Review

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.[1,2] 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.[3,4] 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.[5,6] 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.[2] 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;[2,7] 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[8] or cholesterol[9]).

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.[1] 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[2] (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

Panel 1: **Main sources of error in epidemiological studies of the effects of treatment**

Systematic errors

- Biases due to the differences in outcome caused by factors other than the treatment being investigated
- Frequent problem in the interpretation of observational studies
- Can cause either overestimation or underestimation of treatment effects
- Difficult to determine size or direction of bias

Random errors

- Impact of chance on comparisons of outcome between those who did and did not receive the treatment
- Frequent problem in the interpretation of clinical trials
- Can prevent real effects of treatment being detected or their size being estimated reliably
- Easily quantified

Lancet 2001; **357:** 373–80

Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Clinical Medicine, Radcliffe Infirmary, Oxford OX2 6HE, UK (Prof R Collins MBBS)**; and Institute for International Health, University of Sydney, Sydney, New South Wales, Australia** (Prof S MacMahon PhD)

Correspondence to: Prof Rory Collins

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
  (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)

large numbers of outcomes by doing large individual studies and large meta-analyses[2] (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.[1]

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.[1]

CLINICAL TRIALS: Minimising both systematic and random errors

Avoidance of moderate systematic 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.[10] 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.[11] For example, in a study comparing amniotomy (rupture of membranes) versus oxytocin for induction of labour that was described as randomised, treatment allocation was actually based on whether the woman's date of birth was odd or even. Foreknowledge of this led to women with an "unripe" cervix being far less likely to be recruited if they were to have been allocated amniotomy (ie, had an odd date of birth; table 1).[12] Similarly, in the Captopril Prevention Project (CAPPP) trial,[13] envelopes containing the antihypertensive treatment allocation could be opened before patients were irreversibly entered in the study, and—presumably as a consequence—there were highly significant differences in pre-entry blood pressure (and other characteristics) between the treatment groups, which might have introduced bias.[14]

Cervical "ripeness"	Amniotomy: odd dates of birth (n=110)	Oxytocin: even dates of birth (n=113)
Least	7	28
Intermediate	58	56
Most	45	29

Comparison of the distribution between treatment groups of cervical ripeness before treatment allocation: χ^2_1=16·1 (p<0·0005).

Table 1: **Imbalance in patients' characteristics between treatment groups due to foreknowledge of treatment allocation: trial of amniotomy or oxytocin for induction of labour[12]**

Studies in which treatment has not been properly allocated at random do not necessarily provide misleading evidence about the effects of treatment.[15,16] For example, in the Salk polio vaccine studies of the 1950s,[17] 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.[11,18] 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).[19] 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.[1]

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

Type of study	Poliomyelitis cases/total (rate per 100 000)		Odds ratio (95% CI)
	Vaccine	Control	
Non-randomised	60/231 902 (26)	391/725 173 (54)	0·55 (0·44–0·68)
Randomised*	57/200 745 (28)	142/201 229 (71)	0·43 (0·32–0·56)

*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[17]**

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).[20] 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.[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.[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%.[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 the 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 8587 patients allocated aspirin *vs* 1016 among 8600 allocated placebo-control; proportional reduction of 23% [SD 4]; p<0·000001).[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 effects 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).[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.[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.[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.[22]

Astrological birth sign	Vascular death by 1 month		p
	Aspirin	Placebo	
Libra or Gemini	150 (11·1%)	147 (10·2%)	0·5
All other signs	654 (9·0%)	869 (12·1%)	<0·0001
Any birth sign	804 (9·4%)	1016 (11·8%)	<0·0001

Table 3: **Unreliability of "data-dependent" subgroup analyses: ISIS-2 trial of aspirin among over 17 000 patients with suspected acute myocardial infarction[23]**

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).[7,27] 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)[28] 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).[22] 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[29] could lead to the dangerously misleading conclusion that antiplatelet therapy is not beneficial for such patients.[30] 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[30] is not supported by a meta-analysis of all such trials in patients with ventricular dysfunction.[31]

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

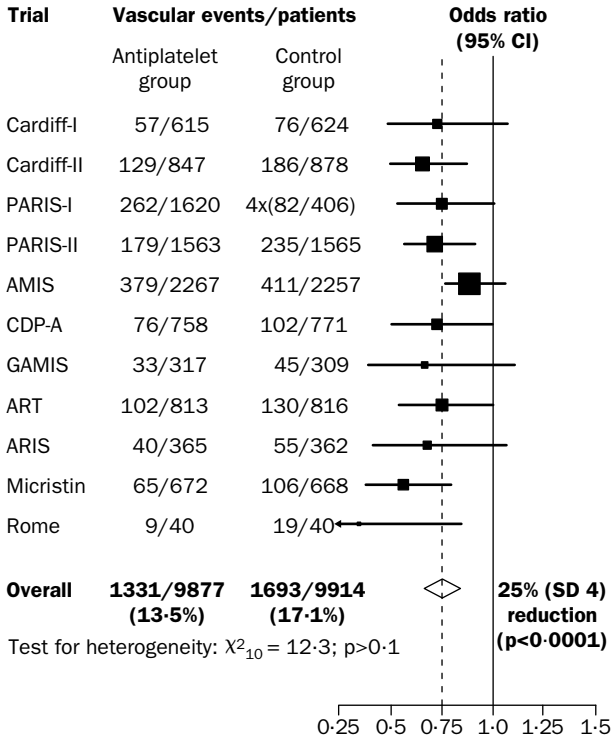

Figure 1: **Clear demonstration of worthwhile benefits in meta-analysis of available trial data, by contrast with failure of individual trials to provide convincing evidence**

Vascular events (ie, death, myocardial infarction, or stroke) in collaborative meta-analysis of trials of prolonged antiplatelet therapy after myocardial infarction.[22] Numbers in the control group of one trial with a deliberately uneven allocation have been adjusted so that the overall numbers allocated antiplatelet 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.

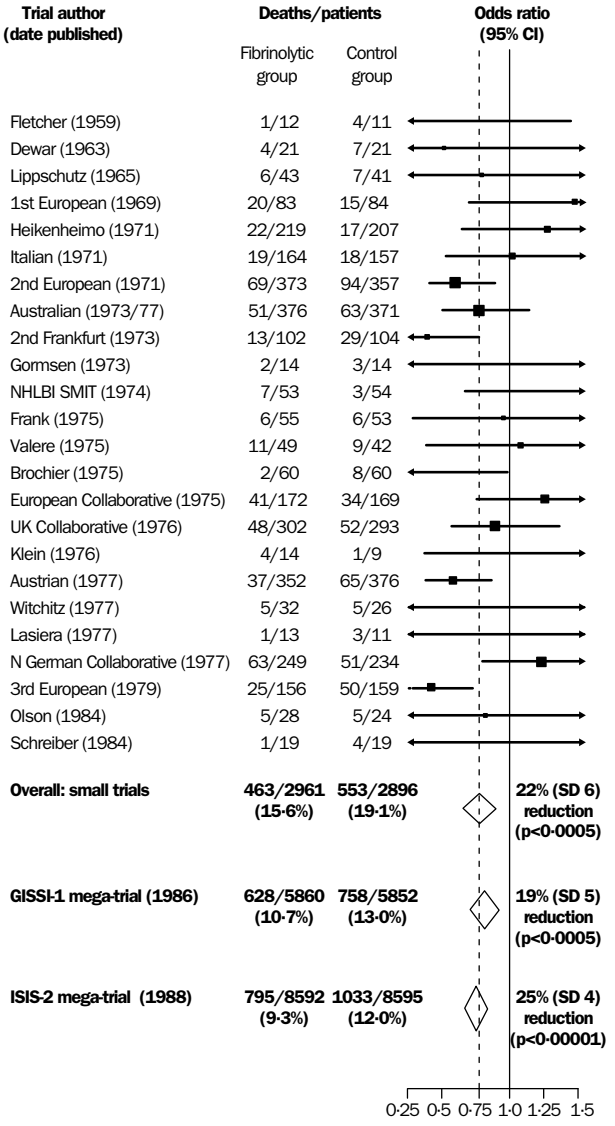

Figure 2: **Clear demonstration of worthwhile benefits in mega-trials, by contrast with failure of previous much smaller trials**

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.[34] Conventions as in figure 1.

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 who would otherwise have been classified as "stage I" will be reclassified as "stage II", biasing any comparisons with women in the lumpectomy alone group for whom the staging was less careful.[32] Similarly, in randomised trials of treatment with 3-hydroxy-3-methylglutaryl-coenzyme A reductase inhibitors (statins) versus no such treatment, comparisons between the coronary disease rates seen among patients who achieved large cholesterol reductions and those who achieved small reductions[33] are potentially biased. For, groups of patients defined by the difference in post-randomisation cholesterol-lowering response to treatment cannot be guaranteed—and, indeed, are unlikely—to differ only randomly from each other (eg, factors related to the apparent biochemical response might also be related to outcome). Hence, inferences drawn from such non-randomised comparisons of "responders" versus "non-responders" could be seriously misleading.

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).[2] 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.[34] 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-1[24] and ISIS-2[23] "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

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).[35] By contrast, the effects in several subsequent, much larger, randomised trials[36,37] 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).[38] 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).[39] 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,[40,41] of intravenous nitrates[42,43] or magnesium[44,45] for the emergency treatment of heart attacks, of heparins[46,47] or calcium antagonists[48,49] for the emergency treatment of strokes, and of vitamin E for the prevention of coronary disease.[50–52]

There are several possible explanations for such discrepancies. One theoretical possibility is that the benefits of the treatment are confined to particular categories of patients that were selectively included in the small trials. Often this possibility can be investigated by separate analyses in these selected subgroups within the large randomised trials: for example, with antiplatelet therapy in pre-eclampsia, such analyses did not identify any particular category of woman in which the effects were as great as those reported in the small trials.[36] Similarly, when considering the disappointing results of a large trial of extracranial-intracranial bypass surgery for symptomatic carotid stenosis,[53] some neurosurgeons suggested that the findings might not be generally relevant because many of the patients thought to benefit from surgery had been excluded from the trial.[54] But, when those categories of patient were defined, it could be shown that the results within the trial among such patients were no more promising.[55] Another, perhaps more plausible, explanation for the failure of large trials to confirm reports of extreme results in selected small trials is that other small trials with unpromising results might be less likely to have been published because they were less remarkable[56] (eg, for pre-eclampsia, at least as many women had been randomised in other small trials of antiplatelet therapy that had not had their results published[36]). Moreover, since large trials of a particular question are often done because the results of initial small trials (or small meta-analyses) are extremely promising, the "hypothesis-generating" trial evidence might well provide an overestimate of the real effects of treatment (especially if those trials were stopped prematurely because of extreme results[57]), whereas the subsequent large trials would not.

Generalisation from clinical trials to clinical practice

Clinicians are used to dealing with individual patients, and might feel that the results of trials somehow deny their individuality. This is almost the opposite of the truth, since one of the main reasons for doing randomised trials is because patients are 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).[2,58]

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.[2] This is clearly the right way to interpret the astrological 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.[59,60] By contrast, a meta-analysis of the major randomised trials of fibrinolytic therapy has provided especially strong evidence of overall benefit,[25] 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.[25,61] 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.[62] 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 strokes would be expected to be more informative about the net effect on the risk of stroke in different situations[63] 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).[25]

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 studies[1]), 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 difference[2,68]—as with the initial results for aspirin in pre-eclampsia,[35] vesnarinone in heart failure,[38] magnesium in heart attacks,[44] and heparin in stroke.[46] 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 events with antibiotics,[70] of dementia with anti-hypertensive therapy,[71] or of either pre-eclampsia[72] or vascular complications in endstage renal disease[73] 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 estimates[74] (as, for example, with the claim that lowering blood pressure could prevent 19 cases of dementia per 1000 patients treated for 5 years,[71] 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

1 MacMahon S, Collins R. Reliable assessment of the effects of treatment on mortality and major morbidity, II: observational studies. *Lancet* 2001; **357:** 455–62.
2 Collins R, Peto R, Gray R, Parish S. Large-scale randomized evidence: trials and overviews. In: Weatherall DJ, Ledingham JGG, Warrell DA, eds. Oxford Textbook of Medicine, Volume 1. Oxford: Oxford University Press, 1996: 21–32.
3 Teo KK, Yusuf S, Furberg CD. Effects of prophylactic antiarrhythmic drug therapy in acute myocardial infarction: an overview of results from randomized controlled trials. *JAMA* 1993; **270:** 1589–95.
4 The Cardiac Arrhythmia Suppression Trial II investigators. Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. *N Engl J Med* 1992; **327:** 227–33.
5 HIV Trialists' Collaborative Group. Zidovudine, didanosine, and zalcitabine in the treatment of HIV infection: meta-analyses of the randomised evidence. *Lancet* 1999; **353:** 2014–25.
6 Early Breast Cancer Trialists' Collaborative Group. Favourable and unfavourable effects on long-term survival of radiotherapy for early breast cancer: an overview of the randomised trials. *Lancet* 2000; **355:** 1757–70.
7 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.
8 MacMahon S, Peto R, Cutler J, et al. Blood pressure, stroke, and coronary heart disease. Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet* 1990; **335:** 765–74.
9 Law MR, Wald NJ, Thompson SG. By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *BMJ* 1994; **308:** 367–73.
10 Armitage P. The role of randomization in clinical trials. *Stat Med* 1982; **1:** 345–52.
11 Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998; **317:** 1185–90.
12 Keirse MJNC. Amniotomy or oxytocin for induction of labour: re-analysis of a randomized controlled trial. *Acta Obstet Gynecol Scand* 1988; **67:** 731–35.
13 Hansson L, Lindholm LH, Niskanen L, et al, for the Captopril Prevention Project (CAPPP) study group. Effect of angiotensin-converting-enzyme inhibition compared with conventional therapy on cardiovascular morbidity and mortality in hypertension: the Captopril Prevention Project (CAPPP) randomised trial. *Lancet* 1999; **353:** 611–16.
14 Peto R. Failure of randomisation by "sealed" envelope. *Lancet* 1999; **354:** 73.
15 Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000; **342:** 1878–86.
16 Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000; **342:** 1887–92.
17 Francis T Jr, Korns RF, Voight RB, et al. An evaluation of the 1954 poliomyelitis vaccine trials: summary report. *Am J Public Health* 1955; **45:** 1–50.
18 Pocock SJ, Elbourne DR. Randomized trials or observational tribulations? *N Engl J Med* 2000; **342:** 1907–09.
19 Chalmers TC, Matta RJ, Smith H Jr, Kunzler A-M. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *N Engl J Med* 1977; **297:** 1091–96.
20 The Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project. *N Engl J Med* 1980: **303:** 1038–41.
21 Cuzick J, Edwards R, Segnan N. Adjusting for non-compliance and contamination in randomized clinical trials. *Stat Med* 1997; **16:** 1017–29.
22 Antiplatelet Trialists' Collaboration. Collaborative overview of randomised trials of antiplatelet therapy. I: Prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. *BMJ* 1994; **308:** 81–106.
23 ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17 187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet* 1988; **2:** 349–60.
24 Gruppo Italiano per lo Studio della Streptochinasi nell'infarto miocardico (GISSI). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. *Lancet* 1986; **1:** 397–402.
25 Fibrinolytic Therapy Trialists' (FTT) Collaborative Group. Indications for fibrinolytic therapy in suspected acute myocardial infarction: collaborative overview of early mortality and major morbidity results from all randomised trials of more than 1000 patients. *Lancet* 1994; **343:** 311–22.
26 Food and Drug Administration. Final rule for professional labeling of aspirin, buffered aspirin, and aspirin in combination with antacid drugs (FR Doc 98–28519). *Federal Register* 1998; **63:** 56802–19.
27 Clarke M, Chalmers I. Discussion sections in reports of controlled trials published in general medical journals: islands in search of continents? *JAMA* 1998; **280:** 280–82.
28 Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994; **309:** 1351–55.
29 Aspirin Myocardial Infarction Study Research Group. A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA* 1980; **243:** 661–69.
30 Cleland JGF, Bulpitt CJ, Falk RH, et al. Is aspirin safe for patients with heart failure? *Br Heart J* 1995; **74:** 215–19.
31 Flather MD, Yusuf S, Køber L, et al, for the ACE-inhibitor Myocardial Infarction Collaborative Group. Long-term ACE-inhibitor therapy in patients with heart failure or left-ventricular dysfunction: a systematic overview of data from individual patients. *Lancet* 2000; **355:** 1575–81.
32 Feinstein AR, Sosin DM, Wells CK. The Will Rogers phenomenon: stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. *N Engl J Med* 1985; **312:** 1604–08.
33 West of Scotland Coronary Prevention Study Group. Influence of pravastatin and plasma lipids on clinical events in the West of Scotland Coronary Prevention Study (WOSCOPS). *Circulation* 1998; **97:** 1440–45.
34 Yusuf S, Collins R, Peto R, et al. Intravenous and intracoronary fibrinolytic therapy in acute myocardial infarction: overview of results on mortality, reinfarction and side-effects from 33 randomized controlled trials. *Eur Heart J* 1985; **6:** 556–85.
35 Imperiale TF, Petrulis AS. A meta-analysis of low-dose aspirin for the prevention of pregnancy-induced hypertensive disease. *JAMA* 1991; **266:** 260–64.
36 CLASP (Collaborative Low-dose Aspirin Study in Pregnancy) Collaborative Group. CLASP: a randomised trial of low-dose aspirin for the prevention and treatment of pre-eclampsia among 9364 pregnant women. *Lancet* 1994; **343:** 619–29.
37 Knight M, Duley L, Henderson-Smart DJ, King JF. Antiplatelet agents for preventing and treating pre-eclampsia. In: Cochrane Library, issue 4. Oxford: Update Software, 2000.
38 Feldman AM, Bristow MR, Parmley WW, et al for the Vesnarinone Study Group. Effects of vesnarinone on morbidity and mortality in patients with heart failure. *N Engl J Med* 1993; **329:** 149–55.
39 Cohn JN, Goldstein SO, Greenberg BH, et al, for the Vesnarinone Trial Investigators. A dose-dependent increase in mortality with vesnarinone among patients with severe heart failure. *N Engl J Med* 1998; **339:** 1810–16.
40 Bucher HC, Guyatt GH, Cook RJ, et al. Effect of calcium supplementation on pregnancy-induced hypertension and preeclampsia: a meta-analysis of randomized controlled trials. *JAMA* 1996; **275:** 1113–17.
41 Levine RJ, Hauth JC, Curet LB, et al. Trial of calcium to prevent preeclampsia. *N Engl J Med* 1997; **337:** 69–76.
42 Yusuf S, Collins R, MacMahon S, Peto R. Effects of intravenous nitrates on mortality in acute myocardial infarction: an overview of the randomised trials. *Lancet* 1988; **1:** 1088–92.
43 Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto Miocardico. GISSI-3: effects of lisinopril and transdermal glyceryl trinitrate singly and together on 6-week mortality and ventricular function after acute myocardial infarction. *Lancet* 1994; **343:** 1115–21.
44 Teo KK, Yusuf S, Collins R, Held PH, Peto R. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. *BMJ* 1991; **303:** 1499–503.
45 ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group. ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58 050 patients with suspected acute myocardial infarction. *Lancet* 1995; **345:** 669–85.
46 Kay R, Wong KS, Yu YL, et al. Low-molecular-weight heparin for the treatment of acute ischemic stroke. *N Engl J Med* 1995; **333:** 1588–93.
47 Gubitz G, Counsell C, Sandercock P, Signorini D. Anticoagulants for acute ischaemic stroke. In: Cochrane Library, issue 4. Oxford: Update Software, 1999.
48 Gelmers HJ. The effects of nimodipine on the clinical course of patients with acute ischemic stroke. *Acta Neurol Scand* 1984; **69:** 232–39.
49 Horn J, Limburg M. Calcium antagonists for acute ischemic stroke. In: Cochrane Library, issue 4. Oxford: Update Software, 2000.
50 Stephens NG, Parsons A, Schofield PM, et al. Randomised controlled trial of vitamin E in patients with coronary disease: Cambridge Heart Antioxidant Study (CHAOS). *Lancet* 1996; **347:** 781–86.

51 GISSI-Prevenzione Investigators (Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto miocardico). Dietary supplementation with n-3 polyunsaturated fatty acids and vitamin E after myocardial infarction: results of the GISSI-Prevenzione trial. *Lancet* 1999; **354:** 447–55.

52 The Heart Outcomes Prevention Evaluation Study Investigators. Vitamin E supplementation and cardiovascular events in high-risk patients. *N Engl J Med* 2000; **342:** 154–60.

53 The EC/IC Bypass Study Group. Failure of extracranial-intracranial arterial bypass to reduce the risk of ischemic stroke: results of an international randomized trial. *N Engl J Med* 1985; **313:** 1191–20.

54 Goldring S, Zervas N, Langfitt T. The extracranial-intracranial bypass study: a report of the committee appointed by the American Association of Neurological Surgeons to examine the study. *N Engl J Med* 1987; **316:** 817–20.

55 Barnett HJM, Sackett D, Taylor DW, et al. Are the results of the extracranial-intracranial bypass trial generalizable? *N Engl J Med* 1987; **316:** 820–24.

56 Dickersin K. How important is publication bias? A synthesis of available data. *AIDS Edu Prev* 1997; **9:** 15–21.

57 Pocock SJ, Hughes MD. Estimation issues in clinical trials and overviews. *Stat Med* 1990; **9:** 657–71.

58 Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991; **266:** 93–98.

59 Thiemann DR, Coresh J, Schulman SP, Gerstenblith G, Oetgen WJ, Powe NR. Lack of benefit for intravenous thrombolysis in patients with myocardial infarction who are older than 75 years. *Circulation* 2000; **101:** 2239–46.

60 Berger AK, Radford MJ, Wang Y, Krumholz HM. Thrombolytic therapy in older patients. *J Am Coll Cardiol* 2000; **36:** 366–74.

61 White HD. Thrombolytic therapy in the elderly. *Lancet* 2000; **356:** 2028–30.

62 Collins R, Peto R, MacMahon S, et al. Blood pressure, stroke, and coronary heart disease. Part 2, short-term reductions in blood pressure: overview of randomised drug trials in their epidemiological context. *Lancet* 1990; **335:** 827–38.

63 Rothwell PM, Warlow CP, on behalf of the European Carotid Surgery Trialists' Collaborative Group. Prediction of benefit from carotid endarterectomy in individual patients: a risk-modelling study. *Lancet* 1999; **353:** 2105–10.

64 Rothwell PM. Can overall results of clinical trials be applied to all patients? *Lancet* 1995; **345:** 1616–19.

65 Early Breast Cancer Trialists' Collaborative Group. Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy: 133 randomised trials involving 31 000 recurrences and 24 000 deaths among 75 000 women. *Lancet* 1992; **339:** 1–15, 71–85.

66 Davies C, McGale P, Peto R. Variation in use of adjuvant tamoxifen. *Lancet* 1998; **351:** 1487–88.

67 Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet* 1998; **351:** 1451–67.

68 Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. An introduction to bayesian methods in health technology assessment. *BMJ* 1999; **319:** 508–12.

69 Layton AM, Ibbotson SH, Davies JA, Goodfield MJD. Randomised trial of oral aspirin for chronic venous leg ulcers. *Lancet* 1994; **344:** 164–65.

70 Gurfinkel E, Bozovich G, Daroca A, Beck E, Mautner B, for the ROXIS Study Group. Randomised trial of roxithromycin in non-Q-wave coronary syndromes: ROXIS pilot study. *Lancet* 1997; **350:** 404–07.

71 Forette F, Seux M-L, Staessen JA, et al, on behalf of the Syst-Eur Investigators. Prevention of dementia in randomised double-blind placebo-controlled Systolic Hypertension in Europe (Syst-Eur) trial. *Lancet* 1998; **352:** 1347–51.

72 Chappell LC, Seed PT, Briley AL, et al. Effects of antioxidants on the occurrence of pre-eclampsia in women at increased risk: a randomised trial. *Lancet* 1999; **354:** 810–16.

73 Boaz M, Smetana S, Weinstein T, et al. Secondary prevention with antioxidants of cardiovascular disease in endstage renal disease (SPACE): randomised placebo-controlled trial. *Lancet* 2000; **356:** 1213–18.

74 Altman DG. Confidence intervals for the number needed to treat. *BMJ* 1998; **317:** 1309–12.

A list of further reading can be found at www.thelancet.com