2.4 Large- scale randomized evidence Trials and me

2.4 Large- scale randomized evidence: Trials and meta-analyses of trials 51

ESSENTIALS Reliable detection or refutation of realistically moderate effects on major outcomes often requires large-scale randomized evidence As long as doctors start with a healthy scepticism about the many ap- parently striking claims and counterclaims that appear in the medical literature, trial results do make sense. The main enemy of common sense is overoptimism: there are a few striking exceptions where treatments for serious disease really do turn out to work extremely well, but in general most claims of vast improvements from new ther- apies turn out to be evanescent. Clinical trials generally need to be able to detect or to refute real- istically moderate (but still worthwhile) differences between treat- ments in long-term disease outcome. Large-scale randomized evidence should be able to detect such effects, but medium-sized trials or medium-sized meta-analyses can, and often do, yield false- negative or exaggeratedly positive results. If the results from such studies seem too good to be true then they probably are: conversely, unpromising evidence can be misleading if it is from a study of inad- equate size, or from one particular subgroup of a large study with a clearly favourable overall result. Realistically moderate expectations of what a treatment might achieve (or, if one treatment is to be compared with another, of how large any difference between the main effects of these two treatments is likely to be) should foster studies that can discriminate reliably be- tween (1) a difference in outcome that is realistically moderate but still worthwhile, and (2) a difference in outcome that is too small to be of any material importance. To assess moderate effects reliably, avoid both moderate biases and moderate random errors To demonstrate or refute realistically moderate differences in out- come, studies must guarantee (1) strict control of bias—which, in general, requires proper randomization and appropriate statistical analysis, with no unduly 'data-dependent' emphasis on specific parts of the overall evidence; and (2) strict control of the play of chance— which, in general, requires large numbers of patients with the out- come of interest, rather than a lot of detail on each patient. The conclusion is obvious: moderate biases and

moderate random errors must both be avoided if moderate benefits are to be assessed reliably. This leads to the need for large numbers of properly randomized pa- tients with properly analysed data, which in turn should lead to some large but simple randomized trials (or ‘mega-trials’) and to large sys- tematic overviews (or ‘meta-analyses’) of all related randomized trials. Other forms of evidence may be untrustworthy Nonrandomized evidence, unduly small randomized trials, unduly small meta-analyses of trials, and undue emphasis on particular sub- groups (or on particular trials) are all inferior sources of evidence about current patient management or as foundations for future research strategies because they often cannot discriminate reliably between moderate (but worthwhile) differences and negligible dif- ferences in outcome, and the mistaken clinical conclusions that they engender could well result in the undertreatment, overtreatment, or other mismanagement of millions of future patients worldwide. Benefits of large-scale randomized evidence Hundreds of thousands of premature deaths each year could be avoided by seeking appropriately large-scale randomized evidence about various widely practicable treatments for the common causes of death, and by disseminating this evidence appropriately. The value of large-scale randomized evidence is illustrated by the trials of fibrinolytic therapy for acute myocardial infarction; of antiplatelet therapy for a variety of vascular conditions; of endocrine therapy for early breast cancer; and of drug therapy for lowering blood pressure. In these examples, proof of benefit that could not have been achieved by either small-scale randomized evidence or non- randomized evidence has led to widespread changes in practice that are now preventing hundreds of thousands of premature deaths each year, and appropriately large-scale randomized evidence could substantially improve the management of many important, but non- fatal, medical conditions. Moderate (but worthwhile) effects on major outcomes are generally more plausible than large effects Some treatments have large, and hence obvious, effects on survival: for example, it was clear without the need for any randomized trials that prompt treatment of diabetic coma or cardiac arrest can save 2.4 Large-scale randomized evidence:
Trials and meta-analyses of trials Colin Baigent, Richard Peto, Richard Gray, Natalie Staplin, Sarah Parish, and Rory Collins

 lives, and more recently the introduction of protease inhibitors for the treatment of HIV infection led to a reduction in AIDS-related morbidity and mortality that was large enough to be obvious even without randomized evidence; indeed, the remarkable effectiveness of antiretroviral drugs can be seen from the sudden reversal, after the mid-1990s, of the upward trend in mortality among US men aged 30–34 (see Fig. 2.4.1), the chief cause of which was HIV/AIDS. However, perhaps in part because of these striking successes, for the past few decades the hopes of large treatment effects on mortality and major morbidity in many serious diseases have been unrealis- tically high. Of course, treatments do quite commonly have large effects on various less fundamental measures: certain drugs clearly reduce blood pressure, blood cholesterol, or blood glucose; many tu- mours or leukaemias can be controlled temporarily by radiotherapy or chemotherapy; and, in acute myocardial infarction, lidocaine (lignocaine) can prevent many arrhythmias and fibrinolytic therapy can dissolve many thrombi. However, although such effects on inter- mediate outcomes may be large, the net effects on mortality may be much more modest. In general, if substantial uncertainty remains about the efficacy of a practicable treatment, its effects on major endpoints are probably either negligibly small, or only moderate, rather than large. Indirect support for this possibly pessimistic conclusion comes from many sources, including: the previous few decades of disappointingly slow progress in the curative treatment of common chronic diseases of middle age; the heterogeneity of each single disease, as

evidenced by the unpredictability of survival duration even when apparently similar patients are compared with each other; the variety of dif- ferent mechanisms in certain diseases that can lead to death, only one of which may be appreciably influenced by any one particular therapy; the modest effects often suggested by meta-analyses (see later) of various therapies, and, in certain special cases, observa- tional epidemiological studies of the strength of the relationship between a particular disease and the factor that the treatment will modify (e.g. blood pressure, blood cholesterol, or blood glucose; see later on in this chapter). Having accepted that only moderate reductions in mortality are likely with many currently unevaluated interventions, how worth- while might such effects be if they could be detected reliably? To some clinicians, reducing the risk of early death in patients with myocar- dial infarction from 10 per 100 patients down to 9 or 8 per 100 pa- tients treated may not seem particularly worthwhile, and if such a reduction was only transient, or involved an extremely expensive or toxic treatment, this might well be an appropriate view. Worldwide, however, several million patients a year suffer an acute myocardial infarction, and if just one million were to be given a simple, nontoxic, and widely practicable treatment that reduced the risk of early death from 10% down to 9 or 8% (that is, a proportional reduction of 10 or 20%), this would avoid 10 or 20 thousand deaths. (For example, at least a million patients a year now receive fibrinolytic therapy for acute myocardial infarction, which is avoiding about 20 000 early deaths a year.) Such absolute gains are substantial, and might consid- erably exceed the number of lives that could be saved by a much more effective treatment of a much less common disease. Reliable detection or refutation of moderate differences requires avoidance of both moderate biases and moderate random errors If realistically moderate differences in outcome are to be reliably detected or reliably refuted, then errors in comparative assessments of the effects of treatment need to be much smaller than the differ- ence between a moderate, but worthwhile, effect and an effect that is too small to be of any material importance. This in turn implies that moderate biases and moderate random errors cannot be tol- erated. The only way to guarantee very small random errors is to study really large numbers, and this can be achieved in two main ways: by making individual studies large, and by combining infor- mation from as many relevant studies as possible in a systematic meta-analysis (Table 2.4.1). However, it is not much use having very small random errors if there are moderate biases, so even the large sizes of some nonrandomized analyses of computerized hos- pital records, where the complex factors involved in the decision to treat a person with a particular drug may not be recorded in suffi- cient detail, cannot guarantee medically reliable comparisons be- tween the effects of different treatments (see later). For example, the choice of treatment may be strongly affected by subtle patient characteristics that are correlated with the prognosis. (A crude 1950 1960 1970 1980 1990 2000 2010 0% 0.1% 0.2% 0.3% 0.4% 0.5% 0.6% Male Female 0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 5-year risk Death rate/1000 UNITED STATES 1950−2015: Males & Females All medical mortality at ages 30−34 Source: WHO mortality & UN population estimates Fig. 2.4.1 Mortality trends in the United States of America among men and women aged 30–34 during the period 1950–2015. Antibacterial drugs caused a big decrease in mortality around the middle of the century in both sexes. The increase in AIDS-related mortality since the early 1980s caused a sharp increase in all-cause mortality, particularly in men, which continued until it was spectacularly reversed by effective antiretroviral drug combinations in the mid-1990s.

 illustration of such problems is provided by the old joke ‘What’s the most dangerous place in the world?’ ‘Bed—look at the number of people who die in bed!’.) Avoiding moderate biases Proper randomization avoids

systematic differences between the types of patient in different treatment groups. The fundamental reason for randomization is to avoid moderate bias, by ensuring that each type of patient can be expected (but for the play of chance) to have been allocated in similar proportions to the dif- ferent treatment strategies that are to be compared. This means that only random differences should affect the final comparisons of outcome. Nonrandomized methods, in contrast, cannot generally guarantee that the types of patient given the new study treatment do not differ systematically in any important ways from the types of patient given any other treatment(s) with which the new study treatment is to be compared. For example, moderate biases might arise if the study treatment was novel and doctors were afraid to use it for the most seriously ill patients, or, conversely, if they were more ready to use it for those who were desperately ill. There may also be other ways in which the severity of the condition differentially affects the likelihood of being assigned to different treatments by the doctor's choice (or by the patient's choice, or by any other non- random procedure). It might appear at first sight that by collecting enough infor- mation about various prognostic features it would be possible to make some mathematical adjustments to correct for any such dif- ferences between the types of patients who, in a nonrandomized study, receive the different treatments that are to be compared. The ill-conceived hope is that such methods, which are often car- ried out on routinely collected healthcare data, might achieve comparability between those entering the different treatment groups, but they cannot be guaranteed to do so, and often fail ser- iously. The difficulty is that some important prognostic factors may be unrecorded, while others may not have been discovered yet or be difficult to assess exactly and hence difficult (or even impossible) to adjust for reliably. Although there are examples of nonrandomized studies in which the estimated effects of treat- ment appear quantitatively close to those observed in analogous randomized trials, there are many examples where they are not, being either quantitatively incorrect—so that drugs appear ei- ther misleadingly promising or of misleadingly low efficacy—or even qualitatively incorrect, when a harmful drug might appear effective (or vice-versa). The machinery of a properly randomized trial No foreknowledge of what the next treatment will be In a properly randomized trial, the decision to enter a patient is made in ignorance of which of the trial treatments that patient will, once entered, be allocated. The treatment allocation is then made known after trial entry has been decided upon. (The purpose of this sequence is to ensure that foreknowledge of what the next treatment is going to be cannot affect the decision as to whether to enter the patient; if it did, those to be allocated one treatment might differ systematically from those to be allocated another.) Ideally, any major prognostic features should also be irreversibly recorded before the treatment is revealed, particularly if these are to be used in any treatment analyses. For, if the recorded value of some prognostic factor might be affected by knowledge of the trial treatment allocation, then treatment comparisons within subgroups defined by that factor might be moderately biased. In particular, treatment comparisons just among 'responders' or just among 'nonresponders' can be extremely misleading un- less the response is assessed before treatment allocation (which it can sometimes be, if all patients have a 'run-in' period on ac- tive treatment before randomization, partly to assess the response to treatment and partly to exclude those who seem during this prerandomization run-in unlikely to participate wholeheartedly in the main post-randomization study). No bias in patient management or in outcome assessment An additional difficulty, in both randomized and nonrandomized comparisons of various treatments, is that there might be system- atic differences in the use of other treatments (including general supportive care) or in the assessment of major outcomes. A non- randomized comparison may well suffer from moderate biases due to such systematic differences in ancillary care or assessment, par- ticularly if it merely involves the

retrospective review of medical records. In the context of a randomized comparison, however, it is generally possible to devise ways to keep any such biases small. For example, placebo tablets may be given to control-allocated patients and certain subjective assessments may be ‘blinded’ (although this is less important in studies assessing mortality). ‘Intention-to-treat’ analyses with no post-randomization exclusions Even in a properly randomized trial, unnecessary biases may be introduced by inappropriate statistical analysis. One of the most important sources of bias in the analysis is undue concentration on just part of the evidence; that is to say, on ‘data-derived subgroup Table 2.4.1 Requirements for reliable assessment of moderate effects: negligible biases and small random errors NEGLIGIBLE BIASES (i.e. guaranteed avoidance of moderate biases) Proper randomization (nonrandomized methods might suffer moderate biases) Analysis by allocated treatment (including all randomized patients: ‘intention-to-treat’ analysis) Chief emphasis on overall results (no unduly data-dependent emphasis on particular subgroups) Systematic overview of all relevant randomized trials (no unduly data-dependent emphasis on particular studies) SMALL RANDOM ERRORS (i.e. guaranteed avoidance of moderate chance fluctuations) Large numbers in any new trials (to be really large, trials should be ‘streamlined’) Systematic overview of all relevant randomized trials (which yields the largest possible total numbers)

54 SECTION 2 Background to medicine analyses’ (see next). Another easily avoided bias is caused by the post-randomization exclusion of patients, particularly if the type (and hence prognosis) of those excluded from one treatment group differs from those excluded from another. Therefore, one of the fundamental statistical analyses of a trial that should be made available is an analysis that compares all those originally allocated one treatment (even though some of them may not have actually received it) with all those allocated the other treat- ment. This is sometimes referred to as an ‘intention-to-treat’ ana- lysis. Additional analyses can also be reported; for example, in describing the frequency of some very specific side effect it may well be preferable to record its incidence only among those who actually received the treatment. (This is because strictly random-ized comparisons may not be needed to assess extreme relative risks.) However, in assessing moderate effects on the main out- come of interest such ‘on-treatment’ analyses can be misleading, and ‘intention-to-treat’ analyses are generally a more trustworthy guide as to whether there is any real difference between the trial treatments in their effects on long-term outcome. No unduly data-dependent emphasis on results in particular subgroups Treatment that is appropriate for one patient may be inappropriate for another. Ideally, therefore, what is wanted is not only an answer to the question ‘Is this treatment helpful on average for a wide range of patients?’, but also an answer to the question ‘For which recog- nizable categories of patient is this treatment particularly helpful?’ However, this ideal is difficult to attain directly because the direct use of clinical trial results to assess proportional risk reductions in particular subgroups of patients is surprisingly unreliable. Of course, patients who already have a very good prognosis anyway and are at low absolute risk cannot have a large absolute benefit (for even if a small risk is halved the absolute benefit is small). Classification of patients as being at low (or high) risk of an adverse disease out- come is often a reliable guide as to which patients can expect little absolute gain even if the trial treatment works as expected (and as to which patients might expect a worthwhile gain). This low-risk/ high-risk split may not require support from formal subgroup analyses—indeed, it could even be damaged by such analyses. For, even if the proportional effects of treatment in specific subgroups are importantly different, standard subgroup analyses are so insensitive that they may well fail to demonstrate these differences. Moreover, even if there are highly significant differences between the propor- tional risk reductions produced by the trial treatment in different

subgroups, and the results seem to suggest that the treatment works in some subgroups but not in others (thereby giving the appearance of a ‘qualitative interaction’), this may still not be good evidence for subgroup-specific treatment preferences. The play of chance often produces qualitatively wrong answers in particular subgroups in trials (or in meta-analyses of trials) that could, if interpreted incau- tiously, lead to millions of people being treated inappropriately, or untreated inappropriately. Questions about such ‘interactions’ between patient character- istics and the effects of treatment are easy to ask, but are surpris- ingly difficult to answer reliably. Apparent interactions can often be produced by the play of chance and, in particular subgroups, can mimic or obscure some of the moderate treatment effects that might realistically be expected. To demonstrate this, a subgroup analysis was performed based on the astrological birth signs of patients ran- domized in the very large Second International Study of Infarct Survival (ISIS-2) trial of the treatment of acute myocardial infarc- tion. Overall in this trial, the 1-month survival advantage produced by aspirin was conclusively demonstrated (804 vascular deaths among 8587 patients allocated aspirin versus 1016 among 8600 al- located no aspirin; 23% proportional reduction, two-sided p value <0.000001). However, when these analyses were subdivided into 12 subgroups by the patients’ astrological birth signs (in mediaeval as- trology, the ‘birth sign’ is determined by the month of birth; for ex- ample, ‘Libra’ means born 24 September to 23 October, and ‘Gemini’ means born 22 May to 21 June) to illustrate the unreliability of sub- group analyses, aspirin appeared totally ineffective for those born under Libra or Gemini (Table 2.4.2). It would obviously be unwise to conclude from such a result that patients born under the astro- logical birth sign of Libra or Gemini should not be given aspirin if they have a heart attack. However, similar conclusions based on ‘exploratory’ data-derived subgroup analyses, which from a purely statistical viewpoint are no more reliable than these, are often re- ported and believed, with inappropriate effects on worldwide clin- ical practice. There are three main remedies for this unavoidable conflict be- tween the reliable subgroup-specific conclusions that doctors and patients want and need, and the statistically unreliable findings that direct subgroup analyses can usually offer. However, the extent to which these remedies are helpful in particular instances is one on which informed judgements differ. First, where there are good a priori reasons for anticipating that the proportional effects of treatment might be very different in different circumstances then a limited number of subgroup analyses may be prespecified in the study protocol, along with a prediction of the direction of such proposed interactions. (For example, it was expected that the benefits of fibrinolytic therapy for acute myocardial infarction would be greater the earlier such patients were treated and so some studies prespecified that the analyses would be subdivided by the number of hours from the onset of symptoms to treatment: see later.) These prespecified Table 2.4.2 False-negative mortality effect in a subgroup defined only by the mediaeval astrological birth sign: the ISIS-2 (1988) trial of aspirin among over 17 000 patients with acute myocardial infarction

Astrological birth sign	No. of 1-month deaths (aspirin versus placebo)	Statistical significance
Libra or Gemini	150 vs. 147	NS
All other signs	654 vs. 869	2p <0.000001
Any birth signa	804 vs. 1016 (9.4%) (11.8%)	2p <0.000001

a Appropriate overall analysis for assessing the true effect in all subgroups.

Mediaeval astrology divides birth dates into 12 ‘birth signs’ (which depend only on the day and month of birth, not the year of birth). To demonstrate the potential unreliability of other subgroup analyses, the ISIS-2 patients were divided into 12 subgroups according to their astrological birth sign, and the apparent effects of aspirin were calculated separately in each of these 12 subgroups. Because of the play of chance, the apparent effects differed from one subgroup to another (ranging from no apparent effect of aspirin in two subgroups (Libra and Gemini: see text for definition) to

aspirin apparently halving mortality in another (Capricorn)).

2.4 Large-scale randomized evidence: Trials and meta-analyses of trials 55 subgroup-specific analyses can then be taken somewhat more ser- iously than other subgroup analyses, but they can still yield im- portantly wrong answers. The second approach is to emphasize chiefly the overall results of a trial (or, better still, of all such trials) for particular outcomes, as a guide to—or at least a context for speculation about—the qualita- tive results in various specific subgroups of patients, and to give less weight to the actual results in each separate subgroup. This is clearly the right way to interpret the astrological 'findings' in Table 2.4.2, but, if used sensibly, it is also likely in many other circumstances to provide the best assessment of whether one treatment is better than another in particular subgroups. The proportional effect of treat- ment as estimated by the overall results may well provide a useful approximation to the proportional effects of treatment in particular subgroups. Of course, any such extrapolation of overall results to particular subgroups needs to be performed in a sensible way, keeping in touch with medical realities. In particular, if a treatment has substantial side effects, it may be obviously inappropriate for low-risk patients. (In this case, the side effects in a particular subgroup and the pro- portional benefit in that subgroup should be estimated separately, but the estimation for both might be more reliable if based on an appropriate extrapolation from the overall results rather than on the results in that one subgroup alone.) The third approach is to be influenced, in discussing the likely ef- fects on mortality in specific subgroups, not only by the mortality analyses in these subgroups but also by the analyses of recurrence- free survival or some other major 'surrogate' outcome. For, if the overall results are similar but much more highly significant for recurrence-free survival than for mortality, subgroup analyses with respect to the former may be more stable and may provide a better guide as to whether there are any major differences between sub- groups in the effects of treatment. The appropriate interpretation of apparently different results in different subgroups of the randomized evidence is still one of the most difficult matters of judgement in the interpretation of random- ized evidence; at present, many clinicians and regulatory agencies pay too much attention to irregularities in apparent effects that are consistent with chance. Avoiding moderate random errors The need for large-scale randomization To distinguish reliably between the two alternatives that (a) there is no worthwhile difference in survival or that (b) treatment confers a moderate, but worthwhile, benefit (e.g. 10 or 20% fewer deaths), not only must systematic errors be guaranteed to be small (see earlier) compared with such a moderate risk reduction, but so too must any of the purely random errors that are produced just by chance. Random errors can be reliably avoided only by studying very large numbers of patients and hence large enough numbers of 'endpoints'. However, it is not sufficiently widely appreciated just how large clin- ical trials need to be in order to detect moderate differences reli- ably. This can be illustrated by a hypothetical trial that is actually quite inadequate—even though by some standards it is moderately large—in which a 20% reduction in mortality (from 10 to 8%) is sup- posed to be detected among 2000 heart attack patients (1000 treated and 1000 controls). In this case, one might predict about 100 deaths (10%) in the control group and 80 deaths (8%) in the treated group. However, if this difference were to be observed it would not be con- ventionally significant ($p = 0.1$); indicating that even if there is no real difference between the effects of the trial treatments, it would still be relatively easy for a result at least as extreme as this to arise by chance alone. Although the play of chance might well increase the difference enough to make it conventionally significant (e.g. 110 deaths vs. 70 deaths, $2p < 0.001$), it might equally well dilute, oblit- erate (e.g. 90 deaths vs. 90 deaths), or even reverse it. The situation in real life is often even worse, as the average trial size may include only a few

dozen events rather than the several hundred (or few thou- sand) that would ideally be needed to guide the future treatment of millions. Mega-trials: How to randomize large numbers One of the chief techniques for obtaining appropriately large-scale randomized evidence is to make trials extremely simple, and then to invite hundreds of hospitals to collaborate. The first of these large streamlined trials (or mega-trials) were the ISIS and GISSI studies of heart attack treatment in the 1980s, and many other mega-trials have now been successfully undertaken, not only in the field of cardiology—where numerous large trials have now been performed—but also in other specialties where treatment might be expected to have only moderate effects on morbidity and mortality from a common disease or injury. Many such mega-trials have pro- duced medically important results that would not otherwise have been reliably obtained. However, in terms of medically significant findings, what has been achieved so far is only a fraction of what would be possible if this research strategy could be more widely adopted. Any obstacle to simplicity is an obstacle to large size, and so it is worth making enormous efforts at the design stage to sim- plify and streamline the process of entering, treating, and assessing patients. Many trials would be of much greater scientific value if they collected 10 times less information, both at entry and during follow-up, on 10 times more patients. Since those responsible for entering patients into trials are generally busy people, it is particu- larly important to simplify the entry of patients, otherwise rapid re- cruitment may prove difficult (see later). Likewise, when allocating resources within large-scale trials, it is important to direct them to where it chiefly matters, namely the recruitment of large num- bers of patients and counting how many suffer the main outcomes of interest, whereas large sums of money may often be wasted on inappropriate audits or unnecessary or excessively frequent meas- urements, the analysis of which will not contribute to answering the main study question. Simplification of entry procedures for trials:
The ‘uncertainty principle’ For ethical reasons, patients cannot have a commonly available treatment chosen at random if either they or their doctor are (for any reasons) already reasonably certain that another treatment is preferable. Hence, randomization can be offered only if both doctor and patient feel substantially uncertain as to which of the trial op- tions is best. The question then arises: ‘Which categories of patients

 about whose treatment there is such uncertainty should be offered randomization?’ The obvious answer is all of them, welcoming the heterogeneity that this will produce. (For example, either the treat- ment of choice will turn out to be the same for men and women, in which case the trial might as well include both, or it will be different, in which case it is particularly important to study both.) In appropri- ately large trials, patient homogeneity is generally a defect while het- erogeneity is generally a strength. Consider, for example, the trials of immediate fibrinolytic therapy for acute myocardial infarction. Some had restrictive entry criteria that allowed inclusion of only those patients who presented between 0 and 6 h after the onset of pain, and those trials contributed almost nothing to the key question of how late such treatment can still be useful. In contrast, the trials with wider and more heterogeneous entry criteria that included some patients with somewhat longer delays between pain onset and hospitalization answered this question prospectively, and were able to show that fibrinolytic therapy can have definite protective effects not only when given 0 to 6 but also when given 7 to 12 hours after the onset of pain (see later). This approach of randomizing the full range of patients in whom there is substantial uncertainty as to which treatment option is best was used in the first Medical Research Council Asymptomatic Carotid Surgery Trial (ACST-1). Narrowing of the carotid artery (which is rapidly detectable by ultrasound) can eventually cause a stroke, or even a succession of strokes. It

can be dealt with surgi- cally by carotid endarterectomy, but in the 1990s there was much uncertainty as to whether such surgery, with its inherent periopera- tive risks, was appropriate for individuals with severe carotid artery narrowing who were currently asymptomatic (that is, had not had a stroke in the past few months). The ACST was therefore designed to compare a policy of immediate carotid endarterectomy with a policy of 'watchful waiting' in asymptomatic patients with substantial ca- rotid artery narrowing. If a patient was prepared at least to consider surgery seriously, then the neurologist and surgeon responsible for that individual's care considered in their own undefined way what- ever medical, personal, or other factors seemed to them to be rele- vant, including, of course, the patient's own preferences and values. Eligibility for randomization was defined by the 'uncertainty prin- ciple' (Fig. 2.4.2): 1. If they or the patient were reasonably certain, for any reason, that they did wish to recommend immediate surgery for that particular patient, the patient was not eligible for entry into the ACST. 2. Conversely, if they or the patient were reasonably certain, for any reason, that they did not wish to recommend immediate surgery, the patient was likewise not eligible for entry into the trial. 3. If, but only if, the doctor(s) and patient were substantially uncertain what to recommend, the patient was automatically eligible for randomization between immediate versus no im- mediate surgery (with all patients receiving whatever their doctors judged to be the best available medical care, which generally included advice to stop smoking, low-dose aspirin, and treatment of hypertension; nowadays, it would probably also include a statin). In ACST-1, there were substantial differences between individual doctors in the types of patients about whom they were uncertain (in terms of the severity of carotid stenosis—which was generally recorded on ultrasound as 70%, 80%, or 90% blockage—age, gen- eral health, and various other characteristics). This guaranteed that no category of patient about which there was widespread uncer- tainty would be wholly excluded, and hence guaranteed that the trial would yield at least some direct evidence in a variety of typical pa- tients. As a result of the wide and simple entry criteria adopted by ACST-1, 3120 patients were randomized (which was more than in any previous vascular surgery trial), so the study was able to provide some clear answers about who needed carotid endarterectomy. In asymptomatic patients younger than 75 years of age, with carotid diameter about 70% or more on ultrasound, immediate carotid end- arterectomy halved the net 5-year stroke risk from about 12% to 6% (even though this 6% included the 3% perioperative hazard). For patients with only moderate carotid artery stenosis on ultrasound, the 5-year risks of carotid stroke (excluding perioperative hazard) were 2% versus 9%, while among those with tighter stenosis the risks were 3% versus 10%, suggesting about as much benefit in moderate as in tight stenosis. The 'uncertainty principle' simultaneously meets the require- ments of ethicality, heterogeneity, simplicity, and maximal trial size, and should be widely used. It states that the fundamental eligibility criterion is that both doctor and patient should be substantially uncertain about the appropriateness of each of the trial treatments for that particular patient. With such uncertainty as the fundamental criterion of eligibility, informed consent can often be simplified. For, the degree of 'informed consent' that is appropriate in a randomized comparison of two established treat- ments governed by the 'uncertainty principle' should probably not differ greatly from that which is applied in routine practice outside trials when treatment is being chosen haphazardly—or, to put it another way, 'double standards' between trial and nontrial situations are inappropriate. The haphazard nature of many non- randomized treatment choices is reflected in the wide variations in practice between and within countries. Even when a practice is similar it may be similarly wrong: for example, before the ISIS-2 results became available (see later), few doctors routinely used fi- brinolytic therapy for acute myocardial infarction. Provided that trials are governed by the 'uncertainty principle', there is an ap- proximate parallel between

good science and good ethics. Indeed, in such circumstances, excessively detailed consent procedures (which can be distressing and inhumane, and so would not be con- sidered appropriate in routine nontrial clinical practice) would not be humane or ethically appropriate in trials. Excessively de- tailed consent procedures are, unfortunately, quite common, but their chief purpose is to protect doctors against lawyers rather than to protect patients against anything. This ‘uncertainty principle’ is just one of many ways of simplifying trials and thereby helping them to avoid becoming enmeshed in a mass of wholly unnecessary traditional complexity. If randomized trials can be substantially simplified (which, it must be admitted, requires a reversal of the current trend towards unnecessary complexity), as has already been achieved for a few major diseases, and hence made very much larger, then they will continue to play an appropriately central role in the development of rational criteria for planning treatment strategies and reducing death and disability.

2.4 Large-scale randomized evidence: Trials and meta-analyses of trials 57 Minimizing both bias and random error: Meta-analyses of randomized trials Archie Cochrane was one of the first people to emphasize the need to organize, by specialty, the results from all relevant randomized trials, and the Cochrane Library brings together in a single place a large number of systematic reviews (many of which include meta- analyses of randomized trials) summarizing the available evidence about a wide range of therapeutic questions. When several trials have all addressed much the same question, the traditional pro- cedure of only a few of them becoming widely known may be a source of serious bias, since chance fluctuations for or against treat- ment may affect which trials become well known and widely cited. To avoid this problem, it is appropriate to base inference chiefly on a meta-analysis of all the results from all of the trials that have ad- dressed a particular type of question (or on an unbiased subset of such trials), and not on some potentially biased subset of these trials. Carotid artery stenosis detected by ultrasound, but, as yet, no clinical evidence of stroke from it. Should patient be offered immediate carotid surgery? Doctor(s) or patient reasonably certain (no matter why) that immediate surgery is not appropriate: Ineligible Doctor(s) or patient reasonably certain (no matter why) that immediate surgery is appropriate: Ineligible Doctor(s) and patient substantially uncertain whether to risk immediate surgery: Uncertainty implies eligibility Telephone to randomize Group 1: Allocated No immediate surgery (unless or until a clear indication is thought to have arisen) Group 2: Allocated Immediate carotid surgery (unless definite contraindication is thought to have been discovered, or patient [or doctor] changes their mind) Over the next few years only a small number get carotid surgery 90% get carotid surgery (median delay: 1 month) Statistical comparisons of various outcomes over 5 to 10 years: 100% of Group 1 vs. 100% of Group 2, i.e. ‘intention to treat analysis’ Fig. 2.4.2 Example of the ‘uncertainty principle’ to define eligibility for trial entry: the chief eligibility criterion for the Asymptomatic Carotid Surgery Trial (ACST) was that doctors and patients should be substantially uncertain whether to risk immediate carotid surgery. Partly because this criterion was appropriately flexible, ACST-1 became the largest-ever trial of vascular surgery, showing that the long-term benefits of carotid artery surgery could eventually outweigh the immediate hazards. ACST-2 (http://www.nds.ox.ac.uk/acst) is now randomizing surgery versus carotid stenting where the doctor(s) and patient are substantially uncertain which to prefer.

58 SECTION 2 Background to medicine Such meta-analyses will also minimize random errors in the assess- ment of treatment since, in general, far more patients are involved in a meta-analysis than in any contributory individual trial. The separate trials may well be heterogeneous in their entry cri-

teria, their treatment schedules, their follow-up procedures, their methods of treating relapse, and so on. In view of this heterogeneity, at one extreme each trial might be considered in virtual isolation from all others, while at the opposite extreme the results from all trials could be combined, largely ignoring any heterogeneity. Both these extreme views have some merit, and the pursuit of each by different people may prove more illuminating than too definite an insistence on any one particular approach. However, the heterogen- eity of the different trials merely argues for careful interpretation of any meta-analyses of different trial results, rather than arguing against meta-analyses. Whatever the difficulties in interpreting meta-analyses may be, without them it is difficult to avoid moderate selective biases and substantial random errors, both of which could obscure any moderate treatment effects, or, conversely, imply an ef- fect where none exists. Which meta-analyses are trustworthy? Since the 1970s, a rapidly increasing number of meta-analyses of the results of randomized trials have been reported, not all of which are trustworthy. When considering how reliable a given one might be there are two fundamental questions: What is the potential for bias, and what is the potential size of purely random errors? To answer the first question, consideration must be given to whether biases might exist within individual trials (e.g. because of an unreliable method of randomization or because of post-randomization exclusions from the main analyses), and whether the subset of trials under consid- eration might be a biased subset of all relevant trials that have been performed (as might arise, for example, if certain trials were aban- doned because of unpromising findings, or remained unpublished for this or any other reason). The simplest approach to meta-analysis is merely to have collected and tabulated the published data from whatever randomized trial reports are easily accessible in the literature, and sometimes this may suffice. At the opposite extreme, extensive efforts may have been made by those organizing the meta-analysis to locate every poten- tially relevant randomized trial, including those never published, to collaborate closely with the trialists to seek individual data on each patient ever randomized into those trials, and then (after extensive checks and corrections of such data) to produce, in collaborative re- analyses with those trialists, agreed analyses and publications. The results of some of the largest such collaborative re-analyses will be described later: the Anti Thrombotic Trialists' (ATT) Collaborative Group, the Fibrinolytic Therapy Trialists' (FTT) Collaborative Group, and the Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Collaboration of the original trialists in the meta- analysis process, with collection of detailed data from each indi- vidual trial participant, can help to avoid or minimize the biases that could be produced by missing trials (e.g. owing to the greater likeli- hood of extremely good, or extremely bad, results being particularly widely known and published), by inappropriate post-randomization withdrawals, or by the failure to allocate treatment properly at random. If randomization was performed properly in the first place, the post-randomization withdrawals can often be followed-up and restored to the study for an appropriate 'intention-to-treat' analysis. Knowledge of the exact methods of treatment allocation (backed up by checks on whether the main prognostic factors recorded are non- randomly distributed between the treatment groups in a particular trial) may help to identify trials that were so improperly randomized that they should be excluded from a meta-analysis of the properly randomized trials. Meta-analyses based on individual patient data may also provide more information about treatment effects than the more usual overviews of grouped data, for they allow more detailed analyses—indeed, if they are really large then they may actually yield statistically reliable subgroup analyses of the effects of treatment in particular types of patient. Conversely, even a perfectly conducted meta-analysis of an inter- vention with moderate effects on a major clinical outcome may not be reliable if the trials were all small. There are two reasons for this. First, when the true effect of an intervention is only moderate, most small

trials will fail to reach statistical significance, and may be less likely to be published (or otherwise available) than the few with results that are misleadingly extreme. Hence, a meta-analysis consisting exclusively of small trials is particularly prone to bias. Secondly, the random errors may be too large to allow reliable in- terpretation. A meta-analysis that includes a total of only 100 deaths will have random errors about as great as a single trial with only 100 deaths. For these reasons small-scale evidence, whether from a meta-analysis or from one trial, is often unreliable and may well be found in retrospect to have yielded wrong answers. What is needed is large-scale randomized evidence; it does not matter much whether the totality of the evidence comes from a properly conducted meta- analysis of several trials or one properly conducted trial with such clear results that no further trials were done. The practical medical value of large-scale randomized evidence will be illustrated by a few examples. Examples of information about the efficacy and safety of widely used drugs that could have been reliably established only by large-scale randomized evidence Evidence of benefit from a single very large trial In the ISIS-2 trial, half of 17 000 patients with suspected acute myocardial infarction were allocated aspirin tablets (162 mg/day for 1 month, which virtually completely inhibits cyclo-oxygenase- dependent platelet activation) and half were allocated placebo tab- lets. Before 1988, when the ISIS-2 results were published, aspirin was not routinely used in the treatment of acute myocardial infarction, and no other major trial had (or has subsequently) compared aspirin with an untreated control group in cases of suspected acute myocar- dial infarction. However, the effects of 1 month of aspirin were so definite in ISIS-2 (804/8587 vascular deaths among those who were allocated aspirin versus 1016/8600 among those who were not) that even the lower 99% confidence limit would have represented a worth- while benefit from this simple and inexpensive treatment (Fig. 2.4.3). As a result, worldwide treatment patterns changed sharply when the ISIS-2 results emerged in 1988, and aspirin is now routinely used for the majority of emergency hospital admissions with suspected acute myocardial infarction not only in Europe and America but throughout Asia. In the United Kingdom, for example, two British

 Heart Foundation surveys found cardiologists reporting that rou- tine aspirin use in acute coronary care had increased from under 10% in 1987 to over 90% in 1989. Worldwide, the annual number of patients with suspected myocardial infarction who would nowadays be given such treatment must be several million a year, suggesting that aspirin is already preventing several tens of thousands of prema- ture deaths each year in this clinical context alone. However, if the ISIS-2 trial had been a factor of 10 smaller (i.e. 1700 instead of 17 000 patients), then exactly the same proportional reduction in mortality as shown in Fig. 2.4.3 would not have been conventionally signifi- cant and, therefore, would have been much less likely to influence medical practice—indeed, the result might by chance have appeared exactly flat, greatly damaging future research on aspirin in this con- text. (In fact, during the early interim monitoring of the ISIS-2 trial results by the independent Data Monitoring Committee there was no apparent difference in mortality based on a few hundred deaths.) Likewise, if the ISIS-2 trial had been nonrandomized, then it might well have produced the wrong answer since, in a nonrandomized study, doctors might tend to give active treatment to patients who are particularly ill, or who are rather different in various other ways from those not given active treatment. In addition, even if a non- randomized study did happen to produce an unbiased answer, it would have been impossible to be sure that it had actually done so, so again a nonrandomized study might have had much less influence on medical practice than ISIS-2. In the ISIS-2 trial, aspirin significantly reduced the 1-month mortality, but it also significantly reduced the

number of nonfatal strokes and nonfatal reinfarctions that were recorded in hos- pital. Combining all these three outcomes into ‘vascular events’ (i.e. stroke, death or reinfarction), 10% of those who were allo- cated aspirin versus 14% of the controls suffered a vascular event in the month after randomization (Table 2.4.3)—an absolute dif- ference of 40 events per 1000 treated (or, perhaps more relevantly, 40 000 per million). The randomized trials of aspirin, or of other antiplatelet regimens, in other types of high-risk patients (e.g. a few years of aspirin for those who have survived a myocardial infarc- tion or stroke) were not as large as ISIS-2, and so, taken separately, most yielded false-negative results. However, when the results from many such trials are combined, statistically definite reductions in ‘vascular events’ are seen (Table 2.4.3). Since such treatments do not appear to increase nonvascular mortality, all-cause mortality Cumulative number of vascular deaths 1000 800 600 400 200 0 Days from randomization 0 7 14 21 28 35 Placebo tablets: 1016 vascular deaths (11.8%) Aspirin: 804 vascular deaths (9.4%) Fig. 2.4.3 Effect of administration of aspirin for 1 month on 35-day mortality in the 1988 ISIS-2 trial among 17 000 patients with acute myocardial infarction. (Absolute survival advantage: 24 (SE 5) lives saved per 1000 patients allocated aspirin, 2p <0.00001. The COMMIT trial in 46 000 such patients has since shown aspirin plus clopidogrel to be slightly more effective than aspirin alone.) Table 2.4.3 Summary results of (a) trials of aspirin (or other antiplatelet drugs), and (b) trials of adding a P2Y12-receptor antagonist to aspirin Type of patient (and study) Mean duration (total randomized) Stroke, heart attack, or vascular death (a) Antiplatelet vs. control Antiplatelet Control Difference Acute heart attack (ISIS-2) 1 month (20 000) 10% 14% 40 per 1000 (2p <0.00001) Acute stroke (CAST) 1 month (40 000) 9% 10% 10 per 1000 (2p = 0.001) Previous heart attack (ATT) 2 years (20 000) 13% 17% 40 per 1000 (2p <0.00001) Previous stroke/TIA (ATT) 2.5 years (23 000) 18% 22% 40 per 1000 (2p <0.00001) Other high-risk (e.g. angina, peripheral vascular disease) 1 year (20 000) 8% 10% 20 per 1000 (2p <0.00001) (b) Aspirin plus P2Y12-receptor antagonist vs. aspirin alone Aspirin + P2Y12-receptor antagonist Aspirin alone Acute coronary syndrome (CURE) 9 months (13 000) 9% 11% 20 per 1000 (2p <0.001) Acute heart attack (CCS2-COMMIT) 1 month (46 000) 9% 10% 10 per 1000 (2p = 0.002) Previous heart attack (PEGASUS-TIMI 54) 33 months (21 000) 8% 9% 10 per 1000 (2p <0.001)

 is also significantly reduced. Subsequently, a combination of aspirin and a P2Y12-receptor antagonist such as clopidogrel (which inhibits platelet activation through different pathways) has been shown to be more effective than aspirin alone in acute myocardial infarc- tion or acute coronary syndrome (Table 2.4.3), and this has paved the way for ‘dual antiplatelet therapy’ to become standard for such conditions. The large-scale randomized evidence on antiplatelet drugs that is summarized in Table 2.4.3 has changed clinical practice world- wide, and may already have affected the treatment of hundreds of millions of patients in ways that, at low cost, have prevented millions of strokes, heart attacks, or vascular deaths. Small randomized trials and small meta-analyses of trials, or nonrandomized studies (how- ever large), could not possibly have provided appropriately reliable evidence about such moderate risk reductions. Evidence of hazard from a single large trial It is not sufficiently appreciated that identical arguments about the need for large-scale randomized evidence apply equally when there is the possibility that a treatment is the cause of a moderate increase in adverse outcomes. The drug niacin decreases blood low-density lipoprotein (LDL) cholesterol and triglycerides, while increasing high-density lipoprotein (HDL) cholesterol. It was shown to be ef- fective for the prevention of coronary events in the Coronary Drug Project conducted in the 1970s, but a more recent trial (AIM-HIGH) had been stopped early because of a lack of efficacy. Many patients are

unable to take niacin because it causes uncomfortable flushing and other adverse effects. The second Heart Protection Study—Treatment of HDL to Reduce the Incidence of Vascular Events (THRIVE) was designed to assess the benefits and hazards of extended-release niacin in - combination with the drug laropiprant, which had been shown to reduce flushing in up to two-thirds of patients and hence, it was hoped, would improve adherence. The trial randomized over 25 000 patients to niacin-laropiprant versus placebo and followed them for about 4 years. There was no significant effect on major vas- cular events, defined in this trial as a major coronary event (nonfatal myocardial infarction or coronary death), stroke of any type, or coronary or noncoronary revascularization (13.2% vs. 13.7%, rate ratio 0.96, 95% confidence interval 0.90–1.03, p = 0.29), but the trial also provided very important information about several adverse ef- fects of niacin. Table 2.4.4 summarizes the results from the trial for selected serious adverse events (which, essentially, means that the event caused hospitalization or, rarely, death). While some of these were known adverse effects of niacin (e.g. diabetes-related, gastrointestinal, musculoskeletal, and skin-related disorders), other adverse events that were in excess in the THRIVE trial had not been suspected. Allocation to niacin-laropiprant re- sulted in moderately raised risks of serious infections (22% rela- tive increase, corresponding to an absolute excess of 14 cases per 1000 participants treated with niacin-laropiprant for 4 years) and of bleeding (38% relative increase, corresponding to an absolute excess of 7 cases per 1000 participants treated with niacin-laropiprant for 4 years). A significant excess of serious infections was also seen with niacin alone in the smaller AIM-HIGH trial of 3500 patients (8.1% vs. 5.8, p = 0.008), but there were relatively few serious bleeding events and the excess observed with niacin alone was not significant (3.4% vs. 2.9%, p = 0.36), although it is consistent with the THRIVE result. The discovery of these new adverse effects in THRIVE, to- gether with the new information about the magnitude of known adverse effects, has led to a reappraisal of the role of niacin in the management of blood lipids. For not only is niacin now known to be ineffective at preventing vascular events (at least in circumstances similar to those tested in THRIVE), but it is also a much more haz- ardous drug than had previously been appreciated. This illustrates the value of large-scale randomized evidence for studying drug safety when, as is typical, the hazards requiring detection are of only moderate magnitude. Definite result from a very large meta-analysis
of trials: Benefit from ‘adjuvant’ therapy with
tamoxifen for patients with hormone-sensitive (ER+) ‘early’ breast cancer By definition, in ‘early’ breast cancer all detectable deposits of dis- ease are limited to the breast and the local or regional lymph nodes, and can be removed surgically. However, experience shows that undetectably small deposits of breast cancer cells may remain else- where that eventually cause clinical recurrence at a distant site, per- haps after a delay of several years, which is then usually followed by death from disease. If the original tumour was ‘ER positive’ (i.e. if the tumour cells were still expressing the oestrogen receptor pro- tein) then the distant deposits of cancer cells that spread from it be- fore it was removed may also be ER positive, and may be continually stimulated by circulating hormones. Therefore, among women who Table 2.4.4 Summary results of the HPS2-THRIVE trial: effects of niacin-laropiprant on selected serious adverse events and diabetes Event type Niacin-laropiprant (n = 12 800) Placebo (n = 12 800) Rate ratio
(95% confidence interval) Absolute excess (%) ± SE Serious adverse event (%) Gastrointestinal event 4.8% 3.8% 1.28 (1.13–1.44) 1.0 ± 0.3 Musculoskeletal event 3.7% 3.0% 1.26 (1.10–1.44) 0.7 ± 0.2 Skin-related event 0.7% 0.4% 1.67 (1.20–2.34) 0.3 ± 0.1 Infection event 8.0% 6.6% 1.22 (1.12–1.34) 1.4 ± 0.3 Bleeding event 2.5% 1.9% 1.38 (1.17–1.62) 0.7 ± 0.2 New-onset diabetes 5.7% 4.3% 1.32 (1.16–1.51) 1.3 ± 0.3 Adapted from The New England Journal of Medicine, The

HPS2-THRIVE Collaborative Group, Effects of Extended-Release Niacin with Laropiprant in High-Risk Patients, 371, 210. Copyright © 2014 Massachusetts Medical Society. Reprinted with permission.

2.4 Large-scale randomized evidence: Trials and meta-analyses of trials 61 have had breast cancer removed by surgery (or by surgery and radio- therapy), there have been many trials of ‘adjuvant’ daily treatment with tamoxifen, a drug that blocks the oestrogen receptor. Some in- volved only 1 to 2 years of treatment, some involved about 5 years, some compared 5 versus 1 to 2 years and more recent trials have compared 10 versus 5 years of tamoxifen. In total, more than 100 000 women have been randomized in several dozen such trials, some of which are still being followed-up for long-term outcomes. Taken separately, most of these tamoxifen trials have been too small to provide reliable evidence about long-term survival. However, if the results of all of them are combined in various ways, some very definite differences emerge: 1 to 2 years of tamoxifen is better than nothing, 5 years is better than 1 to 2 years, and 10 years is better still for delaying or avoiding the recurrence of ER positive breast cancer. Meta-analysis has also established that newer endo- crine therapies (such as aromatase inhibitors) reduce recurrence rates even further, although, unlike tamoxifen, they can only be used in postmenopausal women. Fig. 2.4.4a shows the results from the trials of about 5 years of tamoxifen. Allocation to active treatment produces a 13% absolute difference in the 15-year risk of recurrence (34 vs. 47%), and a 9% absolute difference in survival (25 vs. 34%; both $2p < 0.00001$). Most of the effect on recurrence is seen during the first 5 years, while tamoxifen was still continuing to be given, but most of the effect on breast cancer mortality comes after this period (Fig. 2.4.4a). Indeed, the difference in the 15-year probability of death from breast cancer is about three times as great as that seen after 5 years. Reliable assessment of the moderate improvements in long-term survival in early breast cancer that are produced by tam- oxifen (and by radiotherapy and chemotherapy) would have been impossible without such a meta-analysis of all trials, with updated follow-up data provided periodically, because each of the trials was too small on its own to answer these questions convincingly. Collaborative meta-analyses that involve hundreds of trialists from all around the world can also foster international acceptance of the totality of the randomized evidence. In the case of breast cancer this has, since the mid-1980s, helped lead to widespread adoption in many countries of a succession of improvements in treatment (earlier detection, better local control, progressively better chemo- therapy and, in ER-positive disease, progressively better endocrine therapy). These have, in aggregate, resulted in a sustained fall in na- tional mortality rates (Fig. 2.4.4b). Promising meta-analysis of small trials confirmed by large trials: benefit from fibrinolytic therapy
in acute myocardial infarction If a recent thrombus has just blocked a coronary artery, thereby causing acute myocardial ischaemia or infarction, fibrinolytic drugs (such as streptokinase or tissue plasminogen activator) can some- times rapidly dissolve the thrombus, restoring the flow of blood and reperfusing the heart muscle. These drugs were first introduced into clinical research in the late 1950s, but the trials of fibrinolytic therapy for suspected acute myocardial infarction in the 1960s and 1970s were too small to be statistically reliable (none involved even 1000 patients). So, by the early 1980s the haemorrhagic side effects were obvious, the benefits had not been convincingly dem- onstrated, and such treatments were generally considered to be definitely dangerous, probably fairly ineffective, and hence inappro- priate for routine coronary care. Although meta-analyses published in the mid-1980s of the previous small trials (which had involved a total of only c.6000 patients in 24 trials) indicated a statistically def- inite benefit, they were not really believed by cardiologists and so such treatments were still not widely used. The situation

was saved by two large randomized trials, GISSI-1 and ISIS-2, which together involved about 30 000 patients (and by their aggregation with the seven other randomized trials that each involved more than 1000 patients, yielding a total of 60 000; see next). In ISIS-2, not only were patients randomly allocated to receive aspirin or placebo tablets as described earlier (Fig. 2.4.3), but they were also separately allocated to receive intravenous streptokinase or a placebo infusion. In this 'factorial' design (which allows the separate assessment of more than one treatment without any material loss in the statistical reliability of each comparison), one-quarter of the patients were allocated aspirin alone, one-quarter were allocated streptokinase alone, one-quarter were allocated both streptokinase and aspirin, and one-quarter were allocated neither (that is, they were given placebo tablets and a pla- cebo infusion). Streptokinase, like aspirin, produced a highly signifi- cant reduction in mortality, and the combination of streptokinase and aspirin was highly significantly better than either aspirin alone or streptokinase alone (Fig. 2.4.5). The results shown in Fig. 2.4.5 might suggest that there was no need to collect any more randomized evidence about fibrinolytic therapy, but this ignores the potential hazards of such treatment and the heterogeneity of patients. Taken separately, even ISIS-2 (the largest of these trials) was not large enough for statistically re- liable subgroup analyses, but when the nine largest trials were all taken together, they included a total of about 60 000 patients, half of whom had been randomly allocated fibrinolytic therapy. Those entering a coronary care unit with a diagnosis of suspected or def- inite acute myocardial infarction range from patients who are al- ready in cardiogenic shock with low blood pressure and a fast pulse (half of whom will die rapidly) to those who have merely had a history of chest pain and no very definite changes on their electro- cardiography (ECG) (of whom 'only' a small percentage will die be- fore discharge). Fibrinolytic therapy often causes blood pressure to fall: should it be used in patients who are already dangerously hypo- tensive? It occasionally causes serious strokes: Should it be used in patients who are elderly or hypertensive, and therefore already have an above-average risk of cerebral haemorrhage (or who have only slight changes on their ECG, and therefore have only a low risk of cardiac death)? Finally, if a coronary artery has been totally occluded for long enough, the heart muscle that it supplies will have been ir- reversibly destroyed: How many hours after the heart attack starts is fibrinolytic treatment still worth risking—3? 6? 12? 24? These questions needed to be answered reliably before appropriate and generally accepted indications for and against such an immedi- ately hazardous, but potentially effective, therapy could be devised. To address them, the main fibrinolytic therapy trialists collaborated in a systematic meta-analysis of the randomized evidence, based on individual patient data. On review of the 60 000 patients random- ized between fibrinolytic therapy and control in trials of more than 1000 patients, some of the therapeutic questions were relatively easy to answer satisfactorily. For example, it appears that most of those whose ECG is still normal (or shows a pattern that indicates only a small immediate risk of death) can be left untreated, leaving open the option of starting fibrinolytic treatment urgently if their ECG changes suddenly for the worse over the next few hours or days.

62 SECTION 2 Background to medicine ≈ 5 years tamoxifen vs. Not RECURRENCE ER+ (a) 50 Control 47·3% 33·9% ≈ 5 years tamoxifen 25·0% ≈ 5 years tamoxifen Control 34·3% 15–y gain 13·4% (SE 1·1) Logrank 2p < 0·00001 15–y gain 9·3% (SE 1·1) Logrank 2p < 0·00001 40 30 20 10 0 5 16·6 26·5 8·9 12·2 26·2 18·6 41·0 29·1 10 15 years 0 5 10 15 years % ± SE 0 Recurrence 50 40 30 20 10 % ± SE 0 Breast cancer mortality ≈ 5 years tamoxifen vs. Not BREAST CANCER MORTALITY ER+ UK and USA, 1950–2014: Breast cancer mortality at ages 45–54 (b) Source: WHO mortality & UN population estimates Death rate/10 000 women, age standardized* Further

MODERATE effects are still worthwhile and achievable LARGE effect on UK/USA breast cancer mortality by combining several MODERATE effects. 1950 1970 1990 2010 0 2 4 6 UK USA 0.6% 0.4% 0.2% 0% 10-year risk

- Mean of annual rates in the two component 5–year age groups smoothed in 5-year calendar periods Fig. 2.4.4 (a) Effects of about 5 years of tamoxifen versus not in ER-positive disease: 15-year probabilities of recurrence and of breast cancer mortality (10 000 women in the 2005 worldwide EBCTCG meta-analysis). (b) Female breast cancer mortality in the United Kingdom and United States of America at ages 45 to 54 during the period 1950 to 2014. (The UK breast screening programme has little effect on mortality at these ages.)

2.4 Large-scale randomized evidence: Trials and meta-analyses of trials 63 Conversely, among those who already had ‘high-risk’ ECG changes when they were randomized, the absolute benefit of immediate fi- brinolytic therapy was, if anything, slightly greater than is indicated by Fig. 2.4.5. Age, sex, blood pressure, heart rate, diabetes, and a pre- vious history of myocardial infarction could not identify reliably any subgroup that would not, on average, have their chances of survival appreciably increased by treatment. By contrast, the longer that fibrinolytic treatment for such pa- tients was delayed, the less benefit it seemed to produce. Among the 45 000 whose ECG showed definite ST-segment elevation or bundle-branch block, the benefit was greatest (about 30 lives saved per 1000) among those randomized between 0 and 6 h after the onset of pain (Fig. 2.4.6). However, the mortality reduction was still substantial and signifi- cant (about 20 per 1000; 2p <0.003) for the patients whose hospital admission had been delayed for some hours and who were therefore randomized 7–12 h after the onset of pain. Indeed, even if patients were randomized 13–18 h after the onset of pain, there still appeared to be some net reduction in mortality (about 10 per 1000, but not statistically definite). The regression line in Fig. 2.4.6 reinforces these separate subgroup analyses in a more reliable way. Yet, before these large trials it was forcefully, but mistakenly, argued that such treatments could not possibly be of any worthwhile benefit if given more than about 3 or 4 h after the onset of symptoms. Such detailed inferences are difficult enough with large-scale properly randomized evidence and would be impossible without it. Because of their unknowable biases (see earlier), nonrandomized database analyses are simply not a viable alternative to large-scale randomized evidence. Nor would randomization of ‘only’ several thousand patients have been sufficient. The availability of large-scale randomized evidence, in this case a meta-analysis involving about 6000 deaths among 60 000 patients, has been essential in determining which particular types of patient derive net benefit from fibrinolytic therapy (or from surgical opening of acutely occluded arteries). Promising meta-analysis of small trials refuted by large trials: Lack of significant benefit from magnesium infusion in acute myocardial infarction In animal studies, infusion of a magnesium salt can limit the myo- cardial damage arising from sudden experimental blockage of a cor- onary artery. By the early 1990s, there was considerable optimism that a simple, inexpensive magnesium infusion might prove benefi- cial after acute myocardial infarction. Twelve small trials, involving between them a total of only about 2000 patients, had addressed this question, and their aggregated results indicated a highly statistically significant—but implausibly large—halving of risk (72/1199 deaths among those allocated magnesium versus 151/1191 among the controls, 2p <0.00001). At this time some argued that such results constituted proof beyond reasonable doubt that magnesium was of sufficient value to justify its widespread usage without seeking fur- ther randomized evidence, but others remained

sceptical, arguing that the apparent results were far too good to be true. Two trials, one (LIMIT-2) involving 2000 patients and one (ISIS-4) involving 58 000, were then set up to test the possible effects of mag- nesium more reliably. The first yielded a moderately promising re- sult (Fig. 2.4.7), indicating avoidance of about one-quarter of the early deaths, but with the 99% confidence interval including the pos- sibility that magnesium had no beneficial effect on early mortality. The second (which had continued in spite of intense lobbying of its Data Monitoring Committee to stop the trial), however, yielded a completely unpromising result, so that the overall evidence, by that time based on over 60 000 randomized patients, indicated no net effect on mortality. Nevertheless, some cardiologists remained hopeful that mag- nesium might prove to be effective among specific subgroups. Accordingly, the MAGnesium In Coronaries (MAGIC) trial Percentage dead 10 5 0 0 1 2 3 4 5 Weeks from starting treatment Aspirin only Streptokinase only Routine hospital care alone 13% dead (568/4300) Routine care + combination of both streptokinase and aspirin 8% dead (343/4292) Fig. 2.4.5 Effects of a 1-h streptokinase infusion, and of aspirin for about one month, on 35-day mortality in the 1988 ISIS-2 trial among
17 000 patients with acute myocardial infarction who would not normally have received either treatment, divided at random into four similar groups to receive aspirin only, streptokinase only, both or neither. 40 30 20 10 0 6 12 Hours from symptom onset to randomization 18 Loss of benefit per hour of delay to randomization: 1.6 SD 0.6 per 1000 patients 3000 14 000 12 000 9000 7000 50 24 0 Lives saved per 1000 allocated fibrinolytic therapy (± SD) Fig. 2.4.6 Benefit versus delay (0–1, 2–3, 4–6, 7–12, or 13–24 h) in the nine largest randomized trials of fibrinolytic therapy versus control in patients with acute myocardial infarction. One-month mortality results for 45 000 patients with ST elevation or bundle-branch block when randomized, showing the definite net benefit even for the 9000 randomized 7–12 hours after the onset of pain.

 subsequently randomized 6000 patients, all of whom had received reperfusion therapy within the past few hours, to magnesium versus placebo but this also found no evidence of any net benefit. It is interesting to consider what this sequence of magnesium trial results (Fig. 2.4.7) might mean for those wishing to interpret other randomized evidence. Our interpretation is that if something seems too good to be true then it probably is—or, more formally, that big benefits are much less plausible than moderate benefits. None of the 12 small trials had sufficient power to detect a moderate effect on mortality, and although their aggregated results indicated that mortality could be reduced by more than half, such an effect is too extreme to be plausible, and could be misleading even though it is highly significant. The LIMIT-2 trial then suggested that magnesium might reduce mortality by about a quarter, a result that is somewhat more plausible but not clearly significant. The success of the ISIS-4 and MAGIC trials in refuting the implausibly large benefit suggested by the 13 smaller trials reinforces our point that often, when trying to distinguish between the two medically realistic possibilities of a moderate effect or no effect, only large-scale evidence suffices. Even the LIMIT-2 trial, which recruited 2000 patients, was in retrospect too small. (Another important methodological point is that ‘random effects’ methods for meta-analysis can produce importantly wrong answers: applied to the 15 separate trials in Fig. 2.4.7, a standard ‘random effects’ meta-analysis yields a summary odds ratio of 0.67 [95% CI 0.52–0.85; 2p <0.001], suggesting—clearly incorrectly— that magnesium reduces mortality by about one-third!) Trials in their epidemiological context: Blood pressure, stroke, and heart disease Quantitative epidemiological evidence about the effects of long-term differences in risk factors such as blood pressure or blood choles- terol level can help in interpreting the results from trials of the ef- fects of reducing these risk factors for only a few years. For example, appropriate meta-

analyses of prospective observational epidemio- logical studies indicate that, throughout the range of usual systolic blood pressure in the populations studied (about 115–180 mm Hg systolic blood pressure, SBP), a lower value was associated with a lower risk of ischaemic heart disease, with no apparent ‘threshold’ below which the relationship reversed (Fig. 2.4.8). This analysis suggests that, in later middle age (60–69 years), 10 mm Hg lower Trial name 12 small trials LIMIT– 2 ISIS–4 (1995) MAGIC (2002) Total 99% or 95% confidence intervals 2853/ (8·3%) 34482 2844/ (8·2%) 34487 72/1199 90/1159 2216/29011 475/3113 472/3100 2103/29039 11 8/1157 151/1191 0·44 (SE 0·10) 0·74 (SE 0·13) 1·06 (SE 0·03) 1 ·00 (SE 0·07) 1.00 (SE 0.03) 2p>0·1; NS Magnesium better 0 0.5 1.0 1.5 2.0 Magnesium worse Treatment effect 2p > 0·1; NS, adverse Deaths/Patients Allocated magnesium Allocated control Ratio of death rates Magnesium : Control Fig. 2.4.7 Effect of a magnesium infusion on 1-month mortality among patients with acute myocardial infarction. Ratio of the death rate in the treatment group to that in the control group is plotted for each trial (as a black square with area proportional to the amount of statistical information) along with its 99% confidence interval (horizontal line). A stratified overview of the results of all these trials (and its 95% confidence interval) is represented by an open diamond. 40–49 50–59 60–69 70–79 80–89 Age at risk Effect of 10 mm Hg ↓ SBP 18% ↓ risk 27% ↓ risk 31% ↓ risk 120 140 160 180 Usual systolic BP (mm Hg) Hazard ratio (95% CI) 1 2 4 8 16 32 64 128 Fig. 2.4.8 Stroke and ischaemic heart disease mortality rate in each decade of age versus usual systolic blood pressure (SBP, mm Hg) at the start of that decade, in a systematic overview of 61 prospective studies involving 1 million adults. Hazard ratios are plotted with a group-specific CI derived only from the variance of the log hazard in that one category (including the reference category), and each square has area inversely proportional to the effective variance of the log mortality rate.

2.4 Large-scale randomized evidence: Trials and meta-analyses of trials 65 SBP is associated with about 27% less ischaemic heart disease (IHD) mortality (and about 35% less stroke mortality: Prospective Studies Collaboration, data not shown). By the mid-1990s, several trials had been conducted to deter- mine whether a few years of blood pressure reduction in middle age would reduce the risk of stroke and of coronary heart disease. Partly because of imperfect compliance, the mean difference in SBP between the treatment and control groups in these trials was only about 10 mm Hg. Even if such trial treatments would eventually produce about 27% less coronary heart disease after many years (as seen in observational studies), the effects seen within the 2 or 3 years that are available on average between randomization and death in a 5-year trial might well be somewhat smaller (perhaps only about 15%, for example). But, considered separately, none of the trials recorded enough coronary heart disease events (or enough vascular deaths) for statistically reliable assessment of a 15% risk reduction. For stroke, the trials provide direct and highly significant evi- dence that most, or all, of the risk reduction associated with 10 mm Hg lower usual SBP appears soon after the blood pressure is lowered (Fig. 2.4.9). In contrast, the significant reduction in coronary heart disease seen in the trials (16% SD 4, 95% CI 8–23%; 2p = 0.00001) seems to fall somewhat short of the difference of about 27% suggested by the observational evidence. However, the coronary heart disease reduction in the trials is still substantial and real (2p = 0.00001). Taken together, Figs. 2.4.8 and 2.4.9 suggest that antihypertensive regimens that produce differences of much more than 10 mm Hg SBP will reduce stroke by more than half and heart disease by more than a quarter. They also suggest that the proportional risk reduction produced by a given absolute reduction in SBP will be approximately independent of the initial SBP. Results from large anonymous trials are relevant to real clinical practice A clinician is used to dealing with individual patients, and may feel that the results

of large trials somehow deny their individuality. This is almost the opposite of the truth, for one of the main reasons why trials have to be large is just because patients are so different from one another. Two apparently similar patients may run entirely different clinical courses, one remaining stable and the other progressing rap- idly to severe disability or early death. Consequently, it is only when really large groups of patients are compared that the proportion of patients with truly good and bad prognosis in each can be relied on to be reasonably similar. One commonly hears statements such as: 'If the effect of a treatment isn't obvious in a couple of hundred patients then it isn't worth knowing about'. But, the accumulation since 1980 of large-scale randomized evidence of such moderate effects with treatments for heart disease, stroke, breast cancer, intestinal cancer and various other conditions has transformed medical practice, and may already have avoided millions of catastrophic disabilities, recur- rences of cancer, and premature deaths. It is also said that what is really wanted is not a blanket recommen- dation for everybody, but rather some means of identifying those few individuals who really stand to benefit from therapy. If any cri- teria (e.g. a short-term response to a nonplacebo-controlled course of some disease-modifying agent) can be proposed that are likely to discriminate between people who will and will not benefit, then these can be recorded prospectively at entry and the eventual trial result subdivided with respect to them. However, there is a danger in too detailed an analysis of the apparent responses of small sub- groups chosen for separate emphasis because of the apparently re- markable effects of treatment in these subgroups. Even if an agent brought no benefit, it would have to be acutely poisonous for it not to appear disproportionately beneficial in one or two such subgroups! Conversely, if an intervention really avoids an approximately similar proportion of the risk in each category of patient, it will, by chance alone, appear not to work in some category or categories of patient. The surprising extent to which this happens is evident from the ex- ample in Table 2.4.2. A large anonymous trial will at least still help to answer the practical questions of whether, on average, a policy of widespread treatment (except where clearly contraindicated) is pref- erable to a general policy of no immediate use of the treatment (ex- cept where clearly indicated). Moreover, without really large trials it is difficult to see how else many such questions relating to the effects of treatments on death or disability (or other major outcomes) are to be resolved reliably. Trials are at least a practical way of making some solid progress, and it would be unfortunate if desire for the perfect (that is, knowledge of exactly who will benefit from treat- ment) were to become the enemy of the possible (that is, knowledge of the average direction and approximate size of the effects of treat- ment in many large categories of patient). Trial (or group of trials) Numbers of events treat : control Odds ratios & 95% confidence limits (treat : control) TREATMENT ←BETTER TREATMENT WORSE→ (i) Strokes (ii) Coronary heart disease (CHD) events 38% SD 4 reduction 2p < 0.00001 16% SD 4 reduction 2p = 0.00001 STROKE 35–40% LOWER CHD 20–25% LOWER Reductions in risk associated epidemiologically with a LONG-TERM difference of 5–6 mm Hg DBP: 0.5 1.0 HDFP trial MRC 35–64 trial SHEP trial MRC 65–74 trial 13 other trials ALL TRIALS 102:158 60:109 105:162 101:134 157:272 525:835 275:343 222:234 104:142 128:159 205:226 934:1104 HDFP trial MRC 35–64 trial SHEP trial MRC 65–74 trial 13 other trials ALL TRIALS (Heterogeneity $\chi 4\ 2 = 4.2$, NS) (Heterogeneity $\chi 4\ 2 = 4.2$, NS) Fig. 2.4.9 Reduction in the odds of stroke and coronary heart disease in all unconfounded randomized trials of antihypertensive drug treatment (mean systolic blood pressure differences of about 10 mm Hg for 5 years). Conventions are as for Fig. 2.4.7.

FURTHER READING Adjuvant Tamoxifen: Longer Against Shorter (ATLAS) Collaborative Group (2013). Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years after diagnosis of oestrogen receptor-positive breast cancer: ATLAS, a randomised trial. Lancet, 381, 805–16. Antithrombotic Trialists' Collaboration

(2002). Collaborative meta- analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high-risk patients. BMJ, 324, 71–86. Asymptomatic Carotid Surgery Trial (ACST) (2010). 10-year stroke prevention after successful carotid endarterectomy for asymptom- atic stenosis (ACST-1): a multicentre randomised trial. Lancet, 376, 1074–84. Chalmers I (1994). The Cochrane Collaboration: preparing, maintain ing and disseminating systematic reviews of the effects of health care. Ann N Y Acad Sci, 703, 156–63. Chen ZM et al. for the COMMIT (ClOpidogrel and Metoprolol in Myocardial Infarction Trial) collaborative group (2005) Addition of clopidogrel to aspirin in 45,852 patients with acute myocar- dial infarction: randomised placebo-controlled trial. Lancet, 366, 1607–21. Cochrane AL (1979). 1931–1971: a critical review, with particular ref- erence to the medical profession. In: Teeling-Smith G, Wells N (eds) Medicines for the year 2000, pp. 1–11. Office of Health Economics, London. Collins R, MacMahon S (2001). Reliable assessment of the effects of treatment on mortality and major morbidity I: clinical trials. Lancet, 357, 373–80. Collins R, Peto R (1994). Antihypertensive drug therapy: effects on stroke and coronary heart disease. In: Swales JD (ed) Textbook of hypertension, pp. 1156–64. Blackwell Science, Oxford. Collins R, et al. (1987). Avoidance of large biases and large random errors in the assessment of moderate treatment effects: the need for systematic overviews. Stat Med, 6, 245–50. Collins R, Doll R, Peto R (1992). Ethics of clinical trials. In: Williams CJ (ed) Introducing new treatments for cancer: practical, ethical and legal problems, pp. 49–65. John Wiley & Sons Ltd, Chichester. Dowsett M, et al. (2010). Meta-analysis of breast cancer outcomes in adjuvant trials of aromatase inhibitors versus tamoxifen. J Clin Oncol, 28, 509–18. Early Breast Cancer Trialists' Collaborative Group (EBCTCG) (2005). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the ran- domised trials. Lancet, 365, 1687–717. Early Breast Cancer Trialists' Collaborative Group (EBCTCG) (2015). Aromatase inhibitors versus tamoxifen in early breast cancer: patient-level meta-analysis of the randomised trials. Lancet, 386, 1341–52. Fibrinolytic Therapy Trialists' Collaborative Group (1994). Indications for fibrinolytic therapy in suspected acute myocardial infarction: collaborative overview of early mortality and major morbidity re- sults from all randomised trials of more than 1000 patients. Lancet, 343, 311–22. HPS2-THRIVE Collaborative Group (2014). Effects of extended re- lease niacin with laropiprant in high-risk patients. N Engl J Med, 371, 203–12. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group (1988). Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myo- cardial infarction: ISIS-2. Lancet, 332, 349–60. ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group (1995). ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sul- phate in 58050 patients with suspected acute myocardial infarction. Lancet, 345, 669–85. Law M, Morris J, Wald N (2009). Use of blood pressure lowering drugs in the prevention of cardiovascular disease: meta-analysis of 147 randomised trials in the context of expectations from prospective epidemiological studies. BMJ, 338, b1665. MacMahon S, Collins R (2001). Reliable assessment of the effects of treatment on mortality and major morbidity II: observational studies. Lancet, 357, 455–62. Magnesium in Coronaries (MAGIC) Trial Investigators (2002). Early administration of intravenous magnesium to high-risk patients with acute myocardial infarction in the Magnesium in Coronaries (MAGIC) Trial: a randomised controlled trial. Lancet, 360, 1189–96. MRC Asymptomatic Carotid Surgery Trial (ACST) Collaborative Group (2004). Prevention of disabling and fatal strokes by successful carotid endarterectomy in patients without recent neurological symptoms: randomised controlled trial. Lancet, 363, 1491–502. Prospective Studies Collaboration (2002). Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of indi- vidual data for one million adults in 61 prospective studies. Lancet, 360, 1903–13. Prospective

Studies Collaboration (2007). Blood cholesterol and vas- cular mortality by age, sex and blood pressure: meta-analysis of indi- vidual data from 61 prospective studies with 55,000 vascular deaths. Lancet, 370, 1829–39. Woods KL, et al. (1992). Intravenous magnesium sulphate in sus- pected acute myocardial infarction: results of the second Leicester Intravenous Magnesium Intervention Trial (LIMIT-2). Lancet, 339, 1553–58. Yusuf S, Collins R, Peto R (1984). Why do we need some large, simple randomized trials? Statistics in Medicine, 3, 409–20.

Revision #1
Created 2026-01-22 16:41:01 UTC by Omar Ayman
Updated 2026-01-22 16:41:02 UTC by Omar Ayman