
<td>Ciccone et al., 1992 (48)</td>
<td>Experimental</td>
<td>Mammography</td>
<td>45 mammograms; 7 radiologists</td>
</tr>
<tr>
<td>Cohen et al., 1987 (49)</td>
<td>Experimental</td>
<td>Fine-needle aspiration biopsy</td>
<td>50 specimens examined by 5 observers</td>
</tr>
<tr>
<td>Corley et al., 1997 (50)</td>
<td>Experimental</td>
<td>Histologic diagnosis of pneumonia</td>
<td>39 lung biopsy samples, 4 pathologists</td>
</tr>
<tr>
<td>Cuanon et al., 1980 (51)</td>
<td>Experimental</td>
<td>Tc 99m phosphate myocardial imaging</td>
<td>250 myocardial slides evaluated by 6 observers</td>
</tr>
<tr>
<td>Detrano et al., 1988 (5, 6)</td>
<td>Review</td>
<td>Exercise thallium scintigraphy</td>
<td>56 primary studies</td>
</tr>
<tr>
<td>Detrano et al., 1989 (7)</td>
<td>Review</td>
<td>Exercise electrocardiography</td>
<td>60 primary studies</td>
</tr>
<tr>
<td>Doubleit et al., 1981 (52)</td>
<td>Experimental</td>
<td>Radiography</td>
<td>8 test radiographs; 4 with suggestive and 4 nonsuggestive history</td>
</tr>
<tr>
<td>Eldevik et al., 1982 (53)</td>
<td>Experimental</td>
<td>Myelography and computed tomography</td>
<td>107 patients assessed with and without clinical history</td>
</tr>
<tr>
<td>Elmore et al., 1994 (54)</td>
<td>Experimental</td>
<td>Mammography</td>
<td>150 mammograms, 10 radiologists</td>
</tr>
<tr>
<td>Elmore et al., 1997 (55)</td>
<td>Experimental</td>
<td>Mammography</td>
<td>100 radiographs assessed with and without clinical history</td>
</tr>
<tr>
<td>Froelicher et al., 1998 (28)</td>
<td>Diagnostic accuracy</td>
<td>Electrocardiography and angiographic calipers</td>
<td>814 consecutive patients with angina</td>
</tr>
<tr>
<td>Good et al., 1990 (56)</td>
<td>Experimental</td>
<td>Chest radiography</td>
<td>247 radiographs assessed with and without clinical history</td>
</tr>
<tr>
<td>Lijner et al., 1999 (12)</td>
<td>Review</td>
<td>Various tests</td>
<td>184 primary studies of 218 tests</td>
</tr>
<tr>
<td>Potchen et al., 1979 (57)</td>
<td>Experimental</td>
<td>Chest radiography</td>
<td>3 groups of radiologists; different combinations of data</td>
</tr>
<tr>
<td>Raab et al., 1995 (58)</td>
<td>Experimental</td>
<td>Bronchial brush specimens</td>
<td>100 bronchial brush specimens examined by different observers</td>
</tr>
<tr>
<td>Raab et al., 2000 (59)</td>
<td>Experimental</td>
<td>Bronchial brush specimens</td>
<td>97 specimens, assessed with and without clinical information</td>
</tr>
<tr>
<td>Ransohoff et al., 1978 (19)</td>
<td>Review</td>
<td>Carcinoembryonic antigen and nitroblue tetrazolium tests</td>
<td>17 studies of carcinoembryonic antigen and 16 of nitroblue tetrazolium</td>
</tr>
<tr>
<td>Ronco et al., 1996 (60)</td>
<td>Experimental</td>
<td>Colophistologic and cytologic screening</td>
<td>61 samples examined by cytologists and experts</td>
</tr>
<tr>
<td>Schreiber, 1963 (61)</td>
<td>Experimental</td>
<td>Chest radiography</td>
<td>100 chest radiographs assessed with and without clinical information</td>
</tr>
</tbody>
</table>

* DSM-III = Diagnostic and Statistical Manual of the American Psychological Association, 3rd edition; NA = not applicable; RDC = Research Diagnostic Criteria; ↑ = increased; ↓ = decreased.
Interpretation

Reading processes that involve interpretation of results affect estimates of test performance. Both diagnostic and test review biases were found to increase sensitivity; however, no effect on specificity was noted. An effect on sensitivity would be expected because knowledge of the index test result when interpreting the reference standard (or vice versa) probably increases the agreement between tests. This in turn leads to a greater number of true-positives and true-negative results and would be expected to increase estimates of both sensitivity and specificity. It is unclear why studies did not find significant effects on specificity. Perhaps the effects on specificity are smaller and any effect may therefore not reach statistical significance.

The availability of clinical information to the person interpreting the results of the index test was found to increase sensitivity. Although the evidence for an effect on specificity was minimal, specificity decreased in 2 studies. The provision of clinical information probably has different effects depending on the test being evaluated. Whether clinical information should be available in a particular diagnostic study should be carefully considered in each case. It seems that the best approach to interpreting the results of a diagnostic accuracy study would be to determine whether the clinical information available to those interpreting the results of the index test is the same as the clinical information that would be available when the test is interpreted in practice.

All studies that looked at the effects of observer variation found significant differences among observers in their estimates of test performance. Therefore, the effects of observer variation will inevitably be greater for tests that involve a strong degree of subjective interpretation compared with a fully automated test.

Table 5—Continued

<table>
<thead>
<tr>
<th>Source of Bias or Variation</th>
<th>Factors Investigated</th>
<th>Effect on Sensitivity</th>
<th>Effect on Specificity</th>
<th>Effect on Overall Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inappropriate reference standard</td>
<td>DSM-III instead of RDC as the reference standard</td>
<td>↓</td>
<td>Not reported</td>
<td>NA</td>
</tr>
<tr>
<td>Clinical review bias</td>
<td>Availability of clinical information</td>
<td>↑</td>
<td>None</td>
<td>↑</td>
</tr>
<tr>
<td>Observer variation</td>
<td>Difference between radiologists and orthopedic surgeons</td>
<td>NA</td>
<td>NA</td>
<td>Associated</td>
</tr>
<tr>
<td>Observer variation</td>
<td>Difference between radiologists and orthopedic surgeons</td>
<td>NA</td>
<td>NA</td>
<td>Associated</td>
</tr>
<tr>
<td>Observer variation</td>
<td>Inter- and intraobserver variation</td>
<td>NA</td>
<td>NA</td>
<td>Associated</td>
</tr>
<tr>
<td>Observer variation</td>
<td>Effect of training and experience</td>
<td>↑</td>
<td>↑</td>
<td>NA</td>
</tr>
<tr>
<td>Observer variation</td>
<td>Inter- and intraobserver variation</td>
<td>NA</td>
<td>NA</td>
<td>None</td>
</tr>
<tr>
<td>Observer variation</td>
<td>Interobserver variation</td>
<td>NA</td>
<td>NA</td>
<td>Associated</td>
</tr>
<tr>
<td>Review bias</td>
<td>Lack of blinding, that is, presence of review bias</td>
<td>↑</td>
<td>Not reported</td>
<td>↑</td>
</tr>
<tr>
<td>Review bias</td>
<td>Lack of blinding, that is, presence of review bias</td>
<td>NA</td>
<td>NA</td>
<td>None</td>
</tr>
<tr>
<td>Clinical review bias</td>
<td>Suggestive clinical history</td>
<td>↑</td>
<td>↓</td>
<td>NA</td>
</tr>
<tr>
<td>Clinical review bias</td>
<td>Availability of clinical information</td>
<td>↑</td>
<td>↓</td>
<td>NA</td>
</tr>
<tr>
<td>Observer variation</td>
<td>Interobserver variation</td>
<td>NA</td>
<td>NA</td>
<td>Associated</td>
</tr>
<tr>
<td>Clinical review bias</td>
<td>Availability of clinical information</td>
<td>NA</td>
<td>NA</td>
<td>↑</td>
</tr>
<tr>
<td>Clinical review bias</td>
<td>Availability of clinical information</td>
<td>↑</td>
<td>Not reported</td>
<td>NA</td>
</tr>
<tr>
<td>Clinical review bias</td>
<td>Availability of clinical information</td>
<td>NA</td>
<td>NA</td>
<td>None</td>
</tr>
<tr>
<td>Review bias</td>
<td>Lack of blinding, that is, presence of review bias</td>
<td>NA</td>
<td>NA</td>
<td>↑</td>
</tr>
<tr>
<td>Clinical review bias</td>
<td>Availability of clinical information</td>
<td>↑</td>
<td>Not reported</td>
<td>NA</td>
</tr>
<tr>
<td>Observer variation</td>
<td>Interobserver variation</td>
<td>NA</td>
<td>NA</td>
<td>Associated</td>
</tr>
<tr>
<td>Clinical review bias</td>
<td>Availability of clinical information</td>
<td>NA</td>
<td>NA</td>
<td>↑</td>
</tr>
<tr>
<td>Review bias</td>
<td>Lack of blinding, that is, presence of review bias</td>
<td>↑</td>
<td>↑</td>
<td>NA</td>
</tr>
<tr>
<td>Observer variation</td>
<td>Effect of training and experience (being an “expert”)</td>
<td>↑</td>
<td>Not reported</td>
<td>NA</td>
</tr>
<tr>
<td>Clinical review bias</td>
<td>Availability of clinical information</td>
<td>↑</td>
<td>None</td>
<td>NA</td>
</tr>
</tbody>
</table>
Analysis

Very few studies investigated the effects of biases associated with analysis on test performance. The effect of the exclusion of indeterminate results and the nonarbitrary choice of threshold value remains unclear from the evidence reviewed.

Limitations

The main limitation of our review is the difficulty in identifying articles that examined specific features of the design and conduct of diagnostic studies. Indexing on MEDLINE and other electronic databases focuses on diseases, therapies, and test technologies and not on elements of design. There is no specific way of indexing studies that relate to the diagnostic accuracy of a test (1). In addition, many different names have been used to label the same phenomenon in studies of diagnostic accuracy tests. To try to overcome these difficulties, very broad searches were performed. However, we may have still missed several relevant papers. The information provided in our paper should provide useful examples but may not be comprehensive.

Ideally, we would have liked to provide a quantitative synthesis to assess the magnitude of each of the biases and sources of variation as well as their direction. However, because the studies included were very heterogeneous, a quantitative synthesis was not possible. The studies also measured the effect of the biases and sources of variation in different ways. In particular, diagnostic accuracy and experimental studies looked at the effect of biases and sources of variation within studies, whereas reviews looked at reasons for differences in estimates among studies. It is also likely that different biases and sources of variation will be important in different topic areas. For example, observer variation is likely to be a problem only for studies that involve some degree of subjective interpretation. Also, observer variation is likely to have a greater effect with more subjective interpretations.

Another problem is that sources of bias and variation may act differently depending on the study. For example, for partial verification bias, the effects may differ when the reference standard is not used in selected groups. The group that does not receive verification may, for example, be a random sample of patients, a selected subgroup of patients with negative test results, or all patients with positive test results. All of these situations are called partial verification, but the effects of each situation probably differ. Within a single study, there is only one true effect of a feature, but this true effect may differ depending on the study. Chance and the effect of other factors may obscure the true effect. These factors combine to create difficulty in determining the overall effect of a source of bias or variation.

We included studies that provided both real-life examples of the effects of different biases and sources of variation as well as studies that used modeling to investigate the effects of different biases or sources of variation. When the results of the modeling studies are interpreted, it is important to consider that these studies can provide an indication only of the theoretical effect of a source of bias or variation. The results from these studies need to be supported by additional empirical evidence from real-life examples before more firm conclusions can be drawn (12).

CONCLUSIONS

This paper provides information on the available evidence for the effects of each source of bias and variation in diagnostic accuracy studies. The sources of bias and variation for which there is the most evidence are demographic features, disease prevalence or severity, partial verification bias, clinical review bias, and observer or instrument variation. Some evidence was also available for the effects of distorted selection of participants, absent or inappropriate reference standard, differential verification bias, and review bias. The potential effects of these biases and sources of variation should be considered when interpreting or designing diagnostic accuracy studies. Additional research should be done to investigate potential sources of bias and variation.

From the University of York, York, United Kingdom, and the University of Amsterdam, Amsterdam, the Netherlands.

Disclaimer: The views expressed in this paper are those of the authors and not necessarily those of the Standing Group, the Commissioning Group, or the Department of Health.

Acknowledgments: The authors thank Kath Wright (Centre for Reviews and Dissemination) for carrying out our literature searches. They also

Table 6. Analysis*

<table>
<thead>
<tr>
<th>Study Details</th>
<th>Design</th>
<th>Index Test</th>
<th>Study Sample</th>
<th>Source of Bias or Variation</th>
<th>Factors Investigated</th>
<th>Effect on Sensitivity</th>
<th>Effect on Specificity</th>
<th>Effect on Overall Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detrano et al., 1989 (7)</td>
<td>Review</td>
<td>Exercise electrocardiography</td>
<td>60 primary studies</td>
<td>Handling of indeterminate results</td>
<td>Treatment of equivocal or nondiagnostic tests</td>
<td>NA</td>
<td>NA</td>
<td>None</td>
</tr>
<tr>
<td>Philbrick et al., 1982 (18)</td>
<td>Diagnostic accuracy</td>
<td>Graded exercise test</td>
<td>208 consecutive patients</td>
<td>Handling of indeterminate results</td>
<td>Exclusion of unsatisfactory exercise test results</td>
<td>NA</td>
<td>NA</td>
<td>Unclear</td>
</tr>
</tbody>
</table>

* NA = not applicable.
thank the advisory panel to the review for their help during various stages, including commenting on the protocol and draft report.

Grant Support: Commissioned and funded by the National Health Service R&D Health Technology Assessment Programme (project number 98/27/99).

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Penny Whiting, MSc, Centre for Reviews and Dissemination, University of York, York YO10 5DD, United Kingdom; e-mail, pw2@york.ac.uk.

Current author addresses are available at www.annals.org.

References
Bias and Variation in Diagnostic Studies

1538629
Current Author Addresses: Ms. Whiting and Dr. Kleijnen: Centre for Reviews and Dissemination, University of York, York YO10 5DD, United Kingdom.

Ms. Rutjes and Drs. Reitsma, Glas, and Bossuyt: Department of Clinical Epidemiology & Biostatistics, Academic Medical Center, University of Amsterdam, P.O. Box 22700, 1100 DE Amsterdam, the Netherlands.
Funding: PS is supported by a contract from the Agency for Healthcare Research and Quality to the Southern California Evidence-based Practice Center. The Health Services Research Unit is funded by the Chief Scientist Officer of the Scottish Executive Department of Health. The Health Services Research Unit and Centre for Health Services Research are members of the MRC Health Services Research Collaboration.

Competing interests: JMG is a member of the Guidelines Advisory Committee for the National Institute for Clinical Excellence and the methodological adviser to the Scottish Intercollegiate Guidelines Network. MPE is chairman of the Guidelines Advisory Committee for the National Institute for Clinical Excellence. SHW is a member of the US Preventive Services Task Force and other practice guideline panels involved in updating.

5 European Carotid Surgery Trialists’ Collaborative Group. MRC European Carotid Surgery Trial. Interim results for symptomatic patients with severe (70-99%) or with mild (0-29%) carotid stenosis. Lancet 1991;337:1225-33.
Studies of diagnostic accuracy

Studies of test performance (or accuracy) compare test results between groups of patients with and without the target disease, each of whom undergoes the experi-

Receiver operating characteristic curves

Receiver operating characteristic curves are used in studies of diagnostic accuracy to depict the pattern of sensitivities and specificities observed when the performance of the test is evaluated at several different diagnostic thresholds. Figure 1 is a receiver operating characteristic curve from a study of the detection of endometrial cancer by endovaginal ultrasonography. Women with endometrial cancer are likely to have increased endometrial thicknesses: few women who do not have cancer will have thicknesses above a high threshold whereas few women with endometrial cancer will have thicknesses below a low threshold. This pattern of results is seen in figure 1, with the 5 mm threshold showing high sensitivity (0.98) but poor specificity (0.59) and the 25 mm threshold showing poor sensitivity (0.24) but high specificity (0.98). The overall diagnostic performance of a test can be judged by the position of the receiver operating characteristic line. Poor tests have lines close to the rising diagonal, whereas the lines for perfect tests would rise steeply and pass close to the top left hand corner, where both the sensitivity and specificity are 1. Receiver operating characteristic plots are used in systematic reviews to display the results of a set of studies, the sensitivity and specificity from each study being plotted as a separate point in the receiver operating characteristic space.
computation of a weighted average of the summary statistics across the studies.14 I illustrate the application of three commonly used methods for pooling different summaries of diagnostic accuracy with a case study.

As with systematic reviews of randomised controlled trials, meta-analysis should be considered only when the studies have recruited from similar patient populations (it is problematic to combine studies from general practice with studies from tertiary care), have used comparable experimental and reference tests, and are unlikely to be biased. Even when these criteria are met there may still be such gross heterogeneity between the results of the studies that it is inappropriate to summarise the performance of a test as a single number.

Case study

Detection of endometrial cancer with endovaginal ultrasonography

Smith-Bindman et al published a systematic review of 35 studies evaluating the diagnostic accuracy of endovaginal ultrasonography for detecting endometrial cancer and other endometrial disorders.15 All studies included in the review were of prospective cohort designs and used the results of endometrial biopsy, dilation and curettage, or hysterectomy as a reference standard. Most of the studies presented sensitivities and specificities at several endometrial thicknesses detected by endovaginal ultrasonography (the receiver operating characteristic curve in figure 1 is from one of these studies). The case study is based on the subset of 20 studies from this review that considered the diagnostic accuracy of endovaginal ultrasonography in ruling out endometrial cancer with endometrial thicknesses of 5 mm or less. Figure 2 shows the sensitivities and specificities for the 20 studies.

Sources of heterogeneity

The choice of meta-analytical method depends in part on the pattern of variability (heterogeneity) observed in the results. Heterogeneity can be considered graphically by plotting sensitivities and specificities from the studies as points on a receiver operating characteristic plot (fig 3). Some divergence of the results around a central point is to be expected by chance, but variation in other factors, such as patient selection and features of the study’s design, may increase the observed variability.16

Table 1 Framework for considering study quality and likelihood of bias

<table>
<thead>
<tr>
<th>Study feature</th>
<th>Qualities sought</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample of patients</td>
<td>Consecutive or randomly selected sample, recruited as single cohort unclassified by disease state, recruitment from clinical setting and point in referral process where test would be used, selection and referral processes fully described, clinical and demographic characteristics fully described, complete</td>
</tr>
<tr>
<td>Reference diagnosis</td>
<td>Method and tests described in detail, positive and negative diagnoses clearly described, diagnosis likely to be close to truth, available for all patients, based on same tests and information in all patients, blinding procedures used to prevent knowledge of result of experimental test influencing the reference diagnosis, made before treatment commenced</td>
</tr>
<tr>
<td>Experimental test</td>
<td>Application of test described in detail, positive and negative test results clearly described, blinding procedures used to ensure that test is undertaken without knowledge of reference diagnosis, test undertaken before treatment commenced, results reported for all patients, including those with “grey zone” results</td>
</tr>
</tbody>
</table>

Fig 2 Estimates from 20 studies of sensitivity and specificity of measurement of endometrial thicknesses of more than 5 mm using endovaginal ultrasonography for detecting endometrial cancer.15 Points indicate estimates of sensitivity and specificity. Horizontal lines are 95% confidence intervals for estimates. Size of points reflects total sample size
One important extra source of heterogeneity is variation introduced by changes in diagnostic threshold. Studies may use different thresholds to define positive and negative test results. Some may have done this explicitly—for example, by varying numerical cut-off points used to classify a biochemical measurement as positive or negative, whereas for others there may be naturally occurring variations in diagnostic thresholds between observers, laboratories, or machines. The choice of a threshold may also vary according to the prevalence of the disease—when the disease is rare a more extreme threshold may have been used to avoid large numbers of false positive diagnoses. Unlike other sources of variability, variation of the diagnostic threshold introduces a particular pattern into the receiver operating characteristic plot of study results, such that the points show curvature (fig 1).

If there is no heterogeneity between the studies, the best summary estimate of test performance should be a single point on the receiver operating characteristic graph. The first two methods estimate such a summary, first by pooling sensitivities and specificities then by pooling positive and negative likelihood ratios. The third method is more complex and pools diagnostic odds ratios to take account of possible heterogeneity in diagnostic threshold.

**Pooling sensitivities and specificities**

The pooled estimate of sensitivity is 0.96 (95% confidence interval 0.93 to 0.99) and is depicted by the horizontal line on the receiver operating characteristic plot in figure 3 (left). The overall estimate of mean specificity is lower: 0.61 (0.55 to 0.66).

Heterogeneity is, however, clearly evident in figure 3 (left): although the study points lie reasonably close to the summary sensitivity (test for heterogeneity, P = 0.04), the results of many studies lie some distance from the summary specificity (test for heterogeneity, P < 0.001).

Regardless of the causes of the heterogeneity, the overall high estimate and relative consistency of the sensitivity results does suggest that a negative test result could be of potential clinical use in ruling out endometrial cancer. As there is heterogeneity between specificities, however, it is more appropriate to note the range of specificities (0.27 to 0.88) rather than to quote the average value of 0.61. It is difficult to draw a conclusion about test specificity: the observed values vary considerably and there is no understanding from this analysis as to the reasons for the variation.

**Pooling likelihood ratios**

For the case study the pooled estimate of the positive likelihood ratio was not particularly high (2.54, 2.16 to 2.98), and the values varied significantly between the studies (test for heterogeneity, P < 0.001). In figure 3 (centre) it is clear that the summary positive likelihood ratio lies some distance from many of the values. Again it is debatable whether reporting the average value of such heterogeneous results is sensible, but it is unlikely that a positive test result could provide convincing evidence of the presence of endometrial cancer as the positive likelihood ratios are all below 10 (data not shown).

The negative likelihood ratios show no evidence of significant heterogeneity (test for heterogeneity, P = 0.09), the pooled estimate being 0.09 (0.06 to 0.13), with the summary line on the receiver operating characteristic plot in figure 3 (centre) lying close to the results of most of the studies. This finding again shows...
that a measurement of an endometrial thickness of 5 mm or less made by endovaginal ultrasonography can provide reasonably convincing evidence to rule out endometrial cancer.

Although these conclusions concerning potential diagnostic use are similar to those obtained by pooling sensitivities and specificities, the summaries obtained by pooling likelihood ratios can be more easily interpreted and applied to clinical practice. The box describes how the summary negative likelihood ratio can be applied to estimate the probability of endometrial cancer in a woman with a negative test result.

**Diagnostic odds ratios and summary receiver operating characteristic curves**

If the observed heterogeneity between the studies arises due to variation in the diagnostic threshold, estimates of summary sensitivity and specificity or summary positive and negative likelihood ratios will underestimate diagnostic performance. In this situation the appropriate meta-analytical summary is not a single point in the receiver operating characteristic space but the receiver operating characteristic curve itself. Methods of deriving the best fitting summary receiver operating characteristic curve are necessarily more complex.

How is a summary receiver operating characteristic curve estimated? The simplest approach involves calculating a single summary statistic for each study—the diagnostic odds ratio (box). Each diagnostic odds ratio corresponds to a particular receiver operating characteristic curve. If the studies in a review all relate to the same curve they may have consistent diagnostic odds ratios even if they have variable sensitivities and specificities. Table 2 gives examples of diagnostic odds ratios corresponding to particular sensitivities, specificities, and positive and negative likelihood ratios.

In the case study it is possible that some of the observed heterogeneity could be explained by a threshold effect, perhaps due to differences in calibration of the ultrasound machines. The estimate of the summary diagnostic odds ratio is 28.0 (18.2 to 43.2) and is reasonably consistent across the studies (test for heterogeneity, P = 0.3), suggesting that the points indeed could have originated from the same receiver operating characteristic curve. The summary diagnostic odds ratio can be interpreted in terms of sensitivities and specificities by consulting table 2 (for example, a diagnostic odds ratio of 29 corresponds to a sensitivity of 0.95 and a specificity of 0.60 and to a sensitivity of 0.60 and specificity of 0.95) or by plotting the corresponding summary receiver operating characteristic curve (fig 3 (right)). This method does not yield a unique joint summary estimate of sensitivity and specificity; it is only possible to obtain a summary estimate of one value by specifying the value of the other. This greatly limits its clinical application.

**Discussion**

Systematic reviews of diagnostic accuracy have not, as yet, made the same impression on the practice of evidence based health care as have systematic reviews of randomised controlled trials. Reasons relate to reliability, heterogeneity, and clinical relevance.

**Table 2 Examples of diagnostic odds ratios corresponding to particular pairings of sensitivity and specificity and positive and negative likelihood ratios**

<table>
<thead>
<tr>
<th>Specificity</th>
<th>Sensitivity 0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>0.95</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>9</td>
<td>19</td>
</tr>
<tr>
<td>0.6</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>14</td>
<td>29</td>
<td>149</td>
</tr>
<tr>
<td>0.7</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>21</td>
<td>44</td>
<td>231</td>
</tr>
<tr>
<td>0.8</td>
<td>4</td>
<td>6</td>
<td>9</td>
<td>16</td>
<td>36</td>
<td>76</td>
<td>396</td>
</tr>
<tr>
<td>0.9</td>
<td>9</td>
<td>14</td>
<td>21</td>
<td>36</td>
<td>81</td>
<td>171</td>
<td>891</td>
</tr>
<tr>
<td>0.95</td>
<td>19</td>
<td>29</td>
<td>44</td>
<td>76</td>
<td>171</td>
<td>361</td>
<td>1881</td>
</tr>
<tr>
<td>0.99</td>
<td>99</td>
<td>149</td>
<td>231</td>
<td>396</td>
<td>891</td>
<td>1881</td>
<td>9801</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Negative likelihood ratio</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive likelihood ratio</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Are systematic reviews of diagnostic studies reliable?**

Many meta-analyses of the accuracy of diagnostic tests are hindered by the poor quality of the primary studies; most published evaluations of the accuracy of diagnostic tests having at least one flaw. Headway has been made in understanding the importance of particular features of a study’s design and in improving quality, but for many diagnostic tests few high quality studies have been undertaken and published.

The reliability of a review also depends crucially on whether the included studies are an unbiased selection. As with all reviews, systematic reviews of diagnostic tests are susceptible to publication bias, and this may be a greater problem than for randomised controlled trials. No investigations, however, have been conducted to estimate rates of publication bias for studies of diagnostic accuracy.
How useful are systematic reviews to a practising clinician?

Heterogeneity of the results of studies of diagnostic accuracy is common but in itself does not prevent conclusions of clinical value from being drawn.22 Despite heterogeneity being observed in the case study, it was still possible to draw a conclusion of clinical value—that an endometrial thickness of 5 mm or less can rule out endometrial cancer.

Diagnostic odds ratios and summary receiver operating characteristic curves are, however, often promoted as the most statistically valid method for combining test results when there is heterogeneity between studies, and they are commonly used in systematic reviews of diagnostic accuracy.13 Unfortunately summary curves are of little use to practising healthcare professionals: they can identify whether a test has potential clinical value, but they cannot be used to compute the probability of disease associated with specific test outcomes. Their use is also based on a potentially inappropriate and untested assumption that observed heterogeneity has arisen through variation in diagnostic threshold. In the case study, whereas the diagnostic odds ratio was a reasonably consistent summary statistic across the studies, there was no evidence to suggest that the observed heterogeneity arose through variations in diagnostic threshold (all included studies had a 5 mm threshold for endometrial thickness). Variation in referral patterns, sample selection, and study methods may be more likely explanations for the heterogeneity. There is no clear statistical advantage in using a summary receiver operating characteristic approach to synthesise the results over pooling sensitivity and specificity or likelihood ratios unless there is a threshold effect. Empirical research is urgently required to find out whether the simpler methods for pooling sensitivities, specificities, and likelihood ratios are likely to be seriously misleading in practice and whether apparent threshold effects are really due to variations in diagnostic threshold rather than alternative sources of heterogeneity.

Are studies of diagnostic accuracy clinically relevant?

Systematic reviews of the accuracy of tests do not always answer the most clinically relevant question. New tests are often evaluated for their ability to replace or be used alongside existing tests. The important issues are comparisons of tests or comparisons of testing algorithms: these would be best addressed in properly designed comparative studies, rather than by synthesising studies of diagnostic accuracy separately for each test.

The evaluation of the diagnostic accuracy of a test is also only one component of assessing whether it is of clinical value.23 24 Treatment interventions are recommended for use in health care only if they are shown on average to be of benefit to patients: the same criterion should also be applied for the use of a diagnostic test, and even the most accurate of tests can be clinically useless or do more harm than good. It should always be considered whether undertaking a systematic review of studies of diagnostic accuracy is likely to provide the most useful evidence of the value of a diagnostic intervention.

I thank Rebecca Smith-Bindman for providing the data for the case study.

Competing interests: None declared.

References

8 Deeks JJ, Morris JM. Evaluating diagnostic tests. Bailliere’s Clinical Obstetri-

9 Jaeschke R, Guyatt GH, Sackett DL. For the Evidence-Based Medicine Working Group. Users’ guides to the medical literature. VI. How to use an article about a diagnostic test: B: What are the results and will they help me in caring for my patients? JAMA 1994;271:703-7.

Endpiece

Nothing like experience

We live in an age of mass literacy. We are all writing it or at any rate talking it: the memoir, the apologia, the cv, the cri de coeur. Nothing, for now, can compete with experience—so unspeakably authentic, and so liberally and democratically dispensed. Experience is the only thing we share equally, and everyone senses this.

Martin Amis, All from experience, London: Jonathan Cape, 2000

Education and debate
Under-reporting of clinical trials is unethical

See page 1015

6 years ago the Danish Research Ethics Committee System, after considering what influence the results of existing research should have on the ethical evaluation of proposals for new clinical trials, declared that researchers should review all relevant evidence before they submitted a new protocol for ethical assessment.1 This injunction asks no more of researchers than respect for the principle that science is cumulative. Patients being invited to participate in a clinical trial have a right to expect that its design has been informed by a scientifically defensible review of what is known already.

One of the problems researchers face in complying with this principle, however, is that there is no general access to "all relevant evidence". In today's Lancet, J Pich and colleagues from Spain show that less than a third of clinical trials approved by an ethics committee at a large hospital had been reported in peer-reviewed journals 3 years after completion. These investigators endorse a view published in The Lancet last year2 that ensuring public dissemination of the results of clinical trials falls within the scope of research ethics committees, and that, to the extent that they are not fulfilling these expectations, these committees are failing the public.

Ethicists and members of research ethics committees should certainly be concerned about under-reporting of clinical trials. Not only does under-reporting breach implied contracts with the patients who participate in these studies (who assume that they are contributing to a growth in knowledge); it can also lead to biased and unnecessarily imprecise estimates of the effects of treatments.3,4 Because these unreliable estimates sometimes harm patients,3 Lock and Wells have deemed under-reporting of research to be a form of scientific as well as ethical misconduct.5

Research ethics committees have been challenged previously to consider whether they are behaving unethically in paying so little attention to this problem.2,7,10 There is disappointingly scanty evidence that they and the bioethics community more generally have acknowledged their responsibility to address a problem that is of importance to the patients whose interests they purport to protect. Complaints that research ethics committees lack the capacity to follow up the studies they endorse is not an acceptable excuse for their acquiescence in the under-reporting of trials. These committees are the gatekeepers for clinical research and there is no reason why they should not join with others who have implemented one of the steps needed to reduce the adverse consequences of under-reporting—namely, prospective registration of controlled clinical trials.6 Public registration of all controlled trials at inception will never be a complete cure for under-reporting. However, it does provide a means of identifying studies that should be reported. Furthermore, as Simes showed nearly two decades ago,8,9 comparison of the results of trials that were registered before their results were known with those of studies that were not registered at inception, helps to identify and take account of biased under-reporting.

There has been wide endorsement of the principle of prospective registration of clinical trials over recent years. As made clear in the TrialsCentral website,11 there has also been encouraging progress in making information about planned and ongoing trials publicly available (unlike the information that is to be registered under the European Clinical Trials Directive, which will remain confidential). After the 1997 Food and Drug Administration Modernization Act, for example, the US National Library of Medicine developed and launched an impressive registration system for clinical trials, which is publicly available.12 At about the same time, Current Controlled Trials, working with the British Medical Research Council and others, launched a metaRegister of Controlled Trials and more recently, a system for assigning unique International Standard Randomized Controlled Trials Numbers (ISRCTNs). Both of these are publicly available.13 The European Science Foundation has advised its member organisations to participate in these schemes,14 and has recently emphasised the importance of compliance.15

Even in applying the basic step of registration, however, obstacles remain. Despite clear evidence that progress has been possible without legislation, and although there have been repeated calls for clinical trial registration at meetings organised by the European Commission, the Commission recently rejected an application for resources to promote trial registration across Europe because it deemed progress to be impossible without legislation. Furthermore, it must be acknowledged that the progress made reflects almost entirely registration of non-commercial trials. Despite the fact that basic information about most late-phase drug trials is already effectively in the public domain through presentations at scientific meetings, industry has so far made little contribution to the progress in trial registration, even in the USA, where registration of all trials relevant to serious or life-threatening diseases is required by law.16 The example set by GlaxoWellcome 5 years ago17 has not yet been followed by other companies, even though more than 2 years have passed since the Association of the British Pharmaceutical Industry announced that it had advised all its member companies to emulate GlaxoWellcome’s example.

We endorse Mann’s view7 that those responsible for the work of research ethics committees need to reconsider their apparent acquiescence in this state of affairs. We believe that they should require registration of clinical trials as a
component of ethical review, and ensure that patients are
advised to withhold their consent to participate unless there is
publicly accessible information about the trial and the
investigators have undertaken to make the results of the
study publicly available within a reasonable time after its
conclusion.

GA and IC have worked for registration of controlled trials over many years because this is a responsibility of our salaried positions. We have advised the European Science Foundation, Current Controlled Trials, and other organisations, in Germany, the UK, and other countries, but we have not received fees for this advice. Current Controlled Trials has offered to make a contribution to a research charity chosen by Iain Chalmers when he retires and steps down as chair of its international advisory committee at the end of June.

*Gerd Antes, Iain Chalmers
German Cochrane Centre, Institut für Medizinische Biometrie und
Medizinische Informatik, Universität Freiburg, Freiburg, D-79104,
Germany, and UK RCT Registration Project, c/o UK Cochrane Centre,
Oxford, UK (IC)
(e-mail: antes@cochrane.de)

1 Goldbeck-Wood S. Denmark takes lead on research ethics. BMJ 1998; 316: 1189.
2 Mann H. Research ethics committees and public dissemination of
3 Dickersin K. How important is publication bias? A synthesis of available
4 Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important is publication bias? A synthesis of available
5 Chalmers I. Systematic reviews and registries of ongoing trials for
scientific and ethical trial design, monitoring, and reporting. In: Egger M, Davey Smith G, Altman D, eds. Systematic reviews in health
Group, 1996: xii.
8 Savulescu J, Chalmers I, Blunt J. Are research ethics committees
behaving unethically? Some suggestions for improving performance and
9 Tonks A. A clinical trials register for Europe: stop talking and start
10 Simes RJ. Publication bias: the case for an international registry of
11 TrialsCentral. Your source for online clinical trials information.
12 ClinicalTrials.Gov. Linking patients to medical research.
13 Current Controlled Trials. Featuring the *meta*Register of controlled
14 European Science Foundation Policy Briefing. Controlled clinical trials.
(accessed Feb 17, 2003).
15 European Science Foundation. Press release on registration of
randomised controlled trials: for attention and action by Chief
Executives, ESF Member Organisations. Feb 11, 2003:
16 Münchheimer E, Anderson D. Survey of public information about ongoing
clinical trials funded by industry: evaluation of completeness and

**Taking the sting out of ant stings: venom immunotherapy to prevent anaphylaxis**

See page 1001.

**Ants** (Formicidae), like bees, wasps, and hornets, are insects of the order Hymenoptera. Usually they threaten human beings only by their prodigious numbers and ruthless, organised, and persistent perversiveness. In Australia, the first recorded envenoming by an endemic animal did not involve any of that country’s formidable snakes, spiders, or cnidarians but a mass attack by green
tree-ants (*Ocypus smaragdina*) on the pioneering naturalist Joseph Banks, in 1770.1 However, the bite of a single ant can be traumatic, as in the case of the giant *Paraponera gigantea* (“tocandira”) of Rondonia, Brazil, which is 22 mm long. One sting by some species of ant may provoke fatal anaphylaxis in a sensitised person.

Anaphylaxis is increasingly recognised as an important mode of death in many parts of the world. Apart from drugs and foods, stings by Hymenoptera are among the most familiar precipitants. In England, wasps and hornets (Vespidae) and bees (Apidae) are responsible for four to five of the average of about 20 deaths attributed to anaphylaxis each year, although some cases are probably missed. The USA has a similar pattern among a total of perhaps 50–60 anaphylactic deaths each year. However, between 1920 and 1940, two species of fire ants (*Solenopsis richteri* and *S invicta*) were inadvertently imported from South America. The distribution of *S richteri* has remained relatively restricted around its original port of entry at Mobile, Alabama, whereas *S invicta*, “the ant from hell”, has spread throughout south-eastern USA. Between 0·8% and 40% of Americans are now thought to be hypersensitive to venom from these ants, and more than 50 anaphylactic deaths from the sting of imported fire-ants have been reported since 1972.

In some parts of Australia, ant-sting anaphylaxis has now emerged as an important medical problem, but there it is caused by indigenous rather than imported ants. Of the three species commonly involved, *Myrmecia pilosula* (the “jack-jumper”), has proved the most dangerous. The name refers to its salutary style of locomotion when disturbed. The jack-jumper inhabits Perth, south-eastern Australia from the Eyre Peninsula to Rockhampton, and especially the island of Tasmania. The increase in numbers of these ants has been blamed on the proliferation of human dwellings and the decline of the ant’s major predator, the echidna.

Clarke first publicised the increasing problem of anaphylaxis from the sting of *M pilosula* in Tasmania in 1986. More recently, at the Royal Hobart Hospital Emergency Department, 21–25% of the 324 cases of anaphylaxis treated with epinephrine between 1990 and 1998 were attributable to *M pilosula* stings, compared with only 13% caused by honeybee stings. Between 1980 and 1999, four deaths occurred in a population of 223 000.2 Undoubtedly, the prompt use of self-injectable epinephrine could prevent most cases of severe anaphylaxis, but desensitisation offers even greater insurance against sting reactions. In people who had had anaphylactic reactions to bee and wasp stings, desensitisation, by injection of a series of increasing doses of the appropriate hymenopteran venom, was first validated in 1978.3 The mechanism of desensitisation remains uncertain but earlier notions of an induction of IgG “blocking” antibody have yielded to evidence that, in hypersensitive individuals, immunotherapy causes a switching of the abnormal Th2 cytokine response (interleukin 4 and 13), which stimulates secretion of IgE by B cells, to Th1 responses (interleukin 2 and interferon gamma), which block this production.4

The effectiveness of venom immunotherapy against hymenoptera-sting anaphylaxis is supported by clinical experience, but the most impressive evidence so far that this therapy can prevent severe life-threatening anaphylaxis is now provided by the investigation of jack-jumper stings in Tasmania, reported by Clarke and colleagues in today’s *Lancet*. The rigorous protocol of this double-blind placebo-controlled crossover trial was demanding both on its 68 highly-motivated and