Harveian Oration

De testimonio: on the evidence for decisions about the use of therapeutic interventions

Michael Rawlins

Lancet 2008; 372: 2152–61

National Institute for Health and Clinical Excellence, London, UK; Emeritus Professor, University of Newcastle, Newcastle upon Tyne, UK (Prof Sir M Rawlins FRCP)

Correspondence to: Prof Sir Michael Rawlins, National Institute for Health and Clinical Excellence, 71 High Holborn, London WC1V 6NA, UK **michael.rawlins@nice.org.uk**

William Harvey (1578–1657) was one of a group of 17th century natural philosophers who were no longer prepared to accept the authority of Aristotle, Plato, and Galen as a reliable basis for understanding the natural world. As Harvey himself put it "It is base to receive instructions from others' comments without examination of the objects themselves, especially as the book of nature lies so open and is so easy of consultation."[1]

Although united in their quest to examine "the book of nature" for themselves, natural philosophers of the period were bitterly divided about how it should be done. 350 years later this dispute about the nature of science, and scientific method still persists, particularly in relation to the inductive and deductive approaches to the establishment of scientific knowledge.[2]

Nowhere though is how best to establish scientific knowledge more hotly, and sometimes bitterly, argued than in the nature of the evidence that should support the use of therapeutic interventions. The debate surrounding evidence to support therapeutic use has become particularly apparent with the emergence, over the past 30 years, of what are known variously as rules, levels, or hierarchies of evidence (table 1).[3]

Evidence, in the present context, has only one purpose: it forms the basis for informing decision makers about the appropriate use of therapeutic interventions in routine clinical practice. Such decisions have to be made at various levels but, invariably, with critical consequences for patients, families, and society. Decision makers, for example, determine the appropriateness of treatments that are offered to individual patients; they decide on the range of products to include in a local hospital's formulary; and they may be charged with assessing whether particular interventions are sufficiently safe and effective—as well as cost effective—to be available to entire healthcare systems. Mistakes in decision making may have dramatic repercussions.

Hierarchies place randomised controlled trials (RCTs) at their summit, with various forms of observational studies nestling in the foothills. They are used—as a form of shorthand—to provide some intimation of the strength of the underlying evidence and, particularly by guideline developers, to then grade therapeutic recommendations on the basis of this perceived strength.

The notion that evidence can be reliably placed in hierarchies is illusory. Hierarchies place RCTs on an uncomfortable pedestal[4] for, as I discuss, although the technique has advantages it also has disadvantages. Observational studies have defects but they also have merit. Decision makers need to assess and appraise all the available evidence irrespective of whether it has been derived from randomised controlled trials or observational studies; and the strengths and weaknesses of each need to be understood if reasonable and reliable conclusions are to be drawn. Nor, in reaching these conclusions, is there any shame in accepting that judgments are required about the fitness-for-purpose of the components of the evidence base. On the contrary, judgments are an essential ingredient of most aspects of the decision-making process.[5]

Randomised controlled trials

The introduction of randomised controlled trials (RCTs) in the middle of the 20th century, has had a profound effect on the practice of medicine and its essential features are well described.[4,6,7] An RCT involves comparing the effects of two (or more) interventions that have been allocated randomly to groups of contemporaneously treated patients.

Double-blind RCTs, when properly done and analysed, unquestionably provide confidence in the internal validity of the results[6,8] in so far as the benefits of the intervention are concerned; and the more so if replicated by subsequent studies. Consequently, RCTs are often called the gold standard for demonstrating (or refuting) the benefits of a particular intervention. Yet the technique has important limitations of which four are particularly troublesome: the null hypothesis, probability, generalisability, and resource implications.

The null hypothesis

The analysis of RCTs has traditionally been based on the null hypothesis, which presumes there is no difference between treatments. The null hypothesis is tested by estimating the probability (the frequency) of obtaining a result as extreme, or more extreme, as the one observed were there no difference. If the probability is less than some arbitrary value—usually less than 1 in 20 (ie, $p<0{\cdot}05$)—then the null hypothesis is rejected. This so-called frequentist approach to the design and analysis of RCT has undoubted attractions: the statistical calculations are simple; the methodology has become widely accepted; and the criteria for significance are well established.

The null hypothesis may be irrelevant, though, if there have been previous studies showing that a particular treatment has some benefits. Such a situation can occur during the development of a new drug when preliminary evidence of proof of principle from phase II studies is investigated in larger groups of patients during phase III: at that point, basing the analysis of the results of subsequent phase III studies on the null hypothesis becomes counterintuitive. Equally, the null hypothesis is inappropriate when previously published studies have already shown benefit. Yet surveys over the past 10 years show that 73% of RCTs, published in major journals, persistently fail to make any systematic attempt to set their results in the context of previous investigations.[9]

The null hypothesis is even more awkward for trials seeking to show whether there is no difference (equivalence), no less benefit (non-inferiority), or not less than a prespecified difference (futility), between treatment groups.[10] All require prior assumptions to be made about the extent to which the differences between treatments might be relevant or important.

The null hypothesis may, indeed, be methodologically consistent with the deductive approach to science but, as Rothman[11] puts it: "To entertain the universal null hypothesis is, in effect, to suspend belief in the real world and thereby to question the premises of empiricism."

Probability

In the frequentist approach, if the p value is sufficiently small either the null hypothesis is false or a very rare event has occurred. By convention, a probability of less than 5% (ie, $p<0\cdot05$) is generally used to distinguish between these two possibilities. However, a p value of greater or less than 0·05 neither disproves or proves (respectively) the null hypothesis. Some, though not all the problems with p values can be avoided by expressing results as confidence intervals, which indicate the degree of uncertainty or lack of precision of the estimate of interest. Nevertheless, p values and confidence intervals are closely related.

The difficulties in interpreting frequentist p values become convoluted when seeking to decide, during a clinical trial, whether a study should be terminated prematurely; or how (and whether) to assess outcomes in subgroups of patients once the trial has been completed. A similar problem, which is also discussed later, occurs during the safety analysis of RCTs. In all these instances, repeated tests of statistical significance—adopting the conventional p value ($<0\cdot05$)—is increasingly likely to produce one or more falsely significant results. If, for example, ten separate assumptions are tested, the probability of one being apparently significant (at $p<0\cdot05$) is 40%. This is known as the problem of multiplicity. There are, however, very divergent views among statisticians as to how to deal with this difficulty both in devising stopping rules and in subgroup analyses.[12]

There is a natural desire for investigators, during the course of an RCT, to undertake interim analyses of the accruing data in order to decide whether a trial should continue or be prematurely stopped. Premature termination may be justified on the grounds that the study has already achieved its predefined beneficial endpoints or because of safety concerns in one of the groups. There are, however, serious pitfalls in deciding whether and when to terminate a trial early. If an interim analysis shows an unexpected benefit, it may be difficult to distinguish a true effect from chance (a so-called random high).[13]

	Criteria
1++	High quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias
1+	Well conducted meta-analyses, systematic reviews of RCTs, or RCTs with a low risk of bias
1–	Meta-analyses, systematic reviews of RCTs, or RCTs with a high risk of bias
2++	High-quality systematic reviews of case–control studies or cohort studies; or high quality case–control or cohort studies with a very low risk of confounding, bias, or chance
2+	Well conducted case–control or cohort studies with a low risk of confounding, bias, or chance
2–	Case–control or cohort studies with a high risk of confounding, bias, or chance
3	Non-analytic studies (eg, case reports, case studies)
4	Expert opinion

Table 1: **Levels of evidence**[3]

Various statistical approaches have been developed to resolve this form of multiplicity[12]. Many depend on changing the level of statistical significance as each interim analysis is done, so that for earlier examinations of the data a lower p value is required to reject the null hypothesis. There is, however, no consensus among statisticians about stopping rules.[12,14] A resolution to the problem has, however, become urgent. Because stopping trials early, for benefit, may systematically overestimate treatment effects[12,15] there is a real danger that some claims for efficacy—especially in oncology—may be unwarranted.

Analyses of the effects of an intervention, in subgroups of patients, can be important to establish whether different types of people respond differently.[16] The most common solution to multiplicity in subgroup analyses is to accept as reliable only a limited number of clinically or biologically plausible ones that have been prespecified during the planning stage.[12,17] A definition of what might be regarded as limited is not generally offered. Opinions vary in the assessment of subgroups identified after a trial has been completed. Some eschew post-hoc analyses altogether, whereas others suggest cautious statistical adjustment of p values.[17]

A growing number of statisticians[18] believe that the solution to many of the difficulties inherent in the frequentist approach to the design, analysis, and interpretation of RCTs is the greater use of bayesian statistics. The bayesian approach to probability is named after

Thomas Bayes (1701–61) who was a non-conformist minister in Tunbridge Wells. This notion of probability—subjective or inverse probability—is the likelihood of a hypothesis in view of some data. Thus, although the frequentist approach is about the probability of some data conditional on a specific hypothesis (usually the null hypothesis), the bayesian approach is the reverse (ie, the probability of a hypothesis conditional on the data).

Bayes' theorem relates the probabilities from what is known before (a priori) an experiment—such as an RCT—to the probabilities recalculated after the experiment (a posteriori). The link between the prior and posterior probabilities is the result of the experiment itself. The posterior probability provides an estimate of the probability of a hypothesis conditional on the observed data but taking account of what was already known (the prior) before the experiment was done.[19,20]

Figure 1 shows an application of a bayesian approach, to the analysis of an RCT. The GREAT trial[21] was designed to test the hypothesis that early domiciliary thrombolytic therapy for acute myocardial infarction would be better than later treatment in hospital. The investigators therefore undertook an RCT comparing the effectiveness of thrombolysis, given by general practitioners in patients' own homes, with later treatment once they had reached their local hospital. At 3 months, the relative reduction in all-cause mortality was 49% (p=0·04), for patients treated at home compared with those treated only when they had reached hospital. Although early thrombolysis might well have had survival advantages, a reduction of almost 50% seemed implausible given that hospital thrombolytic therapy, itself, reduces mortality by about 25%.

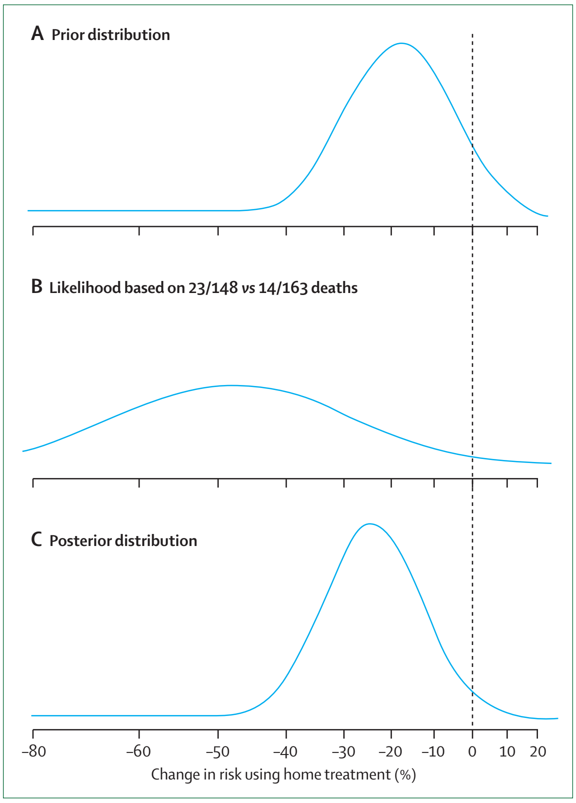

***Figure 1:* Bayesian reanalysis of the GREAT trial showing reduction in mortality with home thrombolytic therapy compared with treatment in hospital**
Prior-probability distribution of home treatment (A). Likelihood-probability distribution from the GREAT trial (B). Posterior-probability distribution of home treatment with Bayes' theorem (C). Reproduced with permission from the British Medical Association.[22]

Pocock and Spiegelhalter[22] therefore undertook a bayesian reanalysis (figure 1) of the GREAT trial. They derived a prior distribution, on the basis of the results of previous RCTs of hospital treatment with thrombolytics, but expressing their belief that a 15–20% reduction in mortality was highly plausible but that extremes of no benefit and a 40% reduction were both unlikely.

The likelihood derived from the original analysis of the GREAT study has been pulled back to provide a formal representation of the belief that the original results were "too good to be true".[22]

As well as avoiding the indiscriminate use of the null hypothesis, bayesian approaches are claimed to overcome problems in the design and analysis of RCTs as well as issues relating to multiplicity.[19,20] Why then are bayesian methods not more commonly used? There seem to be five main reasons.

First, although the subjective approach to probability dates back to the 18th century,[20] some (especially those of a frequentist mindset) regard this interpretation of probability—as a personal belief or judgment—with distaste. They prefer the apparent (but illusory) security of a clear definition of what constitutes an extreme result when tested against the null hypothesis; and they are reluctant to accept that personal belief or judgment should come into play in decision making.

Second, there have been substantial controversies about the derivation of the prior probability. Where there is evidence from previous studies, a so-called clinical prior, is readily available. Where there is no clinical prior, use is made of default priors.[20] Too much has been made of the alleged difficulties in using default priors and bayesians tend to use several of these in the absence (and even, sometimes, in the presence) of clinical priors as part of their sensitivity analyses.

Third, bayesian analyses are computationally complex and are more demanding than the methods used in most frequentist analyses.

Fourth, some statisticians—albeit a dwindling number—are unfamiliar with the techniques of bayesian analysis and are unwilling (or unable) to adapt. Some[23] attribute this variation in skill-mix to a statistician's original choice of university. Others—less kindly—believe it to be generational. As one bayesian explained to me "statisticians who were taught how to use log books and slide rules can't usually do bayesian statistics."

Finally, regulatory authorities have sometimes been hesitant to concede that bayesian approaches might have advantages.[24] There are, however, signs of rising interest particularly in the evaluation of devices.[25] And manufacturers are themselves increasingly adopting bayesian approaches in phase II and III trials.[26,27]

Bayesian approaches to the design and analysis of RCTs are likely to play a much greater part in the future.[18] Eliminating the rigidity of the p value and resolving some of the questions about multiplicity are prizes worth securing. Above all, bayesian approaches might help decision makers draw more appropriate conclusions.

Generalisability

RCTs are generally done in selected populations of patients for a finite—usually relatively brief—period of time. In clinical practice the intervention is likely be used in a more heterogeneous population of patients—often with comorbid illnesses—and frequently for much longer periods. The extent to which the findings from RCTs have external validity and can be extrapolated—or generalised—to wider populations of patients[28,29] has become an increasingly important issue. Table 2 outlines the most important drawbacks.

That there are real concerns over the issue of generalisability is discussed in greater detail elsewhere.[10] Bartlett and colleagues,[30] for example, reviewed the exclusion criteria adopted in RCTs of both statins (27 trials) and non-steroidal anti-inflammatory agents (25 trials). They noted under-representation of women, older people, and ethnic minorities, compared with use in the general population. Similar under-representation has been noted in RCTs of other cardiovascular interventions.[31]

There is therefore uncertainty as to whether the benefits achieved by average patients in RCTs can be extrapolated to average patients undergoing routine clinical care. Does, for example, the under-representation of certain groups in RCTs really matter? Some people presume that the results of RTCs in discrete populations can be reliably extrapolated to the care of patients in general.[7,32] It is argued that, if the pathogenesis of a disease is the same in all subgroups, similar benefits can be expected in wider populations of patients.

The difficulty with this claim is that there is little systematic evidence to support it[29] and some[32] that refutes it. There are, unquestionably, individual studies demonstrating concordance between the beneficial effects seen in RCTs with those noted during conventional medical care. The benefits of anticoagulation in patients with non-valvular atrial fibrillation are a case in point.[33] But the extent to which the differing characteristics of patients treated in RCTs, compared with those undergoing routine clinical care, really matters in relation to the claimed benefits remains unknown. Indeed, as the authors of the CONSORT statement themselves admit,[7] external validity is a matter of judgment.

Although there is optimism, albeit with a fair degree of uncertainty about the generalisability of the results of RCTs in relation to efficacy, experience shows that in the assessment of harms RCTs are weak at providing relevant evidence. RCTs may, as discussed later, detect dramatic safety issues; but they are an unreliable approach.

The custom in clinical trials is to collect and record all the adverse events occurring after randomisation. These data reduce the chance of investigator bias in interpreting the causal nature of any intercurrent illnesses that some patients will inevitably develop during the course of a study. Adverse events include abnormal symptoms and signs, abnormalities detected by routine clinical biochemical tests (full blood counts, urea and electrolytes, liver-function tests, urinalysis, etc), and the results of special monitoring (eg, electrocardiography, echocardiography). Those adverse events causally related to the intervention can, in theory, be identified by simple group comparisons. Although this approach has superficial attractions, there are several problems.

RCTs are designed to ensure that the statistical power will be sufficient to show clinical benefit. Such power calculations do not, however, usually take harms into account.[34] As a consequence, although RCTs can identify the more common adverse reactions, they fail to recognise less common ones or those with a long latency (such as malignant diseases). Most RCTs, even for interventions that are likely to be used by patients for

	Potential problems
Patients	
Age	Effectiveness in younger or older patients
Sex	Effectiveness generally
Severity of the disease	Effectiveness in mild or severe forms of the condition
Risk factors	Effectiveness in patients with risk factors for the condition (eg, smokers)
Comorbidities	Influence of other conditions on effectiveness
Ethnicity	Effectiveness in other ethnic groups
Socioeconomic status	Effectiveness in disadvantaged patients
Treatment	
Dose	Too high a dose used in RCTs
Timing of administration	Influence on adherence (compliance) to treatment regimens
Duration of therapy	Effectiveness during long-term use
Comedication	Adverse interactions
Comparative effectiveness	Effectiveness in comparison with other products used for the same indication
Setting	
Quality of care	Prescription and monitoring by less specialist (expert) healthcare providers

Table 2: **Issues that adversely affect generalisability of results of RTCs**

many years, are only of 6–24 months duration. And, if adverse events are detected at a statistically significant level, it is easy to dismiss them as being due to chance rather than a real difference between the groups.

The analysis of RCTs, for harms, thus poses yet another unresolved multiplicity problem[34]. In large-scale, long-term studies some statistically significant effects will almost inevitably be observed. Distinguishing those that are iatrogenic from those that are intercurrent and non-causal, or just random error, is as much an art as a science. Where the events are typically iatrogenic (eg, anaphylaxis, morbilliform rashes, toxic epidermal necrolysis) a causal relation can be inferred. Similarly, if the adverse events are biologically plausible (eg, breast cancer with hormone replacement therapy), a causal relation might also be inferred. Where these factors do not apply, difficulties in interpretation may arise. Properly done and analysed RCTs can certainly, provide important information about adverse effects. Examples include RCTs of prophylactic antiarrhythmic therapy, with class 1 agents, after myocardial infarction,[35] and of hormone replacement therapy in postmenopausal women.[36] These, though, are exceptions.

Resources

The costs of RCTs are substantial in money, time, and energy. Figure 2 shows the range of costs of 153 RCTs that were completed in 2005–06. These data combine the costs of trials that were funded by the UK National Institute for Health Research and the UK Medical Research Council as well as those incurred by three major pharmaceutical companies in their phase II and III studies. The median cost was £3 202 000 with an interquartile range of £1 929 000 to £6 568 000.

These data are neither comprehensive nor, necessarily, representative of RCTs generally; but they demonstrate that trials can be very expensive undertakings. Costs, too, seem to be rising. One manufacturer estimates that the average the cost per patient, included in trials, has increased from £6300 in 2005 to £7300 in 2006 and £9900 in 2007.

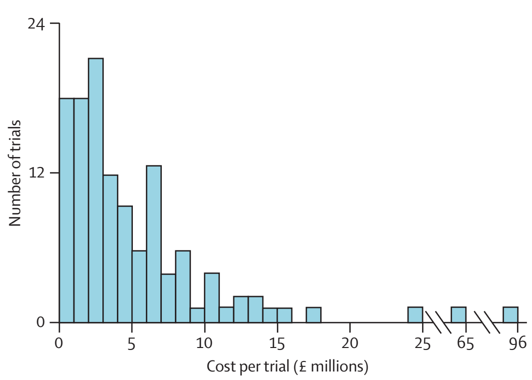

Figure 2: **Range of study costs of individual randomised controlled trials of pharmaceuticals**

Much of these costs is due to the increasing regulatory (and other) requirements imposed on both privately and publicly funded trials over the past few years.[37] Each measure was introduced with the best of intentions. These intentions included the desire to protect patients from unscrupulous investigators and sponsors; to ensure the collection and timely reporting of adverse event data during trials; to audit individual case report forms thus avoiding the consequences of untruthful behaviour by investigators; and so on. But even simple studies, with products that have been available for many years, now place a massive bureaucratic challenge on potential sponsors and investigators irrespective of whether they are based in universities or in the private sector.

Recent proposals by academic clinical investigators indicate that clinical trial costs could be decreased by between 40% and 60%[38] without detriment to their quality. Simple measures to reduce the bureaucratic burden, such as electronic data capture, reduction in the length of case management forms, and modified site management practices, would substantially reduce costs.

Observational studies

The nomenclature describing observational (ie, non-randomised) studies is confused. I eschew a distinction between controlled and uncontrolled studies because all observational studies involve some form of implicit (informal) or explicit (formal) comparisons. Nor do I consider the terms cohort studies or quasiexperimental studies particularly illuminating. Cohort studies include studies that are, in reality, distinct entities, and quasi-experimental is a term that I have never found to be adequately or consistently defined. The panel shows the types of observational studies that have been, and continue to be, used in deriving evidence about the benefits and harms of therapeutic interventions.

The great strength of RCTs is that the allocation of treatments is random so that the groups being compared are similar for baseline factors. In controlled observational trials there is, however, the real danger of selection bias and confounding.[39] There is, indeed, an extensive and sometimes disputatious body of evidence comparing the merits and demerits of randomised and observational studies of the effectiveness of therapeutic interventions.[10]

Attempts at systematic reviews of published comparisons between the two approaches, however, have been bedevilled by two problems. First, there is the difficulty in identifying relevant studies. Because many observational studies have not been consistently tagged in electronic bibliographic databases, it is difficult to ensure that conventional search strategies have identified them in an unbiased manner. Many reviewers have, therefore, relied on personal collections of papers, their own (or others') memories, or studies identified in previous systematic reviews. The possibility of reviewer bias is therefore substantial. The second difficulty is that very few of these reviews have distinguished between the various types of observational designs.[40]

There is general agreement that so called striking effects can be discerned without the need for RCTs.[41–43] There is, though, much less of a consensus about the role of observational studies in defining benefit when the effect size is more modest.[31] There may well be a tendency for observational studies to provide larger treatment effects than do RCTs; although this is not always so. Indeed, in some instances underestimates as well as overestimates have been reported. The magnitude of differences of between RCTs and observational data may also vary with the specific type of design used in the observational studies.[40] Analytical strategies to reduce the effects of selection bias and confounding in observational studies are discussed elsewhere.[44]

Two types of observational study are considered in detail here, because they have been especially important in providing evidence on the benefits and harms of therapeutic interventions. A fuller discussion of the others, which have also made important contributions, can be found elsewhere.[10]

Historical controlled trials

Table 3 lists examples of interventions of unquestioned benefit, as demonstrated by historical controlled trials where comparisons are made between a new intervention and past experience with the condition. In the past, the use of historical controls has been much criticised.[6] During the late 1980s, however, clinical trialists became less hostile to the concept. Prompted by the emerging AIDS epidemic they accepted[45] that some of the traditional approaches to clinical trial design were unnecessarily rigid. Byar and colleagues[45] proposed that historical controlled trials, in support of claims of efficacy for a new drug to treat AIDS, should meet the following specific requirements: there must be no other treatment appropriate to use as a control; there must be sufficient experience to ensure that the patients not receiving treatment will have a uniformly poor prognosis; the therapy must not be expected to have substantial side-effects that would compromise the potential benefit to the patient; there must be a justifiable expectation that the potential benefit to the patient will be sufficiently large to make interpretation of the results of a non-randomised trial unambiguous; and the scientific rationale for the treatment must be sufficiently strong that a positive result would be widely accepted.[45]

My own adaptation of these requirements to historical controlled trials more generally are unashamedly influenced by the considerations outlined by Bradford Hill[46] in distinguishing causal from non-causal associations in epidemiological studies. I therefore believe that historical controlled trials should generally be accepted as evidence for effectiveness provided they meet all of the following conditions. The treatment should have a biologically plausible basis—that is met by all the treatments shown in table 3. There should be no appropriate treatment that could be reasonably used as a control—the term appropriate would exclude, for example, the use of bone marrow transplantation as a control for enzyme replacement therapy in the treatment of Gaucher's disease. The condition should have an established and predictable natural history—I prefer this phraseology to "poor prognosis", since conditions such as port wine stains may greatly impair patients' quality of life without threatening life expectancy. The treatment should not be expected to have adverse effects that would compromise its potential benefits. There should be a reasonable expectation that the treatment effect will be large enough to make the interpretation of the benefits unambiguous—a signal-to-noise ratio of ten or more is strongly suggestive of a genuine therapeutic effect;[41,43] the magnitude of the signal-to-noise ratio representing a dramatic (ie, 10-fold) effect, however, is based on impression and is not (at present) supported by any substantive empirical evidence.

Panel: Types of observational studies

Historical controlled trials
Studies of effects of an intervention among a group of patients treated with an intervention compared retrospectively with a group who had previously received standard therapy (including best supportive care)

Non-randomised contemporaneously controlled trials
A comparison of the outcome(s) of patients receiving one treatment compared with another group of patients (untreated, or treated with an alternative intervention) during the same period

Case–control study
A comparison of the use of an intervention in groups of patients with and without a particular disease or condition

Before and after designs
Observations in groups of patients before and after treatment with an intervention in which patients act as their own controls; this technique has often been used in implicit historical controlled trials where the natural history of the disease or condition is well established and predictable

Case series
The outcomes of a group (series) of patients treated with an intervention during routine clinical practice; although there is no formal control group, implicit or explicit comparisons are invariably made

Case reports
Case reports (anecdotes) of harms to individual patients either reported in a publication or to a central agency (eg, the UK Medicines Healthcare products Regulatory Agency or the US Food and Drugs Administration)

	Indication
Thyroxine (1891)	Myxoedema
Insulin (1922)	Diabetic ketoacidosis
Vitamin B12 (1926)	Pernicious anaemia
Sulphonamides (1937)	Puerperal sepsis
Penicillin (1941)	Lobar pneumonia
Defibrillation (1948)	Ventricular fibrillation
Streptomycin (1948)	Tuberculous meningitis
Ganglion blockers (1959)	Malignant hypertension
Heimlich manoeuvre (1975)	Laryngeal obstruction by a foreign body
Cisplatin plus vinblastine and bleomycin (1977)	Disseminated testicular cancer
Acetylcysteine (1979)	Paracetamol poisoning
Ganciclovir (1986)	Cytomegalovirus retinitis
Laser treatment (2000)	Removal of port wine stains
Imatinib (2002)	Chronic myeloid leukaemia

***Table 3:* Some interventions with effectiveness established through historical controlled trials[3]**

In the future, there will be circumstances when we must continue to be prepared to accept evidence of benefits from historical controlled trials. Interventions falling into this category might, for example, include treatments that completely arrest the progressive neurodegeneration seen in Creutzfeldt-Jakob disease or Huntington's disease. In both these conditions objective, as well as subjective, measures are available to confirm (or refute) claims that progression has been arrested. The fact that clinical investigators in Canada, Europe, and the USA are currently accumulating cohorts of patients with these diseases[10]—specifically for the purpose of providing historical controls for future studies—gives me optimism.

Case–control studies

Case–control studies compare the use of an intervention in groups with and without a particular disease or condition. These studies, like other observational designs, provide information about an association with exposure to a particular intervention but do not necessarily show whether the relation is causal. The problems of selection bias and confounding are no less relevant to the interpretation of case–control studies than they are with other controlled observational designs. They can, however, be lessened by care in their design and analysis.[34]

Case–control studies have been used, though with mixed results, to provide support for demonstrating the benefits of interventions. During the 1980s some observational (mainly case–control) studies suggested that the long term use of hormone replacement therapy was associated with a substantial reduction in ischaemic heart disease. Quantitative overviews in the early 1990s[47,48] indicated that the relative risk in users, compared with non-users, might be associated with a reduction of as much as 50%. Hormone replacement therapies therefore became the most widely prescribed drugs in the USA.[49]

It is now known from the results of several large, well-conducted, RCTs that hormone replacement therapies have no beneficial effect on ischaemic heart disease and that they increase the risk of stroke.[36] The discrepancies between the results of observational studies and RCTs, in the perceived benefits of hormonal replacement therapy, were largely due to selection bias. If the observational studies had taken account of age, socioeconomic status, smoking habits, and duration of use most of the claimed advantages would have disappeared.[50] Some women, though, have paid a high price for this error.

There have, however, been circumstances where case–control studies have provided reliable indications of the benefits of interventions. These include the protective effects of aspirin against acute myocardial infarction,[51] the prevention of neural tube defects by folate,[52] the relation between sleeping posture and sudden infant death syndrome,[53] and the protective effects of non-steroidal anti-inflammatory drugs and colorectal cancer.[54]

We need to develop approaches that allow us to be confident that the results of observational studies generally, and case–control studies in particular, can provide information that permits reasonable assumptions about internal validity.[55] Newer techniques, such as mendelian randomisation,[56] may well assist. More resources, time, and energy to undertake methodological research are needed if causality is to be more securely based on observational evidence.

By contrast with the difficulties in assessing the benefits of interventions with case–control designs, this method has been important in identifying causal relations between specific interventions and their adverse effects (table 4).

Case–control studies have also been useful in providing reassurance that putative adverse effects signalled by spontaneous reporting schemes (see later) do not seem problematic. Examples include suspected associations between bisphosphonates and atrial fibrillation[57] and sympathomimetic bronchodilators with excess asthma deaths.[58]

Selection bias and confounding by indication may still occur in case–control studies designed to investigate harms. For example, in 1974 three case–control studies published simultaneously suggested an association between the use of reserpine for the treatment of hypertension and the subsequent development of breast cancer.[59–61] Other studies, published later, have failed to confirm the original association,[62] which now seems to have resulted from excluding, as controls, patients with cardiovascular disease.[63] Here, a subtle form of selection bias (exclusion bias) was probably responsible for the erroneous conclusions that were originally drawn.

Hierarchies of evidence

The first hierarchy of evidence was published in the late 1970s.[64] Since then many similar hierarchies, of increasing elaboration and complexity, have appeared. A survey in 2002[65] identified 40 such grading systems and study in 2006 identified 20 more.[66]

The hierarchy in table 1, like others, places RCTs at the highest level with a lesser place for those based on observational studies. This hierarchical approach to evidence has not only been adopted by many in the evidence-based medicine and health technology assessment movements, but also it has come to dominate the development of clinical guidelines. Awarding such prominence to the results of RCTs, however, is unreasonable. As Bradford Hill, the architect of the RCT, stated so cogently,[67] "any belief that the controlled trial is the only way would mean not that the pendulum had swung too far but that it had come right off the hook".

As discussed, RCTs are particularly weak in relation to generalisability and most especially in the assessment of harms. Although RCTs can, indeed, identify those adverse effects that occur relatively commonly, and which appear within the short time-scales of their duration, there remain important limitations. Contrary to a recent claim,[68] only observational studies can offer the evidence required for assessing less common or long-latency harms.

Hierarchies cannot, moreover, accommodate evidence that relies on combining the results from RCTs and observational studies. Combining evidence derived from a range of study designs is a feature of decision-analytical modelling as well as in the emerging fields of teleoanalysis and patients' preference trials.[68–71]

Apart from their sheer number, the inconsistencies between hierarchies demonstrate their unsatisfactory nature. These inconsistencies include the variable prominence given to meta-analyses with some positioning them above large, high-quality RCTs, whereas others ignore them. There are also inconsistencies between hierarchies in their grading of observational studies: some give a higher rating to so-called cohort studies than case–control; some regard them to be all equal; and others reverse the order.

Hierarchies attempt to replace judgment with an oversimplistic, pseudoquantitative, assessment of the quality of the available evidence. Decision makers have to incorporate judgments, as part of their appraisal of the evidence in reaching their conclusion.[5] Such judgments relate to the extent to which each of the components of the evidence base is fit for purpose. Is it reliable? Is it generalisable? Do the intervention's benefits outweigh its harms? And so on.

	Adverse effect
Oral contraceptive agents (1967)	Venous thromboembolism
Diethylstilboestrol during pregnancy (1972)	Genital tract carcinoma (in young females)
Aspirin in children (1985)	Reye's syndrome
Tryptophan (1990)	Eosinophilia-myalgia syndrome
Non-steroidal anti-inflammatory drugs (1994)	Upper gastrointestinal bleeding
Hormone replacement therapy (1996)	Venous thromboembolism
Hormone replacement therapy (1997)	Breast cancer
Selective serotonin reuptake inhibitors (1999)	Upper gastrointestinal bleeding
Anticonvulsants (1999)	Stevens-Johnson syndrome and toxic epidermal necrolysis
Olanzapine (2002)	Diabetes
Fluoroquinolones (2002)	Achilles tendon disorders

***Table 4:* Some adverse effects of various drugs confirmed by case–control studies**[3]

Conclusion

Experiment, observation, and mathematics, individually and collectively, have a crucial role in providing the evidential basis for modern therapeutics. Arguments about the relative importance of each are an unnecessary distraction. Hierarchies of evidence should be replaced by accepting—indeed embracing—a diversity of approaches. This is not a plea to abandon RCTs and replace them with observational studies. Nor is it a claim that the bayesian approaches to the design and analysis of experimental and non-experimental data should supplant all other statistical methods. Rather, it is a plea to investigators to continue to develop and improve their methods; to decision makers to avoid adopting entrenched positions about the nature of evidence; and for both to accept that the interpretation of evidence requires judgment.

For those with lingering doubts about the nature of evidence itself I remind them that while Gregor Mendel (1822–84) developed the monogenic theory of inheritance on the basis of experimentation, Charles Darwin (1809–82) conceived the theory of evolution as a result of close observation, and Albert Einstein's (1879–1955) special theory of relativity was a mathematical description of certain aspects of the world around us. William Harvey's discovery of the circulation of the blood—as he described in *De Motu Cordis*—was based on an elegant synthesis of all three forms of evidence.

Acknowledgments

Several friends and colleagues have been extraordinarily generous with their time in providing information and inspiration, as well as reviewing various drafts of this oration. Any merit it may have owes much to their individual and collective contributions. They include Jeffery Aronson (University of Oxford), Deborah Ashby (Queen Mary, University of London), Patricia Beaujouan (Sanofi-Aventis), David Brickwood (Johnson and Johnson), Kalipso Chalkidou (National Institute for Health and Clinical Excellence), Iain Chalmers (James Lind Library), Stephen Evans (London School of Hygiene and Tropical Medicine), Kent Johnson (University of Newcastle, New South Wales, Australia), David Jones (University of Leicester), Jeremy Paterson (University of Newcastle upon Tyne), Stephen Pearson (Institute for Clinical and Economic Review, Massachusetts General Hospital and

Harvard Medical School), Patrick Valance (GlaxoSmithKline), Nancy Wexler (Columbia University and the Hereditary Disease Foundation, USA), Alice Wexler (University of California, Los Angeles and the Hereditary Disease Foundation, USA), and Tony Whitehead (Sanofi-Aventis). Nevertheless, I take sole responsibility for any and all errors of omission and commission.

References

1 Shapin S. The scientific revolution. Chicago and London: University of Chicago Press, 1996.
2 Gower B. Scientific method: an historical and philosophical introduction. London and New York: Routledge, 1997.
3 Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. *BMJ* 2001; **323:** 334–36.
4 Jadad AR, Enkin MW. Randomized controlled trials, 2nd edn. London: BMJ Books, 2007.
5 Rawlins MD, Culyer AJ. National Institute for Clinical Excellence and its value judgments. *BMJ* 2004; **329:** 224–27.
6 Pocock SJ. Clinical Trials. Chichester: John Wiley & Sons, 1983.
7 Altman DG, Schulz KF, Egger M, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001; **134:** 663–94.
8 Moher D, Schulz KF, Altman DG. The CONSORT Statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Ann Intern Med* 2001; **134:** 657–62.
9 Clarke M, Hopewell S, Chalmers I. Reports of clinical trials should begin and end with up-to-date systematic reviews of other relevant evidence: a status report. *J R Soc Med* 2007; **100:** 187–90.
10 Rawlins MD. De testimonio. London: Royal College of Physicians, 2008.
11 Rothman K. No adjustments are needed for multiple comparisons. *Epidemiology* 1990; **1:** 43–46.
12 Schulz KF, Grimes DA. Multiplicity in randomised trials II: subgroups and interim analyses. *Lancet* 2005; **365:** 1657–61.
13 Montori VM, Devereaux PJ, Adhikari NKJ, et al. Randomised trials stopped early for benefit: a systematic review. *JAMA* 2005; **294:** 2203–09.
14 Armitage P, Berry G, Matthews JNS. Statistical methods in medical research. 4th edn. Oxford: Blackwell, 2002.
15 Pocock SJ. When not to stop clinical trials for benefit. *JAMA* 2005; **294:** 2228–30.
16 Pocock SJ, Lubsen J. More on subgroup analyses. *N Engl J Med* 2008; **358:** 2076.
17 Wang R, Lagakos SW, Ware J, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials. *N Engl J Med* 2007; **357:** 2189–94.
18 Ashby D. Bayesian statistics in medicine: a 25 year review. *Stat Med* 2006; **25:** 3589–31.
19 Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomised trials. *J R Soc Stat* [*Ser A*] 1994; **157:** 357–416.
20 Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Bayesian methods in health technology assessment: a review. *Health Technol Assess* 2000; **4:** 38.
21 GREAT Group. Feasability, safety, and efficacy of domiciliary thrombolysis by general practitioners: Grampian region early anistreplase trial. *BMJ* 1992; **305:** 548–53.
22 Pocock SJ, Speigelhalter DJ. Grampian region early anastroplase trial. *BMJ* 1992; **305:** 1015.
23 Bland JM, Altman DG. Bayesians and frequentists. *BMJ* 1998; **317:** 1151.
24 Berry D, Goodman SN, Louis TA, Temple R. Introduction to bayesian methods: floor discussion. *Clin Trials* 2005; **2:** 301–04.
25 Food and Drug Administration. Guidance for the use of bayesian statistics in medical device trials. Bethesda: US Department of Health and Human Services, Food and Drug Administration, Centre for Devices and Radiological Health, 2006.
26 Berry DA. Introduction to bayesian methods III: use and interpretation of bayesian tools in design and analysis. *Clin Trials* 2005; **2:** 295–300.
27 Grieve AP. 25 years of Bayesian methods in the pharmaceutical industry: a personal, statistical bummel. *Pharm Stat* 2007; **6:** 261–61.
28 Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996; **312:** 1215–18.
29 Rothwell PM. External validity of randomised controlled trials: "To whom do the benefits apply?" *Lancet* 2005; **365:** 82–93.
30 Bartlett C, Doyal L, Ebrahim S, Davey P, et al. The causes and effects of socio-demographic exclusions from clinical trials. *Health Technol Assess* 2005, **9:** 38.
31 Heiat A, Gross CP, Krumhplz HM. Representation of the elderly. women, and minorities in heart failure trials. *Arch Intern Med* 2002; **162:** 1682–88.
32 McAlister FA. Applying the results of systematic reviews at the bedside. In: Egger M, Davey Smith G, Altman DG, eds. Systematic reviews in health care: meta-analysis in context. London: BMJ Books, 2001.
33 Kalra L, Yu G, Perez I, Lakhani A, Donaldson N. Prospective cohort study to determine if trial efficacy of anticoagulation for stoke prevention in atrial fibrillation translates into clinical effectiveness. *BMJ* 2000; **320:** 1236–39.
34 Evans SJW. Statistics: analysis and presentation of safety data. In: Talbot J, Waller P, eds. Stephen's detection of new adverse reactions, 5th edn. Chichester: John Wiley & Sons, 2004.
35 Teo KK, Yusuf S, Furberg CD. Effects of prophylactic antiarrhythmic drug therapy in acute myocardial infarction. *JAMA* 1993; **270:** 1589–95.
36 Beral V, Banks E, Reeves G. Evidence from randomised trials on the long-term effects of hormone replacement therapy. *Lancet* 2002; **360:** 942–44.
37 Califf RM. Clinical trials bureaucracy: unintended consequences of well-intentioned policy. *Clin Trials* 2006; **3:** 496–502.
38 Eisenstein EL, Collins R, Cracknell BS, et al. Sensible approaches for reducing clinical trial costs. *Clin Trials* 2008; **5:** 75–84.
39 Rochon PA, Gurwitz JH, Sykora K, et al. Readers guide to critical appraisal of cohort studies, 1: role and design. *BMJ* 2005; **330:** 895–97.
40 Ioannidis JPA, Haidich A-B, Pappa M, et al. Comparison of evidence of treatment effects in randomised and non-randomised studies. *JAMA* 2001; **286:** 821–30.
41 Doll R, Peto R. Randomised controlled trials and retrospective controls. *BMJ* 1980; **280:** 44.
42 MacMahon S, Collins R. Reliable assessment of the effects of treatment on mortality and major morbidity, II: observational studies. *Lancet* 2001; **357:** 455–62.
43 Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *BMJ* 2007; **334:** 349–51.
44 Normand S-LT, Sykora K, Li P, Mamdani M, Rochon PA, Anderson GM. Readers guide to critical appraisal of cohort studies, 3: analytical strategies to reduce confounding. *BMJ* 2005; **330:** 1021–23.
45 Byar DP, Schoenfield DA, Green SB, et al. Design considerations for AIDs trials. *N Engl J Med* 1990; **323:** 1343–48.
46 Hill AB. The environment and disease: association or causation? *Proc R Soc Med* 1965; **58:** 295–300.
47 Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiological evidence. *Prev Med* 1991; **20:** 47–63.
48 Grady D, Rubin SM, Pettiti DB, et al. Hormone therapy to prevent disease and prolong life in postmenopausal women. *Ann Intern Med* 1992; **117:** 1016–37.
49 Herrington DM. Hormone replacement therapy and heart disease: replacing dogma with data. *Circulation* 2003; **107:** 2–4.
50 Petitti DB, Freedman DA. How far can epidemiologists get with statistical adjustment. *Am J Epidemiol* 2005; **162:** 415–18.
51 Boston Collaborative Drug Surveillance Group. Regular aspirin intake and acute myocardial infarction. *BMJ* 1974; **1:** 440–43.
52 Smithells RW, Nevin NC, Seller MJ, et al. Further experience of neural tube supplementation for prevention of neural tube defects. *Lancet* 1983; **1:** 1027–39.
53 Gilbert R, Salanti G, Harden M, See S. Infant sleeping position and sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 to 2002. *Int J Epidemiol* 2005; **34:** 874–87.
54 Rostom A, Dubé C, Lewin G, et al. Nonsteroidal anti-inflammatory drugs and cyclooxygenase-2 inhibitors for the primary prevention of colorectal cancer: a systematic review prepared for the US Preventive Services Task Force. *Ann Intern Med* 2007; **146:** 376–89.

55 Academy of Medical Sciences. Identifying the environmental causes of diseases: how should we decide what to believe and when to take action. London: Academy of Medical Sciences, 2007.

56 Davey Smith G, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol* 2004; **33:** 30–42.

57 Sørensen HT, Christensen S, Mehnert F, et al. Use of biphosphonates among women and risk of atrial fibrillation and flutter: population based case-control study. *BMJ* 2008; **336:** 784–85.

58 Anderson HR, Ayres JG, Sturdy PM, et al. Bronchodilator treatment and deaths from asthma: case-control study. *BMJ* 2005; **330:** 117–24.

59 Boston Collaborative Drug Surveillance Program. Reserpine and breast cancer. *Lancet* 1974; **2:** 669–71.

60 Armstrong B, Skegg D, White G, Doll R. Rauwolfia derivatives and breast cancer in hypertensive women. *Lancet* 1976; **2:** 8–12.

61 Heinonen OP, Shapiro S, Tuominen L, Turunen MI. Reserpine use in relation to breast cancer. *Lancet* 1974; **2:** 675–77.

62 Grossman E, Messerli FH, Goldbourt U. Antihypertensive therapy and the risk of malignancies. *Eur Heart J* 2001; **22:** 1343–52.

63 Horwitz RI, Feinstein AR. Exclusion bias and the false relationship of reserpine and breast cancer. *Arch Intern Med* 1985; **145:** 1873–75.

64 Canadian Task Force on the Periodic Health Examination. The periodic health examination. *CMAJ* 1979; **121:** 1193–254.

65 Agency for Healthcare Research and Quality. Systems to rate the strength of scientific evidence. Washington, DC: US Department of Health and Human Services, 2002.

66 Schünemann HJ, Fretheim A, Oxman AD. Improving the use of research evidence in guideline development, 9: grading evidence and recommendations. *Health Res Policy Syst* 2006; **4:** 21.

67 Hill AB. Heberden Oration 1965b: reflections on the controlled trial. *Ann Rheum Dis* 1966; **25:** 107–13.

68 Freemantle N, Irs A. Observational evidence for determining drug safety. *BMJ* 2008; **338:** 627–28.

69 Weinstein MC, O'Brien B, Hornberger J, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPO Task Force on Good Research Practices—modeling studies. *Value Health* 2003; **6:** 9–17.

70 Wald NJ, Morris JK. Teleoanalysis: combining data from different types of study. *BMJ* 2003; **327:** 616–18.

71 King M, Nazareth I, Lampe F, et al. Impact of participant and physician intervention preferences on randomised trials. *JAMA* 2005; **293:** 1089–99.