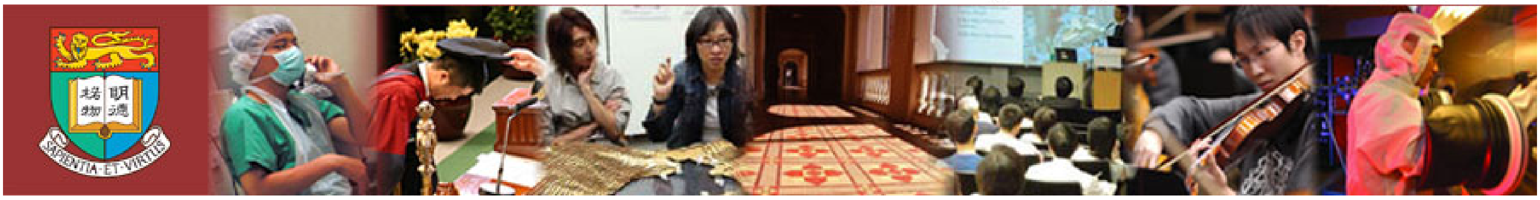


# Appraising published claims about drug treatment to implement best therapy in clinical practice

CR Kumana, IJ Lauder

**The validity and applicability of publications about individual clinical studies and systematic overviews regarding interventions with drugs need to be established and perceived in quantitative terms to implement evidence-based, best current therapy. This requires an understanding of study design, various types of bias, intention to treat analysis, clinical versus statistical significance, and other considerations. The quantitative appreciation of drug effects may be facilitated by arranging results from case-control studies, cohort studies, and controlled trials in suitable contingency tables. Relative risks, relative risk reductions, odds ratios, and absolute risk reductions (in a given period of time), as well as corresponding numbers needing treatment (to prevent one event) may then be calculated. Systematic overviews of multiple clinical trials and assessment of their combined quantitative significance (meta-analyses) were developed to enhance statistical power, to enhance the level of confidence about small differences in effect, and to reconcile conflicting claims. The results of meta-analysis are usually represented by so-called 'forest plots' of point estimates (corresponding to medians) and their respective confidence intervals, as well as a combined point estimate and confidence interval. Heterogeneity (important differences between findings from individual trials) is a special problem that arises with this relatively new tool. Meta-analyses are also specially prone to other sources of bias—a greater likelihood that trials reporting 'favourable' effects are published, covert duplicate inclusion of results from the same patients, and non-blinded meta-analysers.**

*HKMJ 1998; 4:158-68*

*Keywords: Data interpretation, statistical; Drug evaluation; Meta-analysis; Randomized controlled trials; Research design*

## Introduction

The term 'evidence-based medicine' has become a cliché, almost amounting to a slogan for medical respectability. It even conveys the impression that the generations of medical practitioners who can lay no claim to being exposed to the concept must have not provided rational therapy, yet most clinical decisions are unlikely to be based on adequate, sufficiently comprehensive, and scientifically-derived information. No self-respecting doctor knowingly undertakes management contrary to what he or she considers to be the best available evidence and act in the best interests of the patient. It is therefore important to avoid medical management decisions that are inconsistent with such evidence, and necessarily implies that a practitioner has an adequate understanding of relevant research publications and familiarity with the principles of scientific rigour. With respect to drug treatment, for most clinicians there are at least four major challenges that need to be resolved:

(1) How to assess the validity and relevance of claims from specific published studies.
(2) How to perceive drug effects in quantitative terms.
(3) How to sort out the deluge of different claims and counter-claims and extract the most appropriate information on which to base optimum treatment, in the limited time available.
(4) How to translate into clinical practice good intentions to follow what is accepted as best current therapy, against a background of entrenched past habits.

This account addresses the first three of these topics by highlighting a selection of the most salient issues (sometimes illustrated by reference to examples in the medical literature), and provides some pointers

**The University of Hong Kong, Pokfulam, Hong Kong:**
**Department of Medicine**
CR Kumana, BSc, FRCP
**Department of Statistics**
IJ Lauder, MSc, PhD

**Correspondence to: Prof CR Kumana**

on the fourth topic. In particular, it makes no attempt to deal with these subjects comprehensively or to cover statistical methodology.

## Assessing the merit of claims about drug treatment

Whether articles about drug therapy describe original research, or consist of meta-analyses, reviews or editorials, it is the responsibility of medical editors to filter the material for quality.[1] Readers should also appreciate several limitations and constraints on the ultimate selection. Firstly, acceptability for publication will depend on the ability of the editorial process to assess the relative scientific strengths and weaknesses of a submission. Secondly, papers that are currently topical or deal with the special interests of the editorial team, may be more likely to be chosen. Thirdly, there is commonly a bias against the publication of negative findings; authors are reluctant to submit such papers and journals are liable to reject them. Fourthly, it is conceivable that some journals favour submissions from certain institutions or countries. A journal's contents are also influenced by the need to maintain balance, by the perceived clinical (and scientific) importance and originality of a submitted paper, and by the supply of papers. Regrettably, many journals publish poorly-designed and poorly-executed original work with biased claims, which far outnumber soundly conducted investigations. Similarly, ill-conceived, poorly argued, and pedestrian conclusions and recommendations from systematic reviews are all too common.

A study that contains scientifically more rigorous methodology and analysis is more likely to be accepted for publication. Unfortunately, this does not mean that tenuous and possibly misleading drug therapy claims based on relatively weak evidence are never published in prestigious journals. Editors may be more disposed towards subject matter that they consider to be particularly interesting, innovative, or likely to generate controversy and discussion. Moreover, in keeping with most responses involving a biological system, results of drug trials commonly entail a degree of uncertainty. In addition, the peer review process is certainly fallible and what may appear to be the best of study designs can eventually turn out to be flawed. In such cases, the journal's correspondence columns as well as findings from further investigations of the same problem, often turn out to be the final arbiters of usefulness. Notwithstanding these cautions, when planning a clinical study involving drugs, it is useful to consider the interrelated cardinal features of the study's execution that an enlightened doctor might assess. These features have been summarised by Sackett[2] as: (1) the kind of claim being made (study design); (2) the presence of bias; (3) accounting for all patients; (4) consideration of clinical as well as statistical significance; (5) the kind of outcome analysis used; (6) similarity of one's own patients to the study patients; and (7) feasibility of using the treatment in one's own clinical setting. The first four aspects concern validity and are discussed more fully below; all seven are relevant to applicability, there being no applicability for invalid claims.

Whereas the kind of outcome analysis is closely linked to the kind of claim being made, whether the intervention of interest entails therapy or prophylaxis may also be critical. For obvious reasons, an investigation assessing the prevention of rare adverse events (eg infective endocarditis) might require such enormous patient numbers that an adequately-sized outcome study may never be feasible. Under these circumstances, surrogate markers (eg bactericidal antibiotic concentrations of different chemoprophylactic regimens) rather than reduction of event rates are resorted to. Examples of surrogate markers used in clinical studies include serum cholesterol concentration, blood pressure, and CD4 lymphocyte counts, but such studies are regarded as less satisfactory than therapeutic trials of outcome. Outcome studies should address all outcomes of interest (eg trials with lipid-lowering drugs must assess overall mortality and morbidity, and not be confined only to cardiac ischaemic events).

### *Kind of study claim*

According to their liability to less and less bias, study designs may be ranked as increasingly liable to yield valid conclusions[3-5] (Table). This ranking is an oversimplification, as it takes no account of patient numbers or meta-analyses, the type of outcomes being addressed—definite and discrete (eg death), or continuous and subjective variables (eg sense of well-being)—and a host of other factors that may unfairly influence their usefulness. Among the various types of study design, randomised controlled trials (RCTs) are generally regarded as being able to provide the most valid information on which to base clinical decisions about drug therapy or prophylaxis. Yet even RCTs are not necessarily free of bias; however, when anticipated, many biases may be controlled for by external manipulation (eg stratification and randomisation in blocks to account for age, gender, disease severity, or institution). Occasionally, RCTs (and to a lesser extent cohort studies) are also able to yield the most useful

**Table. Claims about therapeutic benefits (or adverse sequelae)**

| Freedom from bias* | Kind of claim (study design) | Comment |
|---|---|---|
| ± | Case report | Benefits plausible, if untreated outcomes typically very poor† |
| + | Case-control study | Commonly entails retrospective analysis |
| ++ | Cohort study | Commonly entails prospective analysis |
| +++ | Controlled trial | Ideally should be randomised and double-blind |

* The number of + symbols indicates relative freedom from bias, while '±' indicates the greatest liability to bias
† After case reports of successful treatment outcomes for diseases that are invariably fatal without treatment, recourse to controlled trials is not necessary (eg antibiotic treatment of infective endocarditis)

and compelling data about possible unexpected drug-related events.[6,7] However, to evaluate the possibility of rarely encountered but serious adverse effects, it may be neither feasible nor ethical to conduct an RCT. On the contrary, case-control studies are nearly always designed with the intention of assessing such adverse sequelae, although very occasionally, they too can be directed at evaluating drug benefits.[8]

### *Bias*

This is a process that tends to produce results that depart systematically from true values, and has been categorized as being due to sampling, the manoeuvre, or measurement. The terms bias, validity, and reliability are closely allied and interrelated. Thus, bias may be real or potential and has a direction (eg unfairly appearing to favour persons receiving active treatment), but the result can still yield a valid conclusion. Reliability encompasses liability to bias as well as both the precision (level of confidence) and repeatability of a given finding.

Bias is often dependent on the appropriateness of the selected controls.[3,9] The use of placebo controls (rather than the best available current therapy) is regarded as unethical if there is already a well-established, beneficial treatment. However, this argument only applies if the latter is already proven to be superior to placebo. Moreover, if such 'established' therapy actually confers more harm than benefit (as happened when certain drugs were used to suppress ventricular arrhythmias in survivors of acute myocardial infarction),[10] the superior results encountered with any new experimental drug might lead to its widespread adoption without being any better than placebo. Conversely, if an established drug therapy is of proven value, it may be unfair to adopt a new and possibly more toxic therapy, just because it was superior to placebo. Precisely this dilemma affects the NINDS-II randomised double-blind trial,[11] which reports that compared with placebo, intravenous tissue plasminogen activator confers an overall benefit in terms of death and dependence (but an increased risk of intracranial bleeds). Since the control group was denied aspirin therapy in the first 24 hours, their treatment must certainly be regarded as suboptimal. Another serious drawback of many studies is the use of historical rather than concurrent (and preferably randomised) controls.[4] Numerous biases can occur in such studies[3] including those due to: (1) treatment being non-blind; (2) variations in unsuspected confounding factors during different periods (eg climatic changes, co-treatment with other drugs, prevailing epidemics); and (3) belief in the superiority of the new treatment, such that patients offered the latest drug may be recruited on less compelling grounds and have a better prognosis than the historical controls.

Apart from isolated case reports, case-control studies are particularly susceptible to bias,[3,12] especially due to sampling. A case-control study is most commonly resorted to as an economical means of investigating possible rare adverse drug effects. For example, in a massive, multicentre, collaborative case-control study in hospitalised patients, it was reported that dipyrone (a widely used analgesic in south-east Asia) did not cause a substantial excess of patients with neutropenia.[13] But the highly symptomatic neutropenia that this agent produces (in sensitive persons) is known to be acute and transient, and this study was conducted in areas where hospitalisation was often delayed and on average, blood counts were performed 5 days after the onset of symptoms.[14] Moreover, during the acute phase of illness, patients could have died prior to being admitted to hospital or diagnosed. Under the circumstances, it is very likely that cases of such short-lived, yet potentially serious neutropenia would be rarely picked up.

### *Accounting for all patients*

Accounting for all of the patients entered in a clinical trial has become a fundamental aspect of proper data evaluation. In randomised trials, inclusion of results from all patients entered (irrespective of whether they are able to complete the intended treatment) is referred to as 'intention to treat analysis'. The importance of this principle was highlighted in the Anturane Reinfarction Trial,[15] which was terminated prematurely as patients taking the active drug sulphinpyrazone (Anturan) appeared to do better than controls. Patients who did not take their treatment for at least 7 days (including some who died) and those who refused or were unable to take it in the long term were excluded from the analysis. The result was an efficacy analysis—an assessment of the treatment effect only among patients who took the treatment as intended. In a subsequent 'intention to treat analysis' (also known as an effectiveness analysis)*, of patients who participated in either arm of the study after randomisation and regardless of subsequent events, no such beneficial impact was evident.[16] Arguably, if an excess of deaths ensued within the first week of starting the active drug, exclusion of such patients from the analysis could well bias the results in favour of active treatment. Moreover, such an analysis might also hide a high drop-out rate from active treatment, that might otherwise detract from the drug's overall effectiveness in patients to whom it was given. Under the circumstances, regardless of the results of any efficacy analysis, an 'intention to treat analysis' should always be undertaken.

### *Clinical versus statistical significance*

The distinction between clinical significance (which depends on clinical judgment) as opposed to statistical significance (an arbitrary level of uncertainty deemed to indicate that an observed difference is not due to chance), is a matter of common sense.[2] By and large, the higher the level of a clinically significant difference ($\delta$ value) that is likely to be encountered, the smaller the number of patients necessary to yield a statistically significant difference. Thus, the desired minimal level of clinical significance is one of the important parameters used by investigators to determine the power of a study—that is, the patient numbers required to yield statistically significant results.

For example, in an RCT entailing antihypertensive therapy, patients treated with a specific drug might achieve average blood pressures that are minimally lower than in those receiving alternative medication. A very small difference might not be judged clinically significant, such that the effort expended in switching the treatment to achieve any minimal theoretical benefit may not be regarded as worthwhile. On the contrary, such a result could very well be statistically significant, provided the trial entailed sufficient patient numbers and/or the treatment responses exhibited a small enough variance. The Medical Research Council's RCT involving more than 17000 patients with mild to moderate hypertension illustrates many aspects of these principles.[17] On average, systolic and diastolic blood pressures encountered over the ensuing years were about 10 mm Hg lower in persons taking bendrofluazide or propranolol than in those treated with placebo. This difference was statistically significant and regarded as clinically significant. In contrast, in all age groups and both sexes, the average blood pressures achieved with thiazide treatment were consistently 2 to 3 mm Hg lower than with propranolol therapy, but the latter differences may not be regarded as clinically significant. Arguably, in contrast to clinical outcome, surrogate markers such as blood pressure should never be regarded as clinically important. But even in terms of outcome, there was a statistically significant difference (stroke rates per 1000 patient years of treatment with bendrofluazide and propranolol being 0.8 and 1.9, respectively; P=0.002), which may nevertheless be regarded as clinically unimportant.

## Quantifying drug effects in clinical studies

In a case-control study, cohort study, or randomised intervention trial, when comparing discrete (as opposed to continuous) binary outcomes associated with active treatment and control intervention, it is useful to construct a 2x2 contingency table (Box 1). The best way of expressing the results can then be appraised, according to the type of investigation under consideration. The corresponding terms are explained in the table legend[18-20] and should normally be stated together with their corresponding confidence intervals. For the purposes of clarification, they are further illustrated and discussed in relation to well-known, recently published studies (Boxes 2 and 3). Since randomised trials and cohort studies both begin with the respective interventions, it is appropriate to approach the analysis from the same perspective and express the results either as absolute risk reductions (ARR) in a given period of time, relative risk (RR) or relative risk reduction (RRR). The deleterious effects of an intervention (whether determined from cohort studies or randomised trials) tend to be reported as RRs, and benefits tend to be expressed as RRRs. An alternative and possibly more meaningful index of drug effect is the number needing to be treated (NNT) to prevent (or cause) one event in a given period of time.

As opposed to studies that involve measurement of continuous variables, this account concentrates on clinical trials that have definitive binary end-points (eg

* A more pedantic definition of effectiveness might be 'the benefits accruing from a given intervention in conditions closely resembling actual clinical practice' but these could hardly be encountered in the setting of a clinical trial.

**Box 1. Application of a 2x2 contingency table for the systematic analysis of binary data**

| | | | Case-control studies<br>↓<br>Patient numbers for adverse outcomes | | Row totals |
|---|---|---|---|---|---|
| | | | Occurred | Did not occur | |
| Randomised trials | → | Test drug | a | b | (a+b) |
| Cohort studies | | Control drug | c | d | (c+d) |
| | | Column totals | (a+c) | (b+d) | |

Absolute risk reduction (ARR) $= (c / [c+d]) - (a / [a+b])$ over n years

Relative risk (RR) $= \dfrac{(a / [a+b])}{(c / [c+d])}$

Relative risk reduction (RRR) $= \dfrac{(c / [c+d]) - (a / [a+b])}{(c / [c+d])}$

Number needing to be treated (NNT) $= \dfrac{1}{(c / [c+d]) - (a / [a+b])}$ for n years

Odds ratio (OR) or risk ratio $= \dfrac{(a / c)}{(b / d)}$

**Box 2. West of Scotland Coronary Prevention Trial with Pravastatin[23]**

| Treatment | Fatal and non-fatal myocardial infarctions over the trial period (5 years, men)<br>Occurred | | Did not occur | | Total | |
|---|---|---|---|---|---|---|
| Pravastatin | 174 | a | 3128 | b | 3302 | (a+b) |
| Placebo | 248 | c | 3045 | d | 3293 | (c+d) |

The crude parameters shown below are unadjusted and are therefore not exactly as published.

ARR $= (c / [c+d]) - (a / [a+b]) = (248 / 3293) - (174 / 3302)$
$= 0.0287$ (ie reduction in risk of event per man treated for 5 years)

RR $= \dfrac{(a / [a+b])}{(c / [c+d])} = \dfrac{(174 / 3302)}{(248/ 3293)} = 70\%$

RRR $= \dfrac{(c / [c+d]) - (a / [a+b])}{(c / [c+d])} = \dfrac{(248 / 3293) - (174 / 3302)}{(248 / 3293)} = 30\%$

NNT $= \dfrac{1}{\text{Reduction in event risk per man treated for 5 years}} = \dfrac{1}{\text{ARR}}$

$= \dfrac{1}{(c / [c+d]) - (a / [a+b])} = \dfrac{1}{(248 / 3293) - (174 / 3302)} = 44$ men treated for 5 years

death or survival) and the relevant parameters (eg RRRs, NNTs, or odds ratios [ORs]) used for statistical analysis. One reason for this emphasis is that these end-points often entail hard outcomes (eg death, stroke, or myocardial infarction) in contrast to continuous variables that commonly involve measurement of surrogate markers (eg blood pressure or serum cholesterol). Thus, conclusions based on studies of such discrete (typically binary) outcomes usually constitute superior grounds for evidence-based clinical decisions. It is possible, however, to apply similar principles to individual studies or meta-analyses in which results are expressed as continuous variables (eg defined health quality adjusted survival times)[21] or as other categorical variables. For the sake of brevity, and as they are rather technical, corresponding methods of statistical modelling[22] are not reviewed here.

While RRRs (and RRs) are independent of the size of each treatment group and are commonly regarded as independent of treatment duration, they nevertheless enable clinicians to discern the relative benefit or risk

**Box 3. Case-control study of myocardial infarction associated with antihypertensive drug therapy**[24]

| | Treated hypertensive patients in a group health cooperative (1986-1993) | | | | | |
|---|---|---|---|---|---|---|
| Treatment | Cases of first MI* | | Controls | | Total | |
| CCBs† | 80 | a | 230 | b | 230 | (a+b) |
| No CCBs | 255 | c | 1165 | d | 1420 | (c+d) |
| Total | 335 | (a+c) | 1395 | (b+d) | | |

The crude ORs calculated below do not correspond exactly to those in the original paper (which were adjusted for other factors).

$$\text{OR (also called risk or hazard ratio)} \quad = \frac{(a / c)}{(b / d)} = \frac{(80 / 255)}{(230 / 1165)} = 1.59$$

*MI myocardial infarction
†CCB calcium channel blocker

associated with alternative treatments. These terms do not reveal the effort that must be expended (over a fixed period of time) to bring about a given absolute effect. They therefore hide very important information that could influence a clinician's decision to start treatment.[25] In Box 2 for instance, if the total number of patients treated with pravastatin and placebo had been 10-fold smaller (330 and 329, respectively) and the number of events encountered remained unchanged, the RRR would also have been unchanged. Obviously, a very much more powerful absolute drug effect might be obscured but would be apparent when using NNT (reciprocal of ARR per person, per fixed period of time) as the calculated value changes drastically from 44 to 4.4 men treated over 5 years.

$$\text{RRR} = \frac{(248 / 329) - (174 / 330)}{(248 / 329)} = 30\%$$

$$\text{NNT} = \frac{1}{(248 / 329) - (174 / 330)} = 4.4 \text{ men in 5 years}$$

Conversely, if an individual's baseline (untreated) risk fails to match that of the average clinical trial patient, the NNT would also be affected; halving the risk would double the corresponding value without necessarily changing the RRR.[26] In addition, apart from emphasising the effort involved in preventing (or inducing) a single event over a given period, the NNT figure helps clinicians to quantify corresponding costs and the likely number of individuals that would be exposed to potential adverse drug effects.[27]

In contrast, case-control studies assemble the data and approach the analysis from the perspective of outcome rather than treatment. It is therefore more appealing to express the findings as an OR, sometimes referred to as a risk or hazard ratio.† Case-control studies are not amenable to estimation of the true RR directly, as the cases and controls are preselected. Odds ratios are also favoured for reporting meta-analysis. This probably ensued after it was realised that it was possible to satisfactorily combine the results from several subgroups of the same case-control study and produce a summative OR, even when such subgroups were regarded as unbalanced due to confounding factors.[20] Fortunately, it is evident that with a little algebraic rearrangement, ORs approached from the perspective of the treatments ie the ratio of the outcome odds to the exposures (as might be reasonable for randomised trials and cohort studies) produce the same expression, viz:

$$\text{From the perspective of outcomes } \frac{(a / c)}{(b / d)} = \frac{ad}{bc}$$

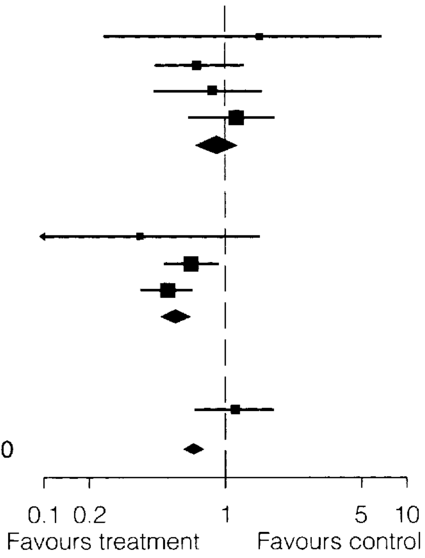
$$\text{From the perspective of treatments } \frac{(a / b)}{(c / d)} = \frac{ad}{bc}$$

Thus, findings from all three types of study (RCTs, cohort studies, and case-control studies) are commonly expressed as ORs.

In randomised trials or cohort studies, however, RRs and ORs should not necessarily be equated, as odds invariably have smaller denominators than do risks, which gives rise to larger values and might lead to overexaggerated impressions of benefit (RRR) or harm (RR).[28] In addition, the concepts of odds and ORs are intuitively difficult to grasp since they can have values from 0 to infinity, whereas RRs are always ≤100%. Any resulting apparent discrepancy is particularly likely if such a trial has small patient numbers and/or the event risk in either treatment arm is great (>20%).

† In the context of ratios, the terms 'odds', 'risk', and 'hazard' are commonly used interchangeably, although the term 'hazard ratio estimator' has a quite different meaning.

| Study | No. of cases / total |  | Peto odds ratio (95% confidence intervals) |
|---|---|---|---|
|  | Thrombolysis group | Control |  |
| Streptokinase vs control |  |  |  |
| Morris 1995 | 6/10 | 5/10 | 1.47 (0.26-8.18) |
| MAST-I 1995 | 97/157 | 106/156 | 0.76 (0.48-1.21) |
| MAST-E 1996 | 124/156 | 126/154 | 0.86 (0.49-1.51) |
| ASK 1996 | 84/174 | 74/166 | 1.16 (0.76-1.78) |
| Subtotal | 311/497 | 311/486 | 0.94 (0.72-1.24) |
| $\chi^2$=2.06 (df=3), Z=0.41* |  |  |  |
| tPA vs control |  |  |  |
| Mori 1992 | 11/19 | 10/12 | 0.32 (0.07-1.48) |
| ECASS 1995 | 198/313 | 220/307 | 0.68 (0.49-0.95) |
| NINDS 1995 | 179/312 | 229/312 | 0.49 (0.35-0.69) |
| Subtotal | 388/644 | 459/631 | 0.57 (0.45-0.72) |
| $\chi^2$=2.38 (df=2), Z=4.73 |  |  |  |
| Streptokinase + aspirin vs aspirin |  |  |  |
| MAST-I 1995 | 99/156 | 94/153 | 1.09 (0.69-1.73) |
| **Total** | **798/1297** | **864/1270** | **0.75 (0.63-0.88)** |
| $\chi^2$=15.08 (df=7), Z=3.47; P<0.05 |  |  |  |



Adapted, with permission.[32]

* $\chi^2$ refers to test for heterogeneity across different trials; Z is test statistic for odds ratio

**Fig. Effect of thrombolysis on death or dependency at end of trial follow-up (subsidiary and overall meta-analyses)**

Odds ratios (95% confidence intervals) are shown for individual trials (the area of each square is proportional to the amount of information contributed), and for subtotals and total (each diamond represents odds ratio and 95% confidence interval). Note that the ASK and NINDS trials have non-overlapping confidence intervals, indicative of heterogeneity

In contrast, as case-control studies are typically used to study rare treatment effects in large numbers of patients, this sort of distortion between ORs and RRs becomes negligible. Other drawbacks of ORs include the fact that they cannot be used to determine NNTs and are not very user-friendly. For all of these reasons, the use of ORs should generally be confined to the reporting of case-control studies and meta-analyses.

## Systematic overviews of multiple clinical trials and meta-analysis

To distil the results of numerous clinical trials and extract the most important information for optimum treatment, it is necessary to refer to systematic reviews.[29-31] Such a review consists of a comprehensive overview of all accessible primary studies on a given topic. These are selected independently of the result using unambiguous criteria and are assessed according to an explicit, reproducible methodology entailing objective measures of outcome. A meta-analysis maybe regarded as a type of systematic review, where there is a mathematical synthesis of the results of two or more primary studies addressing the same hypothesis in the same way. Unlike the RCT, meta-analysis is a relatively new tool and its limitations are only just being understood. The method is expected to undergo considerable refinement and the precision and reliability of any final conclusions derived from such analyses are set to improve. It is also entirely feasible that the process of planning, executing, and analysing multicentre mega-trials and meta-analyses will converge.

In the context of clinical drug trials, the meta-analysis describes the amalgamation of several trials assessing the outcome of the same alternative treatments for a specific condition. Typically, the results of such a meta-analysis are presented in tabular form and as a 'forest plot'. The latter entails plotting individual point estimates of the ORs for each trial with corresponding confidence limits around the line of unity (no effect); the pooled results are represented by a small diamond (Fig).[32] If the diamond overlaps the line of unity, the null hypothesis that the effects of the alternative treatments do not differ is not rejected. The rationale for combining the data from diverse trials is to give a more efficient estimate of any difference in treatment effects. Situations where such amalgamation may be considered beneficial include:

(1) small trials lacking the statistical power to yield a significant result (ie the ability to confidently

detect a genuine difference when it exists) whereas a combination of such trials could detect a real difference between treatments;

(2) when it is necessary to detect a small but clinically significant difference with greater confidence (eg a small difference in death rates); and

(3) when considering how to balance evidence for and against a particular treatment.

From a statistical viewpoint, analysis of aggregated data must be treated with caution due to several possible sources of bias. Various forms of bias that are known to affect meta-analysis and how they may be anticipated and minimised are discussed below.

### *Heterogeneity between trials*

In the context of meta-analysis, 'heterogeneity' between individual clinical trials[30] is usually inferred if there is no overlap between their respective 95% confidence intervals (Fig). Commonly, a variant of the $\chi^2$ statistic is used as a more definitive test. It then becomes incumbent to account for such mathematical heterogeneity on clinical grounds. The problem most often arises due to variations in the methodology both within and between trials, especially with respect to sampling and measured outcomes. Differences that could be responsible include between trial variations in terms of blinding; patient inclusion/exclusion criteria and average age; and the use of current versus historical controls. For a comparison of two treatments, the fixed sample size group comparison design is common and combining trials of the same design appears reasonable. It is not clear how more sophisticated trials, such as those involving group-sequential or fully sequential designs (whose continuation depends on the outcome of interim/continuous analysis) can be included.[33] In the modelling of meta-analysis, it is essential that any possible sources of heterogeneity must be taken into account (eg working with ORs of individual trials rather than an aggregated composite OR). If this is not done, the accuracy of any estimate of the differences between treatment responses is liable to be spurious. To tackle variability both within trials and between trials, a variety of hierarchical modelling processes have been used,[34,35] depending on whether the outcomes of interest were discrete or continuous variables.

A further statistical consideration for hierarchical modelling relates to within-trial parameters of the expected response, which may be regarded as fixed or random.[36,37] According to statistical terminology, in the former type of response modelling, each trial is taken to have a fixed observable mean—sampled from the between-trial model. In the case of random effects, each trial is also considered to have an expected mean response generated from the between-trial model, but the more extreme results are given less value. By conferring relatively less weight to such trials with outlying results, the random effects model makes the distribution of estimated responses appear less erratic. When both models fit the data well, there is little difference between the two models in terms of estimating treatment differences. When there are differences, however, lack of fit of the fixed effects model is usually responsible and commonly due to one or more influential trials with outlying data. Consequently, the random effects model has become more popular, particularly with the advent of powerful computers; but its uncritical use must be avoided. Data should be inspected for trials with outlying responses that strongly influence the overall findings. One way of monitoring the meta-analysis for this effect is to leave out each trial in turn and compare the results with the overall analyses of all trials. More technical consideration of this topic is covered in several reviews.[22,33]

### *Publication bias*

By definition, meta-analysis is performed using published results from trials in a specific area. Such publications are not necessarily a representative sample of all the relevant trials carried out; those yielding negative or non-significant findings may not be submitted or accepted for publication. Meta-analyses of published trials therefore tend to be biased in favour of overestimating treatment differences, which is why there is a move towards including appropriate unpublished data in such reviews.[38-40] Presumably, because smaller trials yielding extreme results are particularly impressive, they are especially prone to such bias and liable to affect corresponding meta-analyses.[41]

### *Covert duplicate publication*

While undertaking a meta-analysis of published reports of RCTs with ondansetron as a postoperative anti-emetic, the researchers noted that data from nine trials was duplicated in 14 others[42]; only one of the latter reports entailed any cross-referencing. Moreover, just as trials with positive findings are more likely to be published, they are also more likely to be duplicated. Thus, a meta-analysis of such duplicated findings would overestimate the drug's anti-emetic effect.

### *Non-blinded meta-analysis*

It makes sense that readers who evaluate retrospective data for the purposes of meta-analysis be blinded (at least initially) to the identity of the treatment groups in each trial and the respective authors and their affiliations. One study found, however, that randomly blinding some of the readers who were evaluating a group of clinical trials had no significant impact on the final summary OR.[43] In addition, the process of blinding was extremely time-consuming. Nevertheless, as the readers participating in this exercise may well have been aware of its intent, their meta-analysing behaviour might have been more exemplary than usual and, interestingly, the trials selected for inclusion in the analysis by the blinded and unblinded readers revealed considerable disagreement.

***Meta-analysis or large randomised controlled trials***

The efficacy of meta-analysis outcomes has been assessed by comparing the results with those of a single, large, RCT that addresses the same question. According to one view, unexplained clinically important differences are extremely uncommon, even when the meta-analyses depend on trials with small patient numbers.[44] This conclusion has been contested and continues to be hotly debated in the literature.[40,41,45,46] The detractors of meta-analyses draw attention to glaring inconsistencies between conclusions drawn from the latter and from individual mega-trials. In one study of the problem,[45] meta-analysis might have led to the adoption of an ineffective treatment in about one third of the instances and the rejection of effective therapy in a similar proportion. However, these arguments depended on the selective use of certain meta-analyses and accepting the premise that randomised controlled mega-trials constitute the gold standard. Notwithstanding these points, a way of anticipating such anomalies has been proposed.[47]

## Towards implementing evidence-based best therapy in clinical practice

Any attempt to implement 'evidence-based' best treatment entails balancing decisions entailing the following three components[48]:

(1) findings from clinically relevant, current, and methodologically sound research, which must usually be extracted from papers published in medical journals;
(2) the clinical experience and expertise of the doctor(s), to cope with the decisional aspects for which there is no direct evidence; and
(3) the whims and individual preferences of patients, families, guardians, and doctors.

With reference to the last two points, real-life situations may sometimes dictate a trial and error approach in individual patients. To enhance the objectivity and evidence base of such efforts, more formal controlled trials in individual patients (so-called 'n of 1 trials') are undergoing development, and their value and precise role are currently being debated.[49,50]

In clinical practice, increasing reliance on the first component as opposed to the second and third is a major challenge. As a first step, it implies an ability to recognise best evidence (from scientific publications). This must be followed by the formulation of unambiguous and feasible treatment guidelines or clinical policies that are appropriate to the local circumstances of the patients under consideration. Regrettably, the acceptance of new policies through educational efforts alone (conducting seminars, issuing clinical practice guidelines or handouts) tends to be limited and transient.[51-54] One approach to overcome this impasse is to impose restrictive regulations (formularies, automatic stop orders or substitution of drugs) on prescribers. Although implemented to varying degrees in many health care environments, these practices are often regarded as an unwarranted interference in clinical autonomy and are liable to stifle initiative. Non-restrictive measures that are effective, usually incorporate education together with other facilitative manoeuvres, such as feedback or concurrent review.[55-57] To be successful and cost-effective, such strategies need to be focused, simple and non-disruptive, ongoing, capable of clearly identifying and targeting specific prescribers, and perceived as emanating from and belonging to the end-users.[58,59]

## References

1. Royal Statistical Society. Checklists for science? RSS News 1993;20:1-2.
2. Sackett DL. How to read clinical journals: V: To distinguish useful from even harmful therapy. CMAJ 1981;124:1156-62.
3. Sackett DL. Bias in analytic research. J Chronic Dis 1979;32: 51-63.
4. Friedman LM, Furberg CD, DeMets DL. Basic study design. In: Fundamentals of clinical trials. Boston: Wright, 1981; 28-39.
5. Yusuf S, Cairns JA, Camm AJ, Fallen EL, Gersh BJ. Grading of recommendations and levels of evidence used. In: Yusuf S, Cairns JA, Camm AJ, Fallen EL, Gersh BJ, editors. Evidence-based cardiology, London: BMJ Books 1998;26.
6. Green KG. Improvement in prognosis of myocardial infarction by long-term beta-adrenoceptor blockade with practolol. A multicentre international study. BMJ 1975;2:735-40.
7. Kumana CR, Chan GT, Yu YL, Lauder IJ, Chan TK, Kou M. Investigation of intravascular haemolysis during treatment of acute stroke with intravenous glycerol. Brit J Clin Pharmacol 1990;293:47-53.
8. Gyorkos TW, Svenson JE, Maclean JD, Mohamed N, Remondin MH, Franco ED. Compliance with antimalarial prophylaxis and the subsequent development of malaria: a

matched case-control study. Am J Trop Med Hyg 1995;53: 511-7.
9. Rothman KJ. Placebo mania. BMJ 1996;313:3-4.
10. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: effect of encanide and flecanide on mortality in a randomised trial of arrhythmia suppression after myocardial infarction. N Engl J Med 1989; 321:406-12.
11. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. N Engl J Med 1995;333:1581-7.
12. Hutchison GB, Rothman KJ. Correcting a bias? [editorial]. N Engl J Med 1978;299:1129-30.
13. The International Agranulocytosis and Aplastic Anemia Study Group: Risks of agranulocytosis and aplastic anemia. A first report of their relation to drug use with special reference to analgesics. JAMA 1986;256:1749-57.
14. Symposium of non-narcotic analgesics today, benefits and risks. In: Brune K, Santoso B, Brogden RN, editors. Discussion of section 2: adverse reaction assessment. Med Toxicol 1986;1(1 Suppl):93S-94S.
15. The Anturane Reinfarction Trial Research Group. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. N Engl J Med 1980;302:250-6.
16. Temple R, Pledger GW. The FDA's critique of the Anturane Reinfarction Trial. N Engl J Med 1980;303:1488-92.
17. Medical Research Council Working Party. MRC trial of treatment of mild hypertension: principal results. BMJ 1985; 291:97-104.
18. Sackett DL. On some clinically useful measures of the effects of treatment. Evidence-Based Medicine 1996;1:37-8
19. Sackett DL, Cook RJ. Understanding clinical trials. BMJ 1994;309:755-6.
20. Sackett DL. Down with odds ratios! Evidence-Based Medicine 1996;1:164-6.
21. Gelber DG, Lenderking WR, Cotton JD, et al. Quality of life evaluation in a clinical trial of ziduvudine therapy in patients with mildly symptomatic HIV infection. Ann Intern Med 1992; 16:961-6.
22. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomised clinical trials. Stat Med 1991; 101:665-77.
23. The West of Scotland Coronary Prevention Study Group. Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. N Engl J Med 1995;333:1301-7.
24. Psaty BM, Heckbert SR, Koepsell TD, et al. The risk of myocardial infarction associated with antihypertensive drug therapies. JAMA 1995;274:620-5.
25. Bucher HC, Weinbacher M, Gyr K. Influence of method of reporting results on decision of physicians to prescribe drugs to lower cholesterol concentration. BMJ 1994;309:761-4.
26. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995;310:452-4.
27. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Deciding whether your treatment has done harm. In: Clinical epidemiology. Boston: Little Brown & Co. 1991:283-301.
28. Davies HT, Crombie IK, Tavakoli M. When can odds ratios mislead? BMJ 1998;316:989-91.
29. Peto R. Why do we need systematic overviews of randomized trials? Stat Med 1987;6:233-40.
30. Greenhalgh T. Papers that summarise other papers (systematic reviews and meta-analyses). BMJ 1997;315:672-5.
31. Chalmers I, Altman DG, editors. Systematic reviews. London: BMJ Publishing Group, 1995.
32. Wardlaw JM, Warlow CP, Counsell C. Systematic review of evidence on thrombolytic therapy for acute ischaemic stroke. Lancet 1997;350:607-14.
33. Hughes MD, Freedman LS, Pockock SJ. The impact of stopping rules on heterogeneity of results in overviews of clinical trials. Biometrics 1992;48:41-53.
34. Collins R, Gray R, Godwin J, Peto R. Avoidance of large biases and large random errors in the assessment of moderate treatment effects: the need for systematic overviews. Stat Med 1987; 6:245-50.
35. Early Breast Cancer Trialists Collaborative Group. Treatment of early breast cancer. Vol 1. Worldwide evidence 1985-90. Oxford: Oxford University Press, 1990:12-18.
36. Greenland S. A critical look at some popular meta-analytic methods. Am J Epidemiol 1994;140:290-6.
37. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. JRSSA 1988;151:419-63.
38. Fullerton-Smith I. How members of the Cochrane Collaboration prepare and maintain systematic reviews of the effects of health care. Evidence-Based Medicine 1995;1:7-8.
39. Meta-analysis under scrutiny [editorial]. Lancet 1997;350:675.
40. Naylor CD. Meta-analysis and the meta-epidemiology of clinical research. BMJ 1997;315:617-9.
41. Pogue J, Yusuf S. Overcoming the limitations of current meta-analysis of randomised controlled trials. Lancet 1998;351: 47-52.
42. Tramèr MR, Reynolds DJ, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. BMJ 1997;315:635-40.
43. Berlin JA on behalf of University of Pennsylvania Meta-analysis Blinding Study Group. Does blinding of readers affect the results of meta-analyses? Lancet 1997;350:185-6.
44. Cappaleri JC, Ionnadis JP, Schmid CH, et al. Large trials vs meta-analysis of smaller trials: how do their results compare? JAMA 1996;276:1332-8.
45. LeLorier J, Grégoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med 1997;337: 536-42.
46. Bailar JC. The promise and problems of meta-analysis. N Engl J Med 1997;337:559-61.
47. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple graphical test. BMJ 1997;315: 629-34.
48. Haynes RB, Sackett DL, Gray JM, Cook DJ, Guyatt GH. Transferring evidence from research into practice: 1. The role of clinical care research evidence in clinical decisions. Evidence-Based Medicine 1996;1:196-8.
49. Johannessen T. Controlled trials in single subjects: 1. Value in clinical medicine. BMJ 1991;303:173-4.
50. Lewis JA. Controlled trials in single subjects. 2. Limitations of use. BMJ 1991;303:175-6.
51. D'Eramo JE, DuPont HL, Preston GA, Smolensky MH, Roht LH. The short- and long-term effects of a handbook on antimicrobial prescribing patterns of hospital physicians. Infect Control 1983;4:209-14.
52. Grimshaw JM, Russell IT. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. Lancet 1993;342:1317-22.
53. Jones S, Pannell J, Barks J, et al. The effect of an educational program upon hospital antibiotic use. Am J Med Sci 1977;273: 79-85.
54. Schroeder SA, Myers LP, McPhee SJ, et al. The failure of physician education as a cost containment strategy. JAMA 1984;252:225-30.

55. Feely J, Chan R, Cocoman L, Mulpeter K, O'Connor P. Hospital formularies: need for continuous intervention BMJ 1990;300:28-9.
56. Heineman HS, Watt VS. All-inclusive concurrent antibiotic usage review: a way to reduce misuse without formal controls. Infect Control 1986;7:168-71.
57. Soumerai SB, Avorn J, Taylor WC, Wessels M, Maher D, Hawley SL. Improving choice of prescribed antibiotics through concurrent reminders in an educational order form. Med Care 1993;31:552-8.
58. Kopelman RE. Performance feedback: objective indicators. In: Kopelman RE. Managing productivity in organizations—a practical, people-oriented perspective. Singapore: McGraw-Hill International, 1986:67-82.
59. Seto WH, Ching TY, Kou M, Chiang SC, Lauder IJ, Kumana CR. Hospital antibiotic prescribing successfully modified by 'immediate concurrent feedback'. Br J Clin Pharmacol 1996;41: 229-34.