THE HARVEIAN ORATION OF 2008

DE TESTIMONIO

On the evidence for decisions about the use of therapeutic interventions

Professor Sir Michael David Rawlins

MD FRCP FFPM FMedSci

Royal College of Physicians

Setting higher medical standards

DE TESTIMONIO

On the evidence for decisions about the use of therapeutic interventions

THE HARVEIAN ORATION

Delivered before the Fellows of
The Royal College of Physicians of London
on Thursday 16 October 2008

by

Professor Sir Michael David Rawlins
MD FRCP FFPM FMedSci

Royal College
of Physicians
Setting higher medical standards

Royal College of Physicians
11 St Andrews Place, London NW1 4LE

Registered Charity No 210508
www.rcplondon.ac.uk

ISBN 978-1-86016-347-0

Printed in Great Britain by Sarum ColourView Ltd, Salisbury, Wiltshire

William Harvey and the Harveian Trust

William Harvey was born in Folkestone on 1 April 1578. He was educated at the King's School, Canterbury, Gonville and Caius College, Cambridge, and the University of Padua, graduating as doctor of arts and medicine in 1602. He became a Fellow of the Royal College of Physicians in 1607 and was appointed to the Lumleian lectureship in 1615.

In the cycles of his Lumleian lectures over the next 13 years, Harvey developed and refined his ideas about the circulation of the blood. He published his conclusions in 1628 in *Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus*, which marks the beginning of clinical science. In it, Harvey considered the structure of the heart, arteries and veins with their valves. By carefully devised experiments and supported by the demonstration of the unidirectional flow of the blood in the superficial veins of his own forearm, he established that the blood circulated, and did not ebb and flow as had been believed for more than 1,000 years.

Harvey was a great benefactor of the College. In 1656 he gave his patrimonial estate of Burmarsh (in Romney Marsh, Kent) to the College to provide for the annual oration and feast. In an indenture dated 21 June 1656, he directed that:

> *to the end friendship between the members of the said College may be the better continued and maintained, there shall be once every month at the meeting of the Censors at the said College some small collation provided . . . and once every year there shall be a general feast kept within the said College for all the Fellows that shall please to come . . . and on the day when such feast shall be kept some one person . . . shall make an oration . . . with an exhortation to the Fellows and Members of the said College to search and study out the secret of Nature by way of experiment; and also for the honour of the profession to continue mutual love and affection amongst themselves without which neither the dignity of the College can be preserved nor yet particular men receive that benefit by their admission into the College which else they might expect.*

Professor Sir Michael Rawlins

MD FRCP FFPM FMedSci

Professor Sir Michael Rawlins

Michael Rawlins has been Chairman of the National Institute for Health and Clinical Excellence (NICE) since its inception in 1999. NICE is responsible for assisting health professionals, in the NHS, to provide patients with the highest attainable standard of care; as well as advising the public health community on measures that are effective and cost effective in the prevention of ill health. He has seen the Institute grow from an organisation with no staff, premises, or bank account and a nominal budget of £8.5 million a year, to a body now employing over 270 people, with offices in London and Manchester, and an annual budget of £35 million, set to more than double over the next few years.

From 1973 to 2006, he was the Ruth and Lionel Jacobson Professor of Clinical Pharmacology at the University of Newcastle, where he undertook research into the safety and efficacy of new and established pharmacological treatments. At the same time he was consultant physician to the Newcastle University Hospitals where he practised clinical pharmacology and general internal medicine. He is now Emeritus Professor at the University of Newcastle and Honorary Professor at the London School of Hygiene and Tropical Medicine, University of London.

He served as Chairman of the Committee on Safety of Medicines (1993–8) and of the Advisory Council on the Misuse of Drugs (1998–2008), and was appointed Knight Bachelor in 1999. He has previously delivered the Bradshaw (1986), William Withering (1994) and Samuel Gee (2006) lectures at the Royal College of Physicians.

Acknowledgements

A number of 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: Jeffrey 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, USA), 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.

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 remarked:

> *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. Robert Boyle (1627–91) and Robert Hooke (1635–1703) believed that it could only be understood by experimentation. Francis Bacon (1561–1626), René Descartes (1596–1650) and Thomas Hobbes (1588–1679) regarded observation to be the most appropriate approach. Isaac Newton (1643–1727), however, considered that only when the 'book of nature' was expressed mathematically could natural philosophers be confident in their knowledge and understanding of the world around them.[1]

Three hundred and fifty 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 this more hotly, and sometimes bitterly, argued than in the nature of the evidence that should support the use of therapeutic interventions. The relative merits of experimentation (randomised controlled trials) and observation have been, and continue to be, debated. Moreover, although the role of mathematics (as biostatistics) is almost universally recognised, the limitations of some current techniques are infrequently discussed outside the biostatistical literature.

The dispute about the evidential basis of modern therapeutics has become particularly apparent with the emergence, over the past 30 years, of what are known variously as 'rules', 'levels' or 'hierarchies' of evidence. A typical example is shown in Table 1.[3]

Such 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.

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

Table 1. A hierarchy of evidence. Reproduced with permission from the BMJ Publishing Group.[3]

Level	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 report, case studies)
4	Expert opinion

RCTs = randomised controlled trials.

patients, decide on the range of products to include in a local hospital's formulary, and may be charged with assessing whether particular interventions are sufficiently safe and effective – as well as cost effective – to be made available to an entire healthcare system. Mistakes in decision making may have dramatic repercussions at all levels.

The notion that evidence can be reliably placed in hierarchies is illusory. Hierarchies place RCTs on an undeserved pedestal for, as I discuss later, although the technique has advantages it also has significant disadvantages.[4] Observational studies too have defects but they also have merit. Decision makers need to assess and appraise all the available evidence irrespective as to whether it has been derived from RCTs 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 judgements are required about the 'fitness for purpose' of the components of the evidence base. On the contrary, judgements are an essential ingredient of most aspects of the decision-making process.[5]

Randomised controlled trials

The introduction of RCTs in the middle of the 20th century has had a profound impact on the practice of medicine and its essential features are well described.[4,6,7] It involves comparing the effects of two (or more) interventions that have been allocated randomly to groups of contemporaneously treated patients. Randomisation means that every patient in a study has a known (usually equal) chance of receiving each of the treatments. As a consequence, so-called selection bias* is minimised[8–11] with both known (and unknown) confounding factors** likely to be distributed in an unbiased manner between the groups.[12] There are claims that in the absence of randomisation, or where randomisation has been inadequate, there may be a tendency to overestimate the size of the effects of treatments.[6,13,14]

Random allocation also allows the application of the underlying theory of random sampling with the differences between treatment groups behaving like the differences between random samples from a single population.[9,15] This allows treatments to be compared as though they were equally effective.[9] It is also partly for statistical reasons that the analysis of RCTs is properly based on 'intention to treat', rather than confined only to those who have successfully completed the study (so-called 'per protocol'). Intention to treat analyses incorporate all randomised patients, irrespective of whether they have completed the trial.

In addition to random allocation and intention to treat analysis, many trials also attempt to 'blind' (or 'mask') investigators and patients as to the treatment that is given. Blinding has two important consequences. Firstly, it minimises selection bias due to the conscious or unconscious preferential allocation of treatments. Secondly, it minimises bias due to a systematic distortion in the outcome measurement(s) in the groups (known as ascertainment bias). Ascertainment bias is likely to be particularly significant where outcomes are assessed subjectively by investigators or patients. Blinding, however, may be less important if the outcomes are objective. And in some circumstances, such as in many surgical trials, it can be virtually impossible.

Double blind RCTs, when properly conducted and analysed, unquestionably provide confidence in the internal validity of the results,[6,11,16] and the more so if replicated by subsequent studies. Thus, both the direction and magnitude of the observed effect, under the particular circumstances of the study, are likely to be reasonably reliable. Consequently, RCTs are often called the 'gold standard' for

**Bias* is a systematic distortion of the estimated intervention effect away from 'the truth' caused by inadequacies in the design, conduct or analysis of the trial.[8] *Selection bias* is a systematic error in creating intervention groups causing them to differ with respect to prognosis. The groups differ in measured or unmeasured baseline characteristics because of the way participants were selected for the study or assigned to their groups.[8]

***Confounding* is a situation in which the estimated effect of an intervention effect is biased because of some difference, apart from the planned intervention, between the groups (eg prognostic factors, concomitant interventions). For a factor to be a confounder it must differ from the comparison groups *and* predict the outcome of interest.

demonstrating (or refuting) the benefits of a particular intervention. Yet the technique has important limitations and imperfections. These include:

- inappropriateness
- utility of the null hypothesis
- theories of probability
- generalisability of the results
- resource implications.

Inappropriateness

There are circumstances when it may be inappropriate either to conceive, or undertake, RCTs. RCTs may be impossible for bioethical or legal reasons. This has been extensively discussed elsewhere[17] and is not considered further. RCTs may be virtually impossible to conduct in the evaluation of treatments for very rare diseases where the numbers of available patients are circumscribed.[18] It took nine years, for example, to enrol 39 patients into an RCT designed to assess the benefit of itraconazole in preventing serious fungal infections in patients with chronic granulomatous disease.[19] Alternative approaches, based on observational methods, are needed if those with rare diseases are not to be denied safe and effective treatments.[18,20–22]

RCTs may be unnecessary when interventions produce very substantial effects. The potential for bias to yield unreliable results of the effects of a treatment is greatest when the intervention produces only a moderate improvement.[23,24] While bias is unlikely to give rise to a ten-fold artifactual difference in a disease outcome, between the treatment groups, it could easily give rise to two-fold differences.[23] Put another way, where the effects of a treatment are large and 'dramatic', conclusions about effectiveness may be obvious.[24,25] This is discussed later in greater detail.

The null hypothesis

The analysis of an RCT 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 as, or more extreme than, the one observed, were the null hypothesis to be true. If the probability is less than some arbitrary value – usually less that 1 in 20 (ie $p<0.05$) – then the null hypothesis is rejected.

This is the 'frequentist' approach to the design and analysis of RCTs and has undoubted attractions: the statistical calculations are relatively 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 demonstrating that a particular treatment has some effect. This 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 is 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.[26] The design, conduct and analyses of RCTs should invariably be based on a full and systematic review of the published and unpublished evidence.[27]

The null hypothesis is even more awkward for trials seeking to show whether there is no difference, or not much of a difference, between treatment groups. Attempts to resolve this include equivalence trials, non-superiority studies and (the ineptly named) futility designs (see Table 2). All require prior assumptions to be made about the extent to which the differences between treatments might be relevant or important.[28–30] None, however, really resolve the underlying difficulty of defining when it is reasonable to accept the null hypothesis. The null hypothesis may, indeed, be methodologically consistent with the deductive approach to science but as Rothman observed: 'To entertain the universal null hypothesis is, in effect, to suspend belief in the real world and thereby to question the premises of empiricism'.[31]

Table 2. Clinical trial designs.

Superiority trial
A study designed to detect a difference between treatments. The usual test of statistical significance evaluates whether the results of the trial are consistent with the assumption of there being no difference (ie the null hypothesis).

Equivalence trial
A study designed to confirm the absence of a pre-specified difference between treatments. The margin of a clinically significant difference is chosen by defining the largest difference that is clinically acceptable.

Non-inferiority trial
A study designed to show that one treatment is no worse than another. It may be either as effective, or more effective, than the comparator.

Futility trial
A study designed to show a pre-specified difference between treatments.

Probability

The p value. As already described, the p value is the probability of observing a result as extreme, or even more extreme, than the one observed, given that the null hypothesis is true. Put another way, if the p value is sufficiently small, either the null hypothesis is false or a very rare event has occurred.[32]

By convention, a probability of less than 5% (ie $p<0.05$) is used to distinguish between 'extreme' and 'non-extreme'. However, a p value of either greater or less than 0.05, as estimated by a frequentist analysis of the results of an RCT, neither disproves or proves (respectively) the null hypothesis. It is often erroneously assumed[33–35] that when $p<0.05$ there is a 95%, or greater, chance that the null hypothesis is incorrect. The p value is calculated on the assumption that the null hypothesis is true, so it cannot be a direct measure of the probability that it is false.[34]

Rejection of the null hypothesis – because the p value is less than 0.05 – does not necessarily mean that the particular treatment is superior to the comparator. Aside from the possibility that the difference is due to chance (random error), very large studies may show small differences – with 'significant' p values – that are clinically inconsequential. Nor does it follow, just because $p>0.05$, that the treatment is ineffective; too small a sample size, for example, can readily produce a false negative conclusion.

Some, though not all, of the problems associated with p values can be avoided by expressing results as confidence intervals. The confidence interval describes the degree of uncertainty, or lack of precision, of the estimate of interest.[36] Nevertheless, p values and confidence interval are closely related.[37]

The philosophy of probability is much debated by statisticians.[38–41] The frequentist approach is based on the work of Ronald Fisher (1890–1962) together with the combined contributions of Jerzy Neyman (1894–1981) and Egon Pearson (1895–1980).* The main alternative – the Bayesian approach – is named after Thomas Bayes (1702–61) who was a non-conformist minister in Tunbridge Wells. His paper on probability, entitled 'An essay towards solving a problem in the doctrine of chance', was published posthumously.

The frequentist approach is concerned with the probability of some data conditional on a specific hypothesis (usually the null hypothesis). The Bayesian

*This is an oversimplification. Fisher was responsible for devising the null hypothesis and its rejection on the basis of the p value. Neyman and Pearson introduced the concepts of type I errors (incorrectly rejecting the null hypothesis) and type II errors (incorrectly accepting the null hypothesis). They also defined an 'alternative hypothesis' when the null hypothesis is rejected on the basis of the p value. The differences in the Fisherian and Neyman–Pearson approaches were more due to personality conflicts than immutable disagreements about the philosophy of probability.[42]

approach to probability – known as subjective or inverse probability – is the likelihood of a hypothesis given specific data.[42] The use of a Bayesian approach to the design and analysis of RCTs is discussed later.

Multiplicity. The difficulties in interpreting frequentist p values become even more convoluted when seeking to decide, during a clinical trial, whether a study should be terminated prematurely. It is equally unsatisfactory in the assessment of the outcome(s) in subgroups of patients once the trial has been completed. A similar problem, which is discussed later, applies to the safety analysis of RCTs.

In all of these instances, repeated tests of statistical significance – adopting the conventional p value (<0.05) – are increasingly likely to produce one or more falsely 'significant' results. This is known as the problem of 'multiplicity'. If d is the number of independent comparisons, the chances that at least one will be found to be significant (where the p value is set at 0.05) is:

$$1 - (1 - 0.05)^{d}$$

Accordingly, if 10 separate assumptions are tested, the probability of one being apparently significant, using p<0.05 as the test for 'significance', is 40%. There are, though, very divergent views among statisticians as to how to deal with this difficulty both in devising stopping rules and in subgroup analyses.[43]

Stopping rules. 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 the trial should continue or be prematurely stopped. Premature termination may be justified on the grounds that the study has already achieved its endpoint(s) for showing benefit, or because of safety concerns in one of the treated groups. Because such interim analyses would necessarily unblind the trial investigators, it is now common practice for this to be undertaken by an independent data monitoring committee which reports its findings and advice to the investigators and sponsors.[44]

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').[45] When this happens, data from future trials will yield a more conservative estimate of the treatment effect with 'regression to the mean'.* Moreover, with repeated examination of the emerging data there is an increasing likelihood of rejecting the null hypothesis

***Regression to the mean* is an empirical phenomenon in which extreme values tend to be followed by more normal ones. The term was originally coined by Francis Galton (1822–1911) after his discovery that the children of tall parents tended to be taller than average but were nevertheless generally shorter than their parents.

at $p<0.05$. In a large trial, a p value of less than 0.05 could be almost guaranteed if the data were analysed often enough.

Various statistical approaches have been developed to resolve this form of multiplicity.[6,7,15,16] Many methods depend on changing the level of statistical significance (ie the p value), as each interim analysis is performed, 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, any more than there is about handling multiplicity in general.[15,43] Some consider that data monitoring committees should adopt formal stopping rules at the start of the study, but allow for flexible implementation. Others favour an open approach that does not formally define stopping rules, but which recognises the effect of repeated inspections of the data.[15]

A resolution to the problem of the early termination of RCTs has become urgent. In a systematic review of RCTs stopped early for benefit, it was noted that not only had this phenomenon become more common, but that the trial reports often failed to provide adequate information about the reasons for stopping.[45] Moreover, many had been prematurely terminated with implausibly large treatment effects ('random highs') and small numbers of events. An even more recent review of prematurely terminated trials in oncology concluded that although studies in this area were better designed than in the past, they are now too often stopped prematurely.[46] The implications of false positive results are of special concern in oncology since many newer (and very expensive) agents provide, at best, only modest benefits. As stopping trials early, for benefit, may systematically overestimate treatment effects, there is a real danger that claims for benefit may be (inadvertently) unwarranted.[43,47]

Analysis of subgroups. Analyses of the effects of an intervention, in subgroups of patients, can be important in order to establish whether different types of people respond differently.[48] This, formally, requires a test of treatment by subgroup interaction. The results may demonstrate whether treatment benefits are confined to certain categories of patients,[49,50] or whether they are more cost effective in some compared to others.[51]

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 pre-specified during the planning stage.[7,43,50,52] 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 suggest cautious statistical adjustment of the p value;[7,50] others consider that no adjustment is needed.[31]

The Bayesian approach. A growing number of statisticians 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.[53]

In its more simple expression, Bayes' theorem relates the probabilities from what is known before (*a priori*) an experiment, such as an RCT, to the probabilities re-calculated 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 performed.

Bayesian statistics are widely used outside medicine. Spam filters in email systems, for example, rely on estimates of Bayesian probabilities to distinguish genuine from unwanted messages. Bookmakers have been instinctive Bayesians for generations: a horse's form book corresponds to the prior; the result of its last outing on the race course is the experiment; and the posterior odds, for today's race, are the resultant. And bookmakers, of course, expect to average a 10% return on their turnover from a day at the races. Bayesian analyses are often depicted, graphically, as probability distributions and an example of this, based on the results of the ISIS-2 trial, is shown in Fig 1.[54]

In the ISIS-2 trial thrombolytic therapy (intravenous streptokinase), given early to patients with acute myocardial infarctions (MIs), reduced the 35-day mortality by 25% (95% CI 18% to 35%) compared to placebo. The curve in Fig 1 shows the

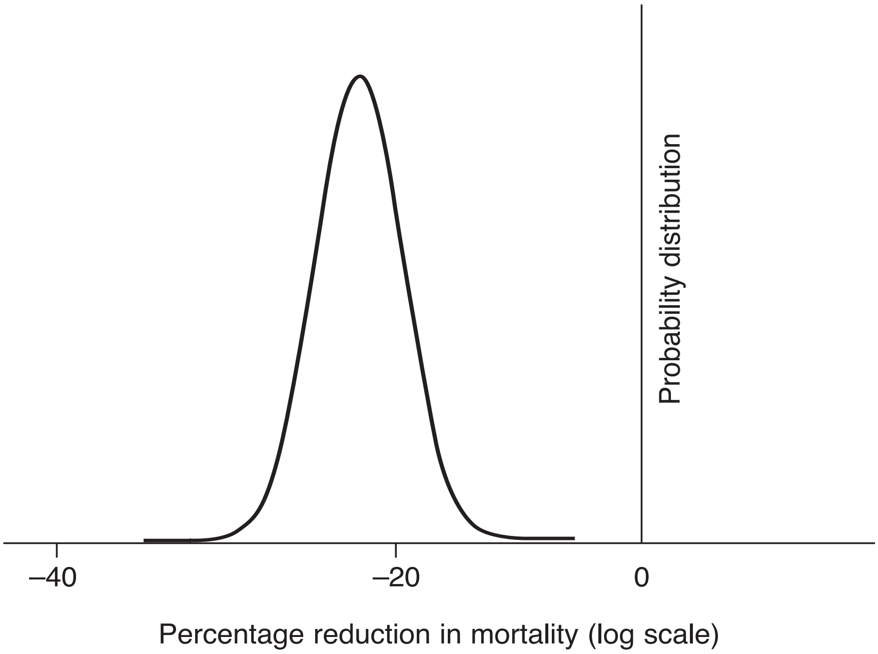

Fig 1. Probability distribution of the percentage reduction in mortality from vascular death following treatment with thrombolysis after acute myocardial infarction. Data from the ISIS-2 trial.[54]

probabilities for all values of the reduction in mortality (compared to placebo). Because it covers the whole of the probability distribution, its area is 1 with a maximum probability corresponding to a 35-day mortality reduction of 25%. There is no p value and a decision maker examines the curve and makes a judgement about what conclusion is reasonable in the light of what it shows.

An application of a Bayesian approach to the analysis of an RCT is shown in Fig 2. The GREAT trial was designed to test the hypothesis that early domiciliary thrombolytic therapy for acute MI would be better than later treatment once patients had reached hospital.[55] The investigators therefore undertook an RCT comparing the effectiveness of thrombolysis given by general practitioners (GPs) in the patients' own homes with later treatment once they had arrived at their local hospital. The investigators observed a relative reduction in all cause mortality of 49% (p=0.04) for patients treated at home, compared to 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%.[56]

Pocock and Spiegelhalter therefore undertook a Bayesian re-analysis (Fig 2).[56] They derived a prior distribution (Fig 2a), based on the results of previous RCTs of hospital treatment with thrombolytics, but expressing their belief that a 15 to 20% reduction in mortality was highly plausible but that extremes of no benefit and a 40% reduction were unlikely. Fig 2b shows the probability distribution of the results of the GREAT trial. Fig 2c represents the posterior distribution obtained by multiplying the prior and the likelihood.

Two points are worth making about Fig 2. First, the probability distribution of the results of the GREAT trial (Fig 2b) is very 'flat'. This provides a simple graphical representation of the degree of uncertainty around the magnitude of the claimed benefit of domiciliary thrombolysis. Secondly, the posterior distribution in Fig 2c shows that domiciliary thrombolytic therapy, for acute MI, is most likely to produce approximately a 24% risk reduction (compared to hospital treatment) with a 95% credible interval* of 0% to 43%. The probability derived from the original analysis of the GREAT study has therefore been 'pulled back' to provide a formal representation of the belief that the original results were 'too good to be true'.[56]

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. These include power calculations and issues relating to both multiplicity in investigating subgroups and interim analyses.[40,57,58–61]

*The *credible interval* estimates the boundaries of the probability distribution.

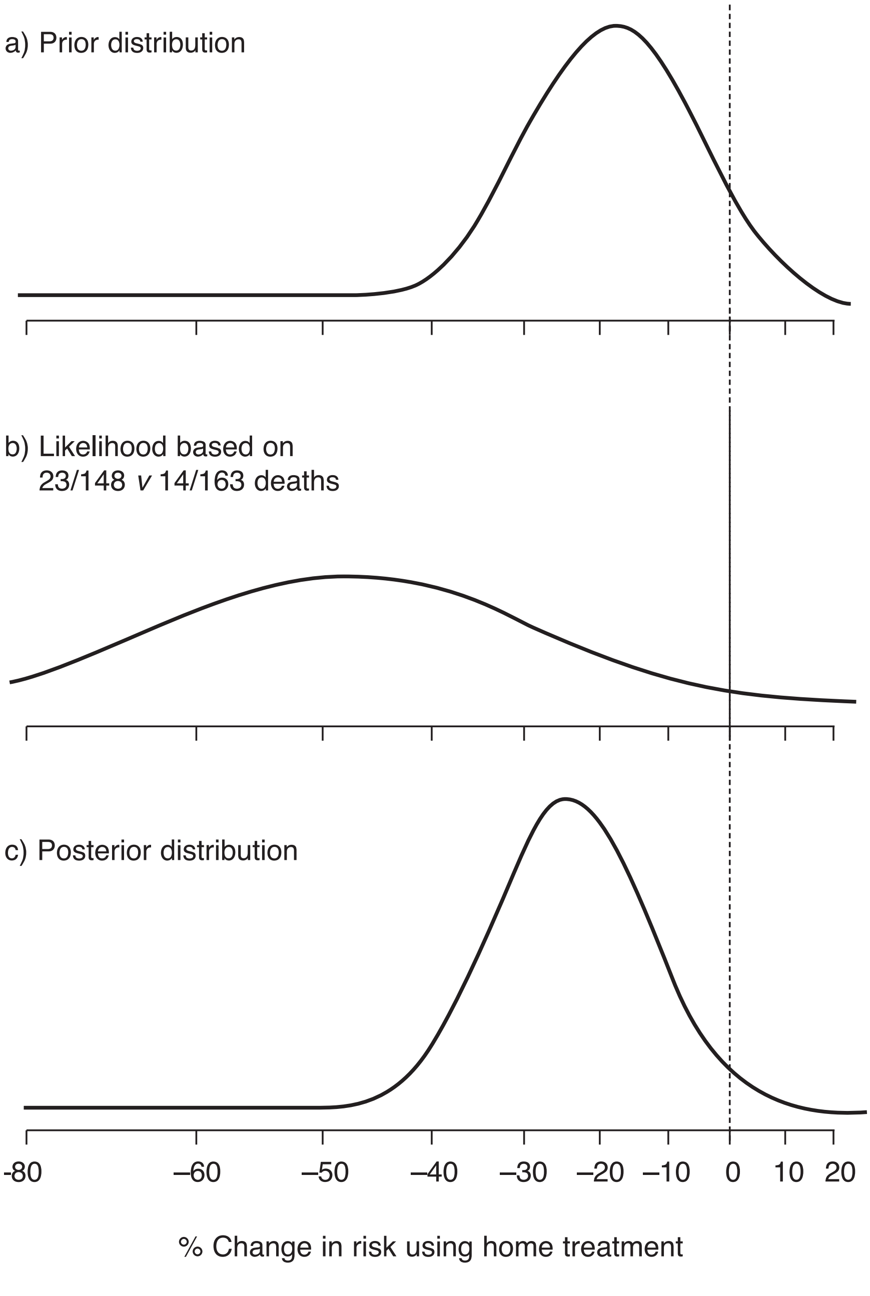

Fig 2. Bayesian re-analysis of the GREAT trial showing change (reduction) in mortality from home thrombolytic therapy compared with treatment in hospital. a) Prior probability distribution of home treatment; b) likelihood probability distribution from the GREAT trial; c) posterior probability distribution of home treatment using Bayes' theorem. Reproduced with permission from the BMJ Publishing Group.[55]

Why then are Bayesian methods not more commonly used in biostatistics? There appear to be five main reasons.

Firstly, although the subjective approach to probability dates back to the 18th century,[40] some (especially those of a frequentist mindset) regard this interpretation of probability – as a personal belief or judgement – 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 either personal belief or judgement should come into play in decision making. Perhaps, surprisingly, given that judgement plays such an important role in the practice of clinical medicine, it is my strong impression that clinical investigators are much more reluctant to accept a subjective approach to the interpretation of probability than statisticians.

Secondly, there have been substantial controversies about the derivation of the prior probability. Where there is evidence from one or more phase II studies during a drug development programme, or from the results of previously published studies (both experimental and observational), a so-called clinical prior is readily available.[59,62] For example, a clinical prior, adjusted to take account of the potential for improvement in survival as a result of early (domiciliary) thrombolysis, was used in the re-analysis of the GREAT study discussed earlier. Difficulties arise, however, when there is no clinical prior. Use then has to be made of 'default' priors.[40] I believe that too much has been made of the alleged difficulties in using these. Bayesians tend to use a number of default priors in the absence, and even, sometimes, in the presence, of clinical priors as part of their sensitivity analyses.

Thirdly, Bayesian analyses are computationally complex. Although my statistical colleagues claim that they are now relatively simple, they are nevertheless much more demanding than the methods used in most frequentist analyses.

Fourthly, some statisticians – albeit a dwindling number – are unfamiliar with the techniques of Bayesian analysis and are unwilling (or unable) to adapt. Some generously attribute this variation in skill-mix to a statistician's original choice of university.[63] 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, and very importantly, regulatory authorities have sometimes been hesitant to concede that Bayesian approaches may have advantages.[64] Pharmaceutical companies have therefore, understandably, been reluctant to submit license applications with the results of their trials based on Bayesian analyses. There are, though, signs of rising interest particularly in the evaluation of devices.[65] And manufacturers, themselves, are increasingly adopting Bayesian approaches in phase II and III trials.[57,66–68] Perhaps as a result of these initiatives, there are now documented instances where Bayesian statistics have been successfully deployed in regulatory submissions.[57,67,68]

Bayesian approaches to the design and analysis of RCTs are likely to play a much greater role in the future[53] and perhaps in combination with a frequentist approach to estimating priors.[69] Eliminating the rigidity of the p value, and answering 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 undertaken in selected patient populations 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 co-morbid 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 patient populations has become an increasingly important issue.[11,70] The most significant problems are outlined in Table 3.

That there is a real concern over the issue of generalisability is undoubted.[11,71–76] There have been few systematic attempts, however, to assess its extent or

Table 3. Adverse influences on the generalisability of the results of randomised controlled trials (RCTs).

Factors	Issues	Potential problems
Patients	Age	Effectiveness in younger or older patients
	Gender	Effectiveness generally
	Severity of the disease	Effectiveness in milder or severer forms of the condition
	Risk factors	Effectiveness in patients with risk factors for the condition (eg smokers)
	Co-morbidities	Influence of other conditions on effectiveness
	Ethnicity	Effectiveness in other ethnic groups
	Socio-economic 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
	Co-medication	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

significance.[70] Bartlett and colleagues[74] 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-representations have been observed in RCTs of other cardiovascular interventions.[77] A comparison between the results of the major RCT demonstrating the benefits of drotrecogin alfa, in the treatment of sepsis in intensive care units and use in UK clinical practice, again showed striking differences in the treated populations.[78]

Assessment of benefit. There is 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? There is a presumption, by some, that the results of RCTs in discrete patient populations can, other things being equal, be reliably extrapolated to the care of patients in general.[8,79] It is argued that, if the pathogenesis of a disease is the same in all subgroups, similar benefits can be expected in wider patient populations.

The problem with this claim is that there is little systematic evidence to support it and some that refutes it.[70,79] There are, unquestionably, individual studies demonstrating concordance between the beneficial effects seen in RCTs and those observed during conventional medical care. The benefits of anticoagulation in patients with non-valvular atrial fibrillation (AF), as discussed later, are a case in point.[80] And a 2008 systematic review[81] claimed to show that the benefits observed in RCTs were similar to those receiving the same treatments outside such trials. But the extent to which the differing characteristics of patients treated in RCTs, compared to those undergoing routine clinical care, really matters – in relation to the claimed benefits – remains uncertain. Although the CONSORT statement,* on the reporting of RCTs, indicates the importance of considering the generalisability of the results, it provides little assistance as to how this might best be done.[16] Indeed, as the CONSORT group themselves admit, 'External validity is a matter of judgement'.[8]

So-called pragmatic RCTs, in which exclusion criteria are kept to a minimum, might provide grounds for adducing an overall effect size in an 'average' general population.[82] Such studies would be very unlikely, though, to provide reasonably robust estimates of effectiveness among the range of potential issues in Table 3 leaving aside the problems

*The *CONSORT statement* is a checklist and flow diagram for reporting the results of RCTs.[16] It was drawn up by a group of statisticians and clinical trialists and its reporting requirements have been supported by many healthcare journals.

of multiplicity. Nor is there any prospect of mounting individual RCTs to study, specifically, each of the possible factors outlined in Table 3.

Informed commentators suggest that better reporting of the characteristics of trial participants, coupled with the much greater use of databases and registries, would enable outcomes during routine use to be better evaluated.[11,78] In the meantime, decision makers must continue to use their scientific and clinical judgement since it appears that little progress has been made over the past 35 years. As Archie Cochrane pointed out in 1971: 'Between measurements based on RCTs and benefit in the community there is a gulf which has been much under-estimated'.[83]

Assessment of harms. 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 reliable evidence. A survey published in 1998 revealed that, of the new active substances licensed as medicines in the UK between 1972 and 1994, only one had been withdrawn for lack of efficacy, but 22 for safety reasons.[84] RCTs may, as discussed later, detect 'dramatic' harms, but they are an unreliable approach to the definitive identification of harms.

In RCTs it is now customary to collect and record all the adverse events* occurring after randomisation. This reduces 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 demonstrate clinical efficacy. Power calculations do not, however, usually take harms into account.[86] As a consequence, although RCTs can identify the more common adverse reactions, they singularly fail to recognise less common ones. As a rule of thumb,[87] the number of patients exposed to a drug must be three times the reciprocal the incidence of a particular adverse event to have a 95% chance of seeing it just once. In other words, if an adverse event associated with a particular intervention has an incidence of 1% then 300 patients must be studied to have a 95% chance of observing it just once. Even then, this may not be evidence that it is occurring more

*An *adverse event* is any unfavourable outcome occurring during or after exposure to an intervention but is not necessarily caused by it.[85]

frequently than in the comparison group, unless there is convincing data to show that the expected number (in the comparison group) was close to zero.

With large RCTs, including, for example, 3,000 participants, there is a reasonable prospect that adverse effects occurring at a rate of 1:100 will be recognised. Serious adverse effects, occurring at a rate of 1:1,000, are likely to be unrecognised. The lack of power of most RCTs to detect less common adverse effects is compounded where these have a long latency (such as malignancies). Most RCTs, even for interventions that are likely to be used by patients for many years, are of only six to 24 months duration.

The analysis of RCTs, for harms, poses yet another unresolved problem of multiplicity.[86,87] In large-scale, long-term studies it will be almost inevitable that some statistically significant adverse events will 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 relationship might be inferred. Similarly, if the adverse events are biologically plausible (eg breast cancer with hormone replacement therapy (HRT)), a causal relationship might also be inferred. Where these factors do not apply, difficulties in interpretation may arise.

The results of a large long-term placebo controlled RCT of the effects of pravastatin after acute MI illustrate this problem.[88] A greater incidence (p=0.002) of breast cancer was noted in women treated with pravastatin compared to those on placebo. A later systematic review, combining the data from seven large RCTs and nine observational studies, failed to show an increased risk of breast cancer after statin use. The original findings can now be reasonably attributed to random error.[89]

Properly conducted and analysed RCTs can occasionally provide important information about adverse effects. Examples include RCTs of prophylactic antiarrhythmic therapy, with class 1 agents, after MI,[90] and of HRT in postmenopausal women.[91] The former showed an increased mortality in those randomised to active treatment with antiarrhythmic therapy; an overview of the latter showed a causal association with breast cancer and stroke. These, though, are exceptions. Although better reporting of data about harms would unquestionably enhance the contributions of RCTs to the assessment of harms, the underlying problems remain.[92]

The role of RCTs, in the detection of adverse effects, reached a nadir with rofecoxib (a cyclooxygenase-2 (COX-2) inhibitor). A randomised comparison of the gastro-intestinal toxicity of rofecoxib and naproxen (the VIGOR trial) revealed a statistically significant excess of MIs in patients treated with rofecoxib.[93] The authors attributed this to a protective effect of naproxen, as a result of inhibition of platelet aggregation, rather than an adverse effect of rofecoxib. Although this was one possible biologically plausible explanation, there was another that the authors failed to discuss. COX-2 inhibitors not only lack an effect on platelet aggregation but also inhibit the production of prostacyclin.[94] This effect could also, equally plausibly, have accounted for the increased risk of adverse cardiovascular events with rofecoxib. A later RCT using placebo as a comparator unequivocally confirmed the association between rofecoxib and an increased risk of adverse vascular events.[95]

The failure, in the case of rofecoxib, to use biological plausibility appropriately was compounded by a lack of disclosure (a form of publication bias). It transpired, at US Food and Drug Administration hearings in 2001, that the sponsor (Merck) had been aware of potential myocardial toxicity before the VIGOR trial and had established an internal committee to reselect cases in the VIGOR study for adjudication.[96] It now appears that the cardiovascular risk described in the VIGOR study was probably an underestimate.[97] It is also now known that Merck had been aware, in 2003, of an excess cardiovascular risk with rofecoxib from the results of trials in Alzheimer's disease, which also revealed an excess of deaths due to cardiac disease.[98]

The sorry tale of rofecoxib is not an isolated example of the failure to place data about harms, from RCTs, in the public domain. In a survey of 192 RCTs, covering seven therapeutic domains, less than half (46%) provided information on the frequency of specific reasons for safety withdrawals.[99] In this same survey, the severity of clinical adverse effects and laboratory abnormalities were defined in only 39% and 29% of trials (respectively).

Of equal concern must be the RCTs which are prematurely terminated for safety reasons, or which show unacceptable toxicity on completion, but which remain unpublished because they never form the basis of an application to a drug regulatory authority. This latter difficulty was highlighted by a comparison of the published and unpublished data on the safety and efficacy of serotonin reuptake inhibitors (SSRIs) in the treatment of childhood depression.[100]

As part of the preparation for a National Institute for Health and Clinical Excellence (NICE) clinical guideline on the management of childhood depression, Whittington and colleagues undertook a systematic review of published RCTs of the efficacy and safety of SSRIs in children and young people.[100] Though unlicensed in this age group they were aware that they were often used 'off label'. They therefore contacted the manufacturers of all SSRIs asking for details of any unpublished trials. None responded. By a more than happy coincidence, the UK's Committee on Safety of Medicines had just completed its own review of SSRIs in children and young people, and had placed details of unpublished studies on its website. Using this additional information, the guideline developers re-assessed the risk:benefit of each of the available SSRIs using published, as well as unpublished, data. Their conclusions changed substantially and only fluoxetine was recommended for use.[100]

If the committee had not disclosed details of these unpublished studies, it is likely that NICE's clinical guideline would have recommended the use of treatments that were at best ineffective and at worst dangerous.

There remain problems of both probity and science in the assessment of safety from RCTs. Although some of the difficulties associated with multiplicity might be resolved by the greater use of Bayesian techniques, these will obviously be unable to address problems of concealment.[53]

Resources

The costs of RCTs are substantial in money, time and energy. Fig 3 shows the range of costs of 153 RCTs that were completed in 2005 and 2006. This data combines the costs of trials that were funded by the National Institute for Health Research and the Medical Research Council (MRC) as well as those incurred by three major pharmaceutical companies in their phase 2 and 3 studies. The median cost was £3,202,000 with an interquartile range of £1,929,000 to £6,568,000.

I do not claim that these data are either comprehensive or necessarily representative of RCTs generally, but they demonstrate that trials can be very expensive undertakings. The average cost per patient also appears to be rising. One manufacturer has estimated that the average cost per patient, included in trials, has increased from £6,300 (in 2005), to £7,300 (in 2006), to £9,900 (in 2007).

Much of the rise in costs, over the past few years, has been due to the increasing regulatory (and other) requirements imposed on privately and publicly funded trials.[101] Each measure was introduced with the best of intentions. These included

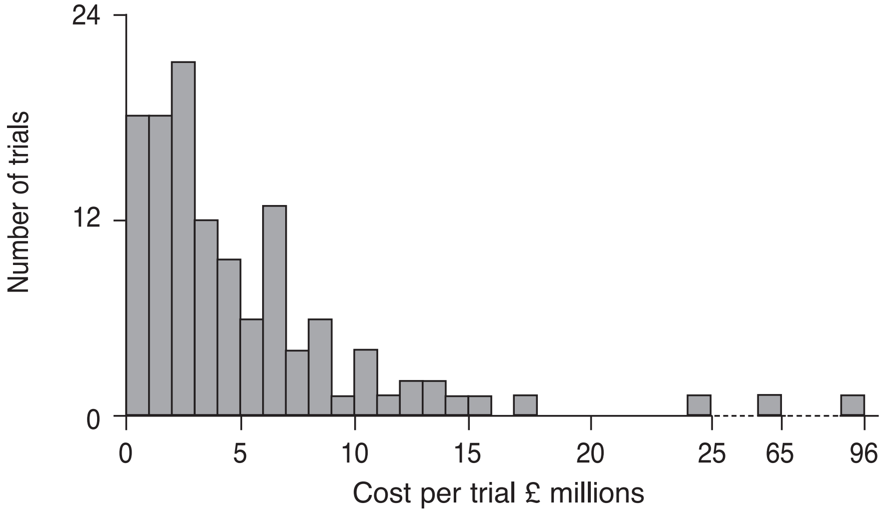

Fig 3. Range of study costs of individual randomised controlled trials (pharmaceuticals).

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. Consequently, 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 the NHS or in the private sector.

Recent proposals by an international group of academic clinical investigators indicate that clinical trial costs could be decreased by between 40% and 60% without detriment to their quality.[102,103] 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.

RCTs also impose a substantial carbon footprint. With commendable honesty, the investigators of the CRASH trial* undertook a carbon audit of their own study.[104,105] During a one-year audit period, the total emission of greenhouse gases amounted to 126 tonnes (carbon dioxide equivalents). If the audit year was representative of the six years of the study, the trial was responsible for about 630 tonnes of carbon dioxide equivalents (corresponding to 525 round-trip flights from London to New York for one passenger). The authors concluded that simplified trial design, reduced bureaucracy and more videoconferencing would reap substantial savings.[105] There is a striking concordance between the measures that would reduce the burdens that RCTs place on healthcare systems and the planet.

Observational studies

The nomenclature describing observational (non-randomised) studies is confused. I eschew a distinction between 'controlled' and 'uncontrolled' studies because all observational studies involve implicit (informal) or explicit (formal) comparisons. Nor do I consider the terms 'cohort studies' or 'quasi-experimental studies' particularly illuminating. The former combines study designs that are, in reality, distinct entities; the latter is a term that I have never found to be adequately or consistently defined.

There are, I believe, five distinct varieties of observational studies that have been, and continue to be, used in deriving evidence about the benefits and harms of therapeutic interventions:

***CRASH* was a large, multicentre RCT designed to investigate whether the administration of corticosteroids, in the immediate aftermath of traumatic brain injury, would improve recovery. The trial showed that corticosteroids had no beneficial effect and possibly a detrimental one on survival.[104]

- historical controlled trials
- non-randomised, contemporaneous controlled trials
- case-control studies
- before-and-after designs
- case series and case reports.

There is extensive, and sometimes disputatious, literature comparing the merits and demerits of randomised and observational studies of the effectiveness of therapeutic interventions.[13,106–120] Attempts at systematic reviews of published comparisons between the two approaches, however, have been bedevilled by two particular problems. The first 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 not inconsiderable. The second is the difficulty that very few of these reviews have distinguished between the various types of observational designs.

There is general agreement that 'dramatic' effects can be discerned without the need for RCTs.[10,23–25,113] There is, though, much less of a consensus about the role of observational studies in defining benefit when the effect size is more modest.[108,113,120,121] There may, indeed, be a tendency for observational studies to provide larger treatment effects than RCTs. This has not, though, been an invariable finding. Indeed, in some instances, underestimates as well as overestimates have been reported. The magnitude of differences between RCTs and observational data may also vary with the specific type of design used in the observational studies.[112]

The greatest strength of an RCT is that the allocation of treatments is random so that the groups being compared are similar for baseline factors. This may not be the case in observational trials where there is a real danger of selection bias and confounding.[121] Various approaches are available for assessing and, if necessary, adjusting the results to take account of these.[122,123]

Historical controlled trials

In a historical controlled trial, a group of patients are treated with the intervention under investigation and compared, retrospectively, with a group who had previously received a standard therapy (including best supportive care). The technique has been used, primarily, to assess an intervention's benefits rather than its harms. Examples of interventions of unquestioned efficacy, as demonstrated by historical controlled trials, are shown in Tables 4 and 5. The controls in these trials may be either implicit or explicit.

Implicit controls represent established knowledge of the natural history of a particular condition. For example, it had been known for many years that untreated

Table 4. Pharmaceutical interventions with established effectiveness based on historical controlled trials.

Intervention (year)	Indication
Thyroxine (1891)	Myxoedema[1]
Insulin (1922)	Diabetic ketoacidosis[2]
Vitamin B_{12} (1926)	Pernicious anaemia[3]
Physostigmine (1934)	Myasthenia gravis[4]
Sulfonamides (1937)	Puerperal sepsis[5]
Penicillin (1941)	Lobar pneumonia[6]
Streptomycin (1948)	Tuberculous meningitis[7]
Ganglion blockers (1959)	Malignant hypertension[8]
Cisplatin plus vinblastine and bleomycin (1977)	Disseminated testicular cancer[9]
N-acetylcysteine (1979)	Paracetamol[10]
Ganciclovir (1986)	CMV retinitis[11]
Imiglucerase (1990)	Gaucher's disease[12]
Imatinib (2002)	Chronic myeloid leukaemia[13]
Imatinib (2005)	Gastrointestinal stromal tumours[14]

Table 5. Procedures with established effectiveness based on historical controlled trials.

Procedure (year)	Indication
Tracheostomy (1546)	Tracheal obstruction[15]
Blood transfusion (1818)	Haemorrhagic shock[16]
Defibrillation (1948)	Ventricular fibrillation[17]
Heimlich manoeuvre (1975)	Laryngeal obstruction by a foreign body[18]
Fundoplication (1993)	Gastrooesphageal reflux[19]
Parental kiss (2000)	Nasal obstruction by a foreign body[20]
Laser treatment (2000)	Removal of port wine stains[21]

myxoedema, pernicious anaemia and Addison's disease all led to inexorable death. The introduction of replacement therapy with thyroxine, vitamin B12 and cortisone (respectively) resulted in complete cures. No RCT was necessary and confidence in the benefits of these treatments is as absolute now as it was when they were first introduced.

Explicit controls are specific groups of patients in whom the progress of the condition has previously been observed and documented. In some instances their case records are retrieved specifically for the purpose of constructing a comparator group. Examples include studies of the effectiveness of N-acetylcysteine for the treatment of paracetamol poisoning, of ganciclovir to treat AIDS-related cytomegalovirus retinitis, or of imatinib for the treatment of chronic myeloid leukaemia (Table 4).

In the past, the use of historical controls has been the subject of considerable criticism.[6] It has been suggested that explicit historical controls are less likely to have clearly defined selection criteria; that they may differ, in some way (such as severity), from those treated with the new intervention; or that there may have been deliberate or subconscious restrictions on the choice of patients selected for the old or new treatments. Moreover, the retrieval of clinical data for the historical group may be incomplete; the response criteria may have changed; and the ancillary care, for patients with the condition, may have also changed since the time the cohort was originally assembled.

One of the most controversial, and inappropriate, uses of historical data was the use of high dose chemotherapy supported by autologous bone marrow transplantation in people with advanced breast cancer.[124] This therapeutic regimen was introduced, during the 1980s, on the basis of claims of greater responsiveness when compared with that in an explicit historical control population treated with conventional dose chemotherapy.[125,126] Yet, despite warnings about the numerous biases in such historical comparisons, and the substantially greater treatment-related mortality in women receiving high-dose chemotherapy and autologous bone marrow transplantation (5% to 15% *v* 1%), over 40,000 American women were treated with this regimen during the 1990s, and at an estimated cost of $3.4 billion.[124,127]

In 2000 the results of an RCT involving over 550 women became available.[128] This showed that high-dose chemotherapy plus autologous bone marrow transplantation provided no increase in the overall three year survival, or in the median time to disease progression, compared with conventional-dose chemotherapy. Selection bias between the groups was probably responsible for the inappropriate conclusions drawn from the original observations.

During the late 1980s, clinical trialists became less hostile to the concept of historical controlled trials. Prompted by the emerging AIDS epidemic, they accepted that 'some of the traditional approaches to clinical trial design were unnecessarily rigid'.[129] An important paper, by 22 of the world's most respected and experienced clinical investigators, described the criteria under which non-randomised controlled trials might be regarded as evidence to support the use of particular interventions in treating AIDS.[129] These comprised the following specific requirements: (1) There must be no other treatment appropriate to use as a control; (2) There must be sufficient experience to ensure that the patients not receiving treatment will have a uniformly poor prognosis; (3) The therapy must not be expected to have substantial side effects that would compromise the potential benefit to the patient; (4) 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; (5) The scientific rationale for the treatment must be sufficiently strong that a positive result would be widely accepted.

My own adaptation of these requirements is unashamedly influenced by the considerations outlined by Bradford Hill in distinguishing causal from non-causal associations in epidemiological studies.[130] I consider historical controlled trials should be accepted as evidence for effectiveness, provided they meet all of the following conditions:

1. The treatment should have a biologically plausible basis. This is met by all the treatments shown in Tables 4 and 5.
2. 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 an alternative to enzyme replacement therapy in the treatment of Gaucher's disease.
3. The condition should have an established and predictable natural history. I prefer this phraseology to 'poor prognosis'. Conditions such as port wine stains may significantly impair patients' quality of life without threatening life expectancy.
4. The treatment should not be expected to have adverse effects that would compromise its potential benefits. This has to be a sine qua non.
5. There should be a reasonable expectation that the magnitude of the benefits of the treatment will be large enough to make the interpretation of the benefits unambiguous. A 'signal-to-noise' ratio of 10 or more appears to be strongly suggestive of a genuine therapeutic effect.[23,25] The magnitude of the 'signal-to-noise' ratio representing a 'dramatic' (ie 10-fold) response, however, is based on impression and is not (at present) supported by any substantive empirical evidence.

The use of high-dose chemotherapy plus autologous bone marrow transplantation, in the 1990s, for the treatment of advanced breast cancer complied with only one of these conditions (biological plausibility).

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 the UK, Europe and North America are currently accumulating cohorts of patients with both these diseases – specifically for the purpose of providing historical controls for future studies – gives me optimism.[131,132]

Non-randomised, contemporaneous controlled trials

In non-randomised, contemporaneous controlled trials, the fate of patients receiving one treatment is compared to that of a group of untreated patients, or those treated with an alternative intervention, during the same time period. This design feature avoids some of the difficulties encountered in the interpretation of historical controlled trials.[6] Nevertheless, because treatments are not randomly allocated, the potential for selection bias and confounding remains.

Assessment of benefit. A recent example of a non-randomised, contemporaneous controlled trial indicates the value of this technique in circumstances where an RCT would have been impractical. Petersen and colleagues investigated the protective effects of antibiotics against the most serious complications of common respiratory tract infections.[133] The study was specifically designed to show whether, despite the evidence from numerous RCTs that antibiotics for these indications have little or no value, antimicrobial chemotherapy prevents rare but severe complications such as quinsy and mastoiditis. Patients were recruited from UK primary care practices contributing to the General Practice Research Database and the investigators were able to study over 1 million episodes of sore throat, and over 400,000 cases of otitis media, diagnosed between 1991 and 2001.[134] From these data, the authors estimated that the number of children needed to treat,* to avoid one episode of either quinsy or mastoiditis, was 4,300 (95% CI 2,522 to 14,586) and 4,064 (2,393 to 13,456) respectively.

Because this was an observational study, with no random allocation of treatments between the groups, it is possible that there was a degree of selection bias in the prescribing of antibiotics.[133] In particular, more severely affected patients might have been treated with antibiotics more readily than those less severely affected.

*The *number needed to treat* describes the number of patients who would be required to be treated to achieve one beneficial outcome.

Consequently the results may have underestimated the benefits of antibiotics in these indications.[133] Nevertheless, it is clear from the numbers needed to treat that the benefits are minimal, especially when set against the likelihood of gastrointestinal and other adverse effects, sensitising many patients to these drugs, and contributing to antibacterial resistance in the community.

Assessment of harms. Non-randomised contemporaneous controlled trials can also be valuable in the assessment of harms, provided that efforts are made to take account of the potential problems of selection bias and confounding.

Cimetidine, the first H_2-receptor antagonist, was marketed in the UK in 1976 to treat peptic ulceration. Reports soon appeared, however, of an association with the development of carcinoma of the stomach. Duncan Colin-Jones and his colleagues assembled a cohort of over 9,000 patients who had been prescribed cimetidine by their GPs, together with an age- and sex-matched control group (not receiving cimetidine) drawn from the lists of the same doctors.[135] At one year of follow up, the mortality in the cimetidine takers was double that of the controls, mainly due to increased deaths from malignancies (especially oesophagus, stomach, colon and lung).[136]

The investigators concluded this was likely to be due to selection bias. The most probable explanation was that cimetidine was being used knowingly, or unknowingly, to treat the symptoms of various other diseases, as well as being used to alleviate the adverse effects of interventions such as corticosteroids, non-steroidal anti-inflammatory drugs (NSAIDs) and radiotherapy.[136] The authors' conclusions were confirmed by the observation that, at four years, mortality rates among patients using cimetidine were similar to those in the general population matched for age and sex.

Interestingly, in this same cohort, an excess of hospital referrals for suspected cataracts was observed in patients taking cimetidine compared to the controls.[137] The number of patients undergoing cataract surgery, however, was similar in the two groups. The authors concluded that the findings were probably due to ascertainment bias: patients receiving cimetidine visited their doctors more frequently than the controls, and therefore had the opportunity to raise other clinical complaints with their family doctor.

Non-randomised, contemporaneous controlled trials can, unquestionably, provide valuable evidence about the effectiveness of therapeutic interventions provided the potential for bias is taken into account. More needs to be learned, however, about the most appropriate circumstances and conditions for their use.[108,120]

Case-control studies

Case-control studies compare the use of an intervention in groups with, and without, a particular disease or condition. The approach is widely used in epidemiology to identify risk factors for specific conditions (eg smoking and lung cancer) but has also been used extensively to confirm or refute associations between the use of particular products and their adverse effects. The technique has also been used, more controversially, to identify beneficial effects. Case-control studies are conventionally displayed as a 2 × 2 table (Table 6). From this both the odds ratio (OR) and relative risk (RR) can be estimated:

Table 6. A 2 x 2 table for simple case-control studies.

Exposure status	Cases	Controls
Exposed	a	cc
Unexposed	b	d
Totals	a + b	c + d

$$\mathrm{OR} = a/b \div c/d = ad/bc$$

$$\mathrm{RR} = a/(a + c) \div b/(b + d) = a(b + d)/b(a + c)$$

The OR is therefore the ratio of the odds of an event in the two groups, and the RR is the ratio of the risk in the two groups. The difference between them may be small when the event is rare, but when events are common the differences may be large. An RR or OR of 1 indicates that the risk factor is neither harmful nor beneficial. As usually calculated, an RR or OR of >1 suggests that the risk factor is harmful, and an RR or OR of <1 suggests that the risk factor is beneficial.

Case-control studies provide information about an *association* between exposure to a particular intervention but not necessarily whether the relationship is *causal.* Non-causal associations may particularly be a consequence of chance or bias. Chance is less likely if the association is strong and consistent between different studies.[12,130] Recall bias may occur if patients with an adverse event are more likely to remember exposure to the intervention than controls. It can be minimised by determining exposure from prescription records and by 'blinding' patients (and sometimes even research staff) to the specific intervention and condition of enquiry.

The inevitable problems of selection bias and confounding are no less relevant to the interpretation of case-control studies than to other controlled observational designs. They can, however, be minimised in the design as well as in the analysis of case-control studies.[86] Matching of patients in the two groups is frequently used in an attempt to eliminate confounding but will not be effective unless the matching factor is also taken into account in the analysis. Overmatching, on the other hand, will reduce statistical efficiency and can, itself, introduce bias.

Assessment of benefit. Case-control studies have been used, though with mixed results, to provide support for demonstrating the benefits of interventions.

During the 1980s a number of observational (mainly case-control) studies suggested that the long-term use of HRT was associated with a substantial reduction in ischaemic heart disease. Quantitative overviews in the early 1990s indicated that the relative risk in users, compared to non-users, might be associated with a reduction of as much as 50%.[138,139] The potential for selection bias was apparent, to some, even at that time. Women using HRT might well have been different from non-users, for a number of critical risk factors, including socioeconomic status and smoking. None of these were controlled for in these observational studies, and the Committee on Safety of Medicines declined to advise that the labelling of HRTs should include the indication 'for the prevention of ischaemic heart disease' on a number of occasions. Nevertheless, HRTs became one of the most widely prescribed drugs in both the UK and the USA.[140]

It is now known from the results of a number of large, well-conducted RCTs that HRTs have no beneficial effect in ischaemic heart disease and that they increase the risk of stroke.[91] These RCTs did, however, show increased risks of breast cancer and venous thromboembolism. This confirmed the results of previous case-control studies in which the potential for selection bias and confounding for these *adverse* effects was, in my judgement, minimal.[91]

The discrepancies between the results of observational studies and RCTs, in the perceived benefits of HRT, do indeed appear to have been largely due to selection bias. If the observational studies had been able to take account of age, socioeconomic status, smoking habits and duration of use, most (though not all) of the claimed advantages would have disappeared.[141] Some women, though, have paid a high price for the inappropriate use of this observational data in determining benefit.

The failure of case-control studies to provide reliable information about the benefits of HRT contrasts with the success of observational studies in defining the relationship between maternal folate deficiency and neural tube defects. During the 1960s and 1970s, emerging evidence from case-control studies suggested that maternal vitamin deficiency, specifically folate deficiency, was associated with the development of neural tube defects in the offspring.[142,143,144] Non-randomised trials[145] indicated that vitamin supplementation (which included folic acid) around the time of conception, to women who had previously given birth to one or more infants with a neural tube defect, was associated with a substantial reduction in the incidence of these congenital abnormalities. A small randomised study showed a non-significant benefit in preventing neural tube defects.[146]

The results of a large multi-centre RCT,[147] carried out under the auspices of the

MRC and published in 1991, confirmed that periconceptional folate supplementation reduced the incidence of neural tubes in 'at risk' mothers to an extent similar to that seen in observational studies.

I have no criticism of the decision to undertake that MRC trial despite the ethical and legal issues expressed at the time.[148] The degree of uncertainty about extrapolating from the observational studies, coupled with the problem over the dose of folate and its possible adverse effects on the foetus, made it essential.[149] Moreover, I have the greatest admiration for the fortitude, energy and commitment of its investigators. But a heavy price was paid for our inability to be confident in extrapolating from observational studies. During the 10-year period between the recognition of folate deficiency as a cause of neural tube defects, and the publication of the MRC trial, several thousand pregnancies were aborted or resulted in the births of severely disabled children.

There are other circumstances where case-control studies have provided significant indications of the benefits of interventions. These include the protective effects of aspirin against acute MI,[150] the relationship between sleeping posture and sudden infant death syndrome,[151] and the protective effects of NSAIDs and colorectal cancer.[152] For the future 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.[12] Newer techniques, such as Mendelian randomisation* may well assist.[153] More resources, time and energy to undertake methodological research are needed if causality is to be more securely based on observational evidence.

There is a salutary postscript to the folate story. Ten years ago, as a result of the findings of the MRC trial, the USA and Canada legislated for the fortification of flour with folate. As a consequence, these countries have seen a 28% and a 40% reduction (respectively) in the incidence of neural tube defects.[154] In Britain, 17 years after the publication of the results of the MRC's folate trial, folate supplementation is still 'under consideration'.[155]

Assessment of harms. In contrast to the difficulties in assessing the benefits of interventions using case-control designs, this method has been extremely important in identifying causal relationships between specific interventions and their adverse effects. Examples are shown in Table 7. Case-control studies have also been useful in providing reassurance that putative adverse effects 'signalled' by spontaneous reporting schemes do not appear to be problematic. Examples of this include

**Mendelian randomisation* exploits the idea that a genotype affecting the phenotype of interest is assigned randomly at meiosis and independently of confounding factors.[153]

suspected associations between bisphosphonates and AF;[156] and sympathomimetic bronchodilators with excess asthma deaths.[157]

Although selection bias and confounding by indication are less likely to prejudice the results of case-control studies for harms than for benefits, they may still do so. For example, three case-control studies published simultaneously in 1974 suggested an association between the use of reserpine, for the treatment of hypertension, and the subsequent development of breast cancer.[158–160] Other studies, published later, failed to confirm the original association which now appears to have resulted from excluding, as controls, those patients with cardiovascular disease.[161,162] Here, a subtle form of selection bias (exclusion bias) was probably responsible for the erroneous conclusions that were originally drawn.

Before-and-after designs

Observations among groups of patients, before and after treatment, form the basis of many historical controlled trials. In this design, patients are their own controls. Where a 'dramatic' response is observed, and the requirements discussed earlier are met, the benefits can be presumed. Many of the examples given in Tables 4 and 5 are of this type.

Table 7. Some adverse effects confirmed by case-control studies.

Intervention (year of publication)	Adverse effect
Oral contraceptive agents (1967)	Venous thromboembolism[22]
Diethylstilboestrol during pregnancy (1972)	Genital tract carcinoma (in young females)[23]
Aspirin in children (1985)	Reye's syndrome[24]
L-tryptophan (1990)	Eosinophilia-myalgia syndrome[25]
Non-steroidal anti-inflammatory drugs (1994)	Upper gastrointestinal bleeding[26]
Hormone replacement therapy (1996)	Venous thromboembolism[27,28]
Hormone replacement therapy (1997)	Breast cancer[29]
Selective serotonin reuptake inhibitors (1999)	Upper gastrointestinal bleeding[30]
Anticonvulsants (1999)	Stevens-Johnson syndrome and toxic epidermal necrolysis[31]
Olanzapine (2002)	Diabetes[32]
Fluoroquinolones (2002)	Achilles tendon disorders[33]

Before-and-after designs, in conditions with a fluctuating natural history, are of little value. Spontaneous improvement, random symptom fluctuation, regressions to the mean, and patients' politeness – all contributing to the so-called 'placebo effect' – negate their findings. There have, for example, been numerous before-and-after studies of the effects of dopamine antagonists in the treatment of the choreiform movement disorders characteristic of symptomatic Huntington's disease.[163] The influences of random fluctuation, together with other factors contributing to the placebo effect, make it difficult to draw any reliable conclusions.

Case series and case reports

Case series. Healthcare systems, in most developed countries, collect information about their activities, which is primarily used to inform the planning and management of services; to carry out evaluative research and clinical audit; and to provide individual (or groups of) clinicians with estimates of the outcomes of their activities.[164,165] Numerous registries containing details of individual patients have been established by learned societies or other groups of interested health professionals.[164]

Registries containing patient level information about both interventions and outcomes can provide data supporting evidence-based therapeutics in two principle ways.[165] They can (at least in theory) be used to access information about the generalisability of the results of RCTs; and they can offer further evidence about an intervention's safety.

Generalisability. As already discussed, one of the most significant limitations of RCTs is their uncertain generalisability. An example of the use of a case series, to address this problem, was a study of the effectiveness of anticoagulation in non-valvular AF among patients undergoing routine clinical care. Although these patients were on average seven years older, and comprised 33% more women, than those in RCTs the incidence of stroke, and major and minor bleeding, were strikingly similar to those of the pooled RCT results.[80] This provided valuable reassurance about the generalisability of a complicated, and potentially dangerous, intervention when used as part of routine care. Systematic studies examining generalisability, however, appear to be limited, and much more remains to be done in this field. The introduction of the electronic patient record, in Britain, might facilitate this.[166]

Assessment of harms. Case-series, both general and condition specific, can play a useful role in characterising the harmfulness of therapeutic interventions. Three general databases, in particular, have had special value in pharmacovigilance. The General

Practice Research Database[134] and the Medicines Monitoring Unit (MEMO)[167] are large multipurpose databases which contain individual patient data, about the use of interventions as well as outcomes. Both have been used for assessing safety in treated populations, and both have also been successfully used to assemble patient cohorts for other types of observational studies (eg case-control studies). The prescription-event monitoring scheme is different.[168] Organised by the Drug Safety Research Unit, at Southampton, the scheme depends on notifications about patients receiving a particular (usually new) product, identified from NHS prescriptions. Simple questionnaires are sent to the prescribing physician about five months after the first prescription. These request the reporting of any medical 'event' of sufficient significance to have been entered into the patient's notes. Numerous products have now been investigated through this scheme.

Condition-specific databases can also make an important contribution to patient safety. The UK Epilepsy and Pregnancy Register, for example, has collected data since 1996 about pregnant women with epilepsy – whether or not they were taking antiepileptic drugs – and who were referred before the outcome of their pregnancy was known.[169] This design feature therefore avoided the selective reporting of pregnancies with adverse outcomes (a form of ascertainment bias) and the registry has provided important information about the risks of major congenital malformations with antiepileptic drugs individually as well as in combination. The results have been used in the development of national guidelines on the management of epilepsy in pregnancy as well as to inform potential parents of the extent of the risks of major congenital malformations in mothers taking antiepileptic drugs.[170]

Case reports. Reports of individual cases of suspected adverse effects of drugs have long played a substantial role in providing 'signals' about the safety of medicines. The first indications of the teratogenicity of thalidomide,[171] for example, came from a letter to the *Lancet.* [172]

Since that time drug regulatory authorities, in most developed countries, have established formal spontaneous schemes for the reporting of adverse effects to marketed medicines.[173] Although in most countries the scheme is voluntary for healthcare professionals, there is a legal obligation on manufacturers to report any suspected adverse reactions of which they become aware.

Such schemes suffer from disadvantages. Partly because they are voluntary and partly because of the difficulties in distinguishing iatrogenic from non-iatrogenic conditions, under-reporting is substantial. As a consequence, spontaneous reporting schemes are susceptible to reporting bias, especially where there has been media (professional or lay) interest in a particular safety issue. Nevertheless, and despite these limitations, spontaneous reporting schemes have made, and continue to make, substantial contributions to monitoring the safety of pharmaceutical products.[174,175] The first intimations of extrapyramidal reactions (acute dyskinesias and Parkinsonian symptoms) with the antiemetic metoclopramide, for example, arose from spontaneous reports.[176]

Recent developments in the analysis of spontaneous reports have enhanced their value. The standard method, for many years, was to express the numbers of adverse events in relation to the prescription volume of the product.[177] This had various weaknesses. Apart from the biases associated with the numbers of reports (the numerator), prescription volume (the denominator) was only a poor proxy for the number of users because of the inability to distinguish between first and repeat prescriptions.

An alternative approach is to analyse the data on spontaneous reports without using prescription data. The proportional reporting ratio expresses proportion of all reactions to a particular product in comparison to all drugs in the database.[178] A preliminary study of 15 recently marketed drugs showed that, using this approach, 70% of reports were known adverse reactions, 13% were events likely to be related to the underlying disease, and 17% were signals requiring further evaluation. Other so-called 'disproportionality' methods have also been developed including ones using Bayesian approaches.[179,180]

Another method is the 'self-controlled case series method'.[181] This is a retrospective cohort model applied to a defined observation period, conditionally on the number of events experienced by each individual over the observed time. Individuals thus serve as their own controls. The method has been particularly useful in studying the adverse effects of vaccines[182] but may also have wider applications for monitoring drug safety.[183]

Qualitative evidence

The contribution that qualitative evidence can make in informing decision makers about the use of interventions is becoming increasingly recognised.[184–186] Qualitative research involves the collection, analysis and interpretation of data that are not easily reduced to numbers.[184] It can provide information about patients' preferences, and their attitude to risk, as well as their approaches to trading risk for benefit.[187,188,189] Qualitative research can also enable generalisability to be extended to groups and settings beyond those studied in RCTs.[184]

Qualitative research may also provide important insights into the social values expressed by society as a whole.[190] These play a crucial role in shaping the decisions of bodies, like NICE, when advising on the use of interventions for whole healthcare systems such as the NHS.

Hierarchies of evidence

The first hierarchy of evidence was published in the late 1970s.[191] Since then many similar hierarchies, of increasing elaboration and complexity, have appeared in the literature.[3,192–206] Such hierarchical approaches to evidence have not only been adopted by many in the evidence-based medicine and health technology assessment movements, but they have come to dominate the development of clinical guidelines.[194–197,201,203,205]

The hierarchy in Table 1, like others, places RCTs at the highest level with a lesser place for those based on observational studies. Giving such prominence to the results of RCTs, however, is unreasonable. As Bradford Hill, the architect of the RCT, stated so cogently: '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'. [207]

Unquestionably, RCTs have had a profound influence on the practice of modern medicine. There are, though, other ways to establish the benefits of an intervention, where the effects are substantial and 'dramatic', that are no less robust.[25] It is absurd, for example, to regard the evidence for the benefits of thyroxine in myxoedema, or N-acetylcysteine in paracetamol overdose, as any less secure than the evidence for the benefits of thrombolytic therapy in the treatment of MI. Yet a hierarchy such as that in Table 1 would position thrombolysis as Level 1+++, but relegate thyroxine for myxoedema and N-acetylcysteine for paracetamol poisoning to four levels lower down (Level 2+ at best).

RCTs are particularly weak in relationship to generalisability and most especially in the assessment of harms.[87,208] Although RCTs can, indeed, identify those adverse effects that occur relatively commonly, and which appear within the short timescales of their duration, there remains significant limitations. Contrary to a recent claim, only observational studies can realistically offer the evidence required for assessing less common, or long-latency, harms. [209]

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-analytic modelling as well as in the emerging fields of teleoanalysis* and patient preference trials.[115,210] Decision-analytic modelling is at the heart of health economic analysis. It involves synthesising evidence from sources that include RCTs, observational studies, case registries, public health statistics preference surveys and (at least in the US) insurance claim databases.[211] In teleoanalysis, different categories of evidence (experimental as well as

*The term *teleoanalysis* was coined by Jeffrey Aronson from the Greek 'teleos' (meaning 'complete' or 'thorough').

observational) are combined to obtain a quantitative summary of the relationship between the cause of a disease and the extent to which the disease can be prevented or treated.[210] Patient preference trials are generally undertaken in parallel with RCTs among patients who are reluctant to be randomised. A cohort of patients, using the therapeutic option of their choice, is studied alongside randomised patients.[115] Such designs may be useful in assessing generalisability.

The sheer number, as well as the inconsistencies between them, demonstrate the unsatisfactory nature of hierarchies. A survey in 2002 identified 40 such grading systems and noted marked disparities between them; a 2006 study uncovered 20 more.[212,213] When six prominent hierarchies of evidence were compared, there was poor inter-rater agreement.[202] The inconsistencies between systems include the variable prominence given to meta-analyses: some position them above large, high-quality RCTs while others ignore them.[196,198,200,201] There are also inconsistencies between hierarchies in their grading of observational studies: some give a higher rating to cohort studies than case control; some consider them to be all equal; and others reverse the order. None incorporate qualitative evidence except in relation to 'expert opinion'.

Hierarchies attempt to replace judgement with an oversimplistic, pseudo-quantitative, assessment of the quality of the available evidence. Decision makers have to incorporate judgements, as part of their appraisal of the evidence, in reaching their conclusions.[5] Such judgements relate to the extent to which each of the components of the evidence base is 'fit for purpose'. Is it reliable? Does it appear to be generalisable? Do the intervention's benefits outweigh its harms? And so on. Decision makers have to be teleoanalysts. Although techniques such as Bayesian statistics will undoubtedly assist they will not be a substitute for judgement. As William Blake (1757–1827) observed: 'God forbid that truth should be confined to mathematical demonstration'.[214]

Concluding thoughts

Experiment, observation and mathematics – individually and collectively – have a crucial role to play 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 observational data should supplant all other statistical methods. Rather, it is a plea to investigators to continue to develop and improve their methodologies; to decision makers to avoid adopting entrenched positions about the nature of evidence; and for both to accept that the interpretation of evidence requires judgement.

I am aware that those who develop and use hierarchies of evidence are attempting to replace judgements with what, in their eyes, is a more reliable and robust approach to assessing evidence. All my experience tells me they are wrong. It is scientific judgement – conditioned, of course, by the totality of the available evidence – that lies at the heart of making decisions about the benefits and harms of therapeutic interventions.

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,[215] Charles Darwin (1809–82) conceived the theory of evolution as a result of close observation,[216] and Albert Einstein's (1879–1955) special theory of relativity[217] was a mathematical description of 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.[218]

References (text)

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;323:334–6.

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;329:224–7.

6 Pocock SJ. *Clinical trials: a practical approach.* Chichester: John Wiley & Sons, 1983.

7 Matthews JNS. *An introduction to randomized controlled clinical trials.* London: Arnold, 2000.

8 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.

9 Altman DG, Bland JM. Treatment allocation in controlled trials: why randomise? *BMJ* 1999; 318:1209.

10 Collins R, MacMahon S. Reliable assessment of the effects of treatment on mortality and major morbidity, I: clinical trials. *Lancet* 2001;357:373–80.

11 Jüni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. *BMJ* 2001; 323:42–6.

12 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.

13 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408–12.

14 Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials. *Cochrane Database Systematic Rev* 2007;(2):MR000012.

15 Armitage P, Berry G, Matthews JNS. *Statistical methods in medical research*, 4th edn. Oxford: Blackwell, 2002.

16 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.

17 Beauchamp TL, Childress JF. *Principles of bioethics*. Oxford and New York: Oxford University Press. 2001.

18 Evans CH, Ildstad ST (eds). *Small clinical trials: issues and challenges*. Washinton DC: Institute of Medicine, National Academy Press 2001.

19 Gallin JI, Alling DW, Malech HL *et al*. Itraconazole to prevent fungal infections in chronic granulomatous disease. *N Engl J Med* 2003;348:2416–22.

20 Lagakos SW. Clinical trials and rare diseases. *N Engl J Med* 2003;348:2455–56.

21 Tan S-B, Dear KBG, Machin D. Strategy for randomised clinical trials in rare cancers. *BMJ* 2003;327:47–50.

22 Behera M, Kumar A, Soares HK, Sokol L, Djubegovic B. Evidence-based medicine for rare diseases: implications for data interpretation and clinical trial design. *Cancer Control* 2007; 14:160–6.

23 Doll R, Peto R. Randomised controlled trials and retrospective controls. *BMJ* 1980;280:44.

24 Collins R, Peto R, Gray R, Parish S. Large-scale randomized evidence: trials and overviews. In: Warrell DA, Cox TM, Firth JD, Benz EJ (eds), *Oxford textbook of medicine*, 4th edn. Oxford: Oxford University Press, 2003.

25 Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *BMJ* 2007;334:349–51.

26 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.

27 Chalmers I. The lethal consequences of failing to make full use of all relevant evidence about the effects of medical treatments: the importance of systematic reviews. In: Rothwell PM (ed), *Treating individuals: from randomised trials to personalised medicine*. London: Elsevier, 2007.

28 Ware JH, Antman EM. Equivalence trials. *N Engl J Med* 1997;337:1159–61.

29 Levin B. The utility of futility. *Stroke* 2005;36:2331–2.

30 Fleming TR. Current issues in non-inferiority trials. *Stat Med* 2008;27:317–32.

31 Rothman K. No adjustments are needed for multiple comparisons. *Epidemiology* 1990;1:43–6.

32 Sen S. *Statistical issues in drug development*. Chichester: John Wiley & Sons, 1997.

33 Wulff HR, Anderson B, Brandenhoff P, Guttler F. What do doctors know about statistics. *Stat Med* 1987;6:3–10.

34 Goodman SN. Toward evidence-based medical statistics, 1: the P value fallacy. *Ann Intern Med* 1999;130:995–1004.

35 Windish DM, Huot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. *JAMA* 2007;298:1010–22.

36 Altman DG. *Practical statistics for medical research*. London: Chapman Hall, 1991.

37 Feinstein AR. P-values and confidence intervals: two sides of the same unsatisfactory coin. *J Clin Epidemiol* 1998;51:355–60.

38 Burton PR, Gurrin LC, Campbell MJ. Clinical significance not statistical significance: a simple Bayesian alternative to p values. *J Epidemiol Community Health* 1998;52:318–23.

39 Goodman SN. Toward evidence-based medical statistics, 2: the Bayes factor. *Ann Intern Med* 1999;130:1005–13.

40 Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Bayesian methods in health technology assessment: a review. *Health Technol Assess* 2000;4:1–130.

41 Goodman SN. Introduction to Bayesian methods I: measuring the strength of evidence. *Clin Trials* 2005;2:282–90.

42 Wright DB. *Understanding statistics: an introduction for the social sciences.* London: Sage, 1997.

43 Schulz KF, Grimes DA. Multiplicity in randomised trials II: subgroups and interim analyses. *Lancet* 2005;365:1657–61.

44 Grant AM, Altman DG, Babiker AB *et al.* Issues in data monitoring and interim analysis of trials. *Health Technol Assess* 2005;9:1–238.

45 Montori VM, Devereaux PJ, Adhikari NKJ *et al.* Randomised trials stopped early for benefit: a systematic review. *JAMA* 2005;294:2203–9.

46 Trotta F, Apolone G, Garattini S, Tafuri G. Stopping a trial too early in oncology: for patients or for industry. *Ann Oncol* 2008;doi:10.1093/annonc/mdn042

47 Pocock SJ. When not to stop clinical trials for benefit. *JAMA* 2005;294:2228–30.

48 Pocock SJ, Lubsen J. More on subgroup analyses. *N Engl J Med* 2008;358:2076.

49 Lagakos SW. The challenge of subgroup analysis – reporting without distorting. *N Engl J Med* 2006;354:1667–9.

50 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.

51 National Institute for Health and Clinical Excellence. *Guide to the methods of technology appraisal.* London: NICE, 2008.

52 Fletcher J. Subgroup analyses: how to avoid being misled. *BMJ* 2007;335:96–7.

53 Ashby D. Bayesian statistics in medicine: a 25 year review. *Stat Med* 2006;25:3589-3631.

54 ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected myocardial infarction. *Lancet* 1988;ii:349–60.

55 GREAT Group. Feasability, safety, and efficacy of domiciliary thrombolysis by general practitioners: Grampian region early anistreplase trial. *BMJ* 1992;305:1015.

56 Pocock SJ, Speigelhalter DJ. Grampian region early anastroplase trial. *BMJ* 1992;305:1015.

57 Berry DA. Bayesian clinical trials. *Nature Rev Drug Discov* 2006;5:27–36.

58 Berry DA. A case for Bayesianism in clinical trials. *Stat Med* 1993;12:1377–93.

59 Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomised trials. *J R Stat Soc A* 1994;157:357–416.

60 Parmar MKB, Griffiths GO, Spiegalhalter DJ *et al.* Monitoring large randomised clinical trials: a new approach with Bayesian methods. *Lancet* 2001;358:375–81.

61 Berry DA, Wolff MC, Sack D. Decision making during phase III. Randomised controlled trial. *Control Clin Trials* 1994;15:360–78.

62 Abrams K, Ashby D, Errington D. Simple Bayesian analysis in clinical trials: a tutorial. *Control Clin Trials* 1994;15:349–59.

63 Bland JM, Altman DG. Bayesians and frequentists. *BMJ* 1998;317:1151.

64 Berry D, Goodman SN, Louis TA, Temple R. Introduction to Bayesian methods: floor discussion. *Clin Trials* 2005;2:301–4.

65 Food and Drug Administration. *Guidance for the use of Bayesian statistics in medical device trials. Draft guidance.* Bethesda: US Department of Health and Human Services, Food and Drug Administration, 2006.

66 Berry DA. Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis. *Clin Trials* 2005;2;295–300.

67 Grieve AP. 25 years of Bayesian methods in the pharmaceutical industry: a personal, statistical bummel. *Pharm Stat* 2007;6:261–81.

68 Chang M, Boral A. ABC of Bayesian approaches to drug development. *Pharm Med* 2008;22: 141–50.

69 Efron B. *Bayesians, frequentists and scientists*. Presidential Address to the American Statistical Society, 2004. www-stat.stanford.edu

70 Rothwell PM. External validity of randomised controlled trials: 'To whom do the benefits apply?' *Lancet* 2005;365:82–93.

71 Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;312:1215–8.

72 Altman DG, Bland JM. Generalisation and extrapolation. *BMJ* 1998;317:409–10.

73 Tunis S, Stryer DB, Clancy CM. Practical clinical trials: Increasing the value of clinical research for decision making in clinical and health policy. *JAMA* 2003;290:1624–32.

74 Bartlett C, Doyal L, Ebrahim S *et al.* The causes and effects of socio-demographic exclusions from clinical trials. *Health Technol Assess* 2005;9:1–152.

75 Bonell C, Oakley A, Hargreaves J, Strange V, Rees R. Assessment of generalisability in trials of health interventions: suggested framework and systematic review. *BMJ* 2006;333:346–9.

76 Weiss NS, Koepsell TD, Psaty BM. Generalisability of the results of randomised trials. *Arch Intern Med* 2008;168:133–5.

77 Heiat A, Gross CP, Krumhplz HM. Representation of the elderly. Women, and minorities in heart failure trials. *Arch Intern Med* 2002;162:1682–8.

78 Padkin A, Rowan K, Black N. Using high quality clinical databases to complement the results of randomised controlled trials: the case of recombinant human activated protein C. *BMJ* 2001;323:923–6.

79 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.

80 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;320:1236–9.

81 Vist GE, Bryant D, Somerville L, Birminghem T, Oxman AD. Outcomes of patients who participate in randomized controlled trials compared to similar patients receiving similar interventions who do not participate. *Cochrane Database Systematic Rev* 2008;(3):MR000009.

82 Schartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. *J Chronic Dis* 1967;20:637–48.

83 Cochrane A. *Effectiveness and efficiency: random reflections of health services*. London: Nuffield Provincial Hospitals Trust, 1972.

84 Jefferys DB, Leakey D, Lewis JA, Payne S, Rawlins MD. New active substances authorised in the UK between 1972 and 1994. *Br J Clin Pharmacol* 1998;45:151–6.

85 Higgins JPT, Green S (eds). Cochrane handbook for systematic reviews of interventions 4.2.6 (updated September 2006). In: *Cochrane Library* Chichester: John Wiley & Sons, 2006.

86 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.

87 Report of CIOMS Working Group VI. *Management of safety information from clinical trials*. Geneva: Council for International Organizations of Medical Sciences, 2005.

88 Sacks FM, Pfeffer MA, Moye LA *et al.* The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. *N Engl J Med* 1996;335: 1001–9.

89 Bonovas S, Filiousse K, Tsavaris N, Sitaras NM. Use of statins and breast cancer: a meta-analysis of seven randomised clinical trials and nine observational studies. *J Clin Oncol* 2005; 23:8606–11.

90 Teo KK, Yusuf S, Furberg CD. Effects of prophylactic antiarrhythmic drug therapy in acute myocardial infarction. *JAMA* 1993;270:1589–95.

91 Beral V, Banks E, Reeves G. Evidence from randomised trials on the long-term effects of hormone replacement therapy. *Lancet* 2002;360:942–4.

92 Ioannidis JPA, Evans SJW, Gøtzshe PC *et al.* Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med* 2004;141:781–8.

93 Bombardier C, Laine L, Reicin A *et al.* Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. *N Engl J Med* 2000;343: 1520–8.

94 McAdam BF, Catella-Lawson F, Mardini IA, Kapoor S, Lawson JA. Systemic biosynthesis of prostacyclin by cyclooxygenase (COX-2): the human pharmacology of COX-2. *Proc Natl Acad Sci USA* 1999;96:272–7.

95 Kerr DJ, Dunn JA, Langman MJ *et al.* Refecoxib and cardiovascular adverse events in adjuvant treatment of colorectal cancer. *N Engl J Med* 2005;352:360–9.

96 Boers M. Seminal pharmaceutical trials. *Lancet* 2002;360:100–1.

97 Krumhol H, Ross JS, Presier AH, Egilman DS. What have we learned from Vioxx. *BMJ* 2007;334:120–3.

98 Psaty BM, Kronmal RA. Reporting mortality findings in trials of rofecoxib for Alzheimer disease or cognitive impairment: a case study based on documents from the rofecoxib litigation. *JAMA* 2008;299:1813–7.

99 Ioannidis JPA, Lau J. Completeness of safety reporting in randomised trials. *JAMA* 2001; 285:437–43.

100 Whittington CJ, Kendall T, Fonagy P *et al.* Selective serotonin reuptake inhibitors in childhood depression: systematic review of published versus unpublished data. *Lancet* 2004; 363:1341–5.

101 Califf RM. Clinical trials bureaucracy: unintended consequences of well-intentioned policy. *Clin Trials* 2006;3:496–502.

102 Eisenstein EL, Lemons PW, Tardiff BE *et al.* Reducing the costs of phase III cardiovascular trials. *Am Heart J* 2005;149:482–8.

103 Eisenstein EL, Collins R, Cracknell BS *et al.* Sensible approaches for reducing clinical trial costs. *Clin Trials* 2008;5:75–84.

104 CRASH Trial Collaborators. Final results of MRC CRASH, a randomised placebo-controlled trial of intravenous corticosteroid in adults with head injury – outcomes at 6 months. *Lancet* 2005;365:1957–9.

105 Sustainable Trials Study Group. Towards sustainable clinical trials. *BMJ* 2007;334:671–3.

106 Britton A, McKee M, Black N *et al.* Choosing between randomised and non-randomised studies: a systematic review. *Health Technol Assess* 1998;2.

107 Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998;317:1185–90.

108 McKee M, Britton A, Black N *et al.* Interpreting the evidence: choosing between randomised and non-randomised studies. *BMJ* 1999;319:312–5.

109 MacLehose RR, Reeves BC, Harvey IM *et al.* A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technol Assess* 2000;4: 1–154.

110 Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies and the hierarchy of research designs. *N Engl J Med* 2000;342:1887–92.

111 Benson K, Hartz AJ. A comparison of observational studies and randomised controlled trials. *N Engl J Med* 2000;342:1878–86.

112 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.

113 MacMahon S, Collins R. Reliable assessment of the effects of treatment on mortality and major morbidity, II: observational studies. *Lancet* 2001;357:455–62.

114 Deeks JJ, Dinnes J, D'Amico R *et al.* Evaluating non-randomised intervention studies. *Health Technol Assess* 2003;7:1–173.

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

116 Papanikolaou PN, Christidi GD, Ioannidis JPA. Comparison of evidence of harms of medical interventions in randomized and non-randomized studies. *CMAJ* 2006;174:635–41.

117 Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials. *Cochrane Database Systematic Rev* 2007(2):MR000012.

118 Sackett DL, Richardson WS, Rosenberg WMC, Haynes RB. *Evidence-based medicine: how to practice and teach EBM*. New York: Churchill Livingstone, 1997.

119 Pocock SJ, Elbourne DR. Randomised trials or observational tribulations? *N Engl J Med* 2000;342:1907–9.

120 Vandenbroucke JP. When are observational studies as credible as randomised trials? *Lancet* 2004;363:1728–31.

121 Rochon PA, Gurwitz JH, Sykora K *et al.* Readers guide to critical appraisal of cohort studies: 1. Role and design. *BMJ* 2005;330:895–7.

122 Mamdani M, Sykora K, Li P *et al.* Readers guide to critical appraisal of cohort studies: 2. Assessing potential for confounding. *BMJ* 2005;330:960–2.

123 Normand S-LT, Sykora K, Li P *et al.* Readers guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding. *BMJ* 2005;330:1021–3.

124 Mello MM, Brennan TA. The controversy over high-dose chemotherapy with autologous bone marrow transplantation. *Health Aff* 2001;20:101–17.

125 Peters WP, Shpall EJ, Jones RB *et al.* High-dose combination alkylating agents with bone marrow support as initial treatment for metastatic breast cancer. *J Clin Oncol* 1988;6:1368–76.

126 Williams SF, Mick R, Dresser R *et al.* High-dose consolidation therapy with autologous stem-cell rescue in stage IV breast cancer. *J Clin Oncol* 1989;7:1824–30.

127 Eddy DM. High-dose chemotherapy with autologous bone marrow transplantation for the treatment of metastatic breast cancer. *J Clin Oncol* 1992;10:657–70.

128 Stadtmauer, EA, O'Neil A, Goldstein LJ *et al.* Conventional-dose chemotherapy compared with high-dose chemotherapy plus autologous hematopoietic stem-cell transplantation for metastatic breast cancer. *N Engl J Med* 2000;342:1069–76.

129 Byar DP, Schoenfield DA, Green SB *et al.* Design considerations for AIDS trials. *N Engl J Med* 1990;323:1343–8.

130 Hill AB. The environment and disease: association or causation? *Proc R Soc Med* 1965;58: 295–300.

131 Collinge J. Personal communication. 2008

132 Huntington's Study Group. COHORT (NCT 00313495). wwwhuntington-study-group.org.

133 Petersen I, Johnson AM, Islam A *et al.* Protective effect of antibiotics against serious complications of common respiratory tract infections: retrospective cohort study within the UK General Practice Research Database. *BMJ* 2007;335:982.

134 Walley T, Mantgani A. The UK General Practice Research Database. *Lancet* 1997;350: 1097–9.

135 Colin-Jones DG, Langman MJS, Lawson DH, Vessey MP. Postmarketing surveillance of cimetidine: 12 month mortality report. *BMJ* 1983;286:1713–6.

136 Colin-Jones DG, Langman MJS, Lawson DH, Vessey MP. Postmarketing surveillance of the safety of cimetidine: mortality during second, third, and fourth years of follow up. *BMJ* 1985;291:1084–8.

137 Colin-Jones DG, Langman MJS, Lawson DH, Vessey MP. Postmarketing surveillance of the safety of cimetidine. *QJM* 1985;54:253–68.

138 Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiological evidence. *Prev Med* 1991;20:47–63.

139 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.

140 Herrington DM. Hormone replacement therapy and heart disease. Replacing dogma with data. *Circulation* 2003;107:2–4.

141 Petitti DB, Freedman DA. How far can epidemiologists get with statistical adjustment. *Am J Epidemiol* 2005;162:415–8.

142 Hibbard BM. The role of folic acid in pregnancy. *Br J Obstet Gynaecol* 1964;71:529–42.

143 Hibbard ED, Smithells RW. Folic acid metabolism and human embryopathy. *Lancet* 1965;i: 1254.

144 Elwood JH, Nevin NC. Factors associated with anencephalus and spina bifida in Belfast. *Br J Prev Soc Med* 1973;27:73–80.

145 Smithells RW, Nevin NC, Seller MJ *et al.* Further experience of neural tube supplementation for prevention of neural tube defects. *Lancet* 1983;i:1027–39.

146 Laurence KM, James N, Miller MH, Tennant GB, Campbell H. Double-blind randomised controlled trial of folate treatment before conception to prevent recurrence of neural-tube defects. *BMJ* 1981;282:1509–11.

147 MRC Vitamin Study Research Group. Prevention of neural tube defects: results of the Medical Research Council Vitamin Study Group. *Lancet* 1991;338:131–6.

148 Beardsley T. Spina bifida: MRC folate trial to start at last. *Nature* 1983;303:647.

149 Wald NJ, Polani PE. Neural tube defects and vitamins: the need to do a randomise controlled trial. *Br J Obstet Gynaecol* 1984;91:516–23.

150 Boston Collaborative Drug Surveillance Group. Regular aspirin intake and acute myocardial infarction. *BMJ* 1874;i:440–3.

151 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.

152 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.

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

154 De Wals P, Tairou F, Van Allen MI *et al.* Reduction in neural-tube defects after folic acid fortification in Canada. *N Engl J Med* 2007;357:135–42.

155 Bayston R, Russell A, Wald NJ, Hoffbrand AV. Folic acid fortification and cancer risk. *Lancet* 2007;370:2004.

156 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;336:813–6.

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

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

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

160 Heinonen OP, Shapiro S, Tuominen L, Turunen MI. Reserpine use in relation to breast cancer. *Lancet* 1974;2:675–7.

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

162 Horwitz RI, Feinstein AR. Exclusion bias and the false relationship of reserpine and breast cancer. *Arch Intern Med* 1985;145:1873–5.

163 Bonelli RM, Hofman P. A systematic review of the treatment studies in Huntington's disease since 1990. *Expert Opin Pharmacother* 2007;8:141–53.

164 Black N, Barker M, Payne M. Cross sectional survey of multicentre clinical databases in the United Kingdom. *BMJ* 2004;328:1478.

165 Raftery J, Roderick P, Stevens A. Potential use of routine databases in health technology assessment. *Health Technol Assess* 2005;9:1–92.

166 Black N. Maximising research opportunities of new NHS information systems. *BMJ* 2008: 336:106–7.

167 Evans JMM, MacDonald TM. Record linkage for pharmacovigilance in Scotland. *Br J Clin Pharmacol* 1999;47:105–10.

168 Mann RD. Prescription-event monitoring – recent progress and future horizons. *Br J Clin Pharmacol* 1998;46:195–201.

169 Morrow J, Russell A, Guthrie E *et al.* Malformation risks of antiepileptic drugs in pregnancy: a prospective study from the UK Epilepsy and Pregnancy Register. *J Neurol Neurosurg Psychiatry* 2006;77:193–8.

170 Aylward RLM. Epilepsy: a review of reports, guidelines, recommendations and models for the provision of care for patients with epilepsy. *Clin Med* 2008;8:433–8.

171 Mellin GW, Katzenstein M. The saga of thalidomide. *N Engl J Med* 1962;267:1184–93, 1238–44.

172 McBride WG. Thalidomide and congenital abnormalities. *Lancet* 1961;ii:1358.

173 Rawlins MD. Pharmacovigilance: paradise lost, regained or postponed? *J R Coll Physicians* 1995;29:41–9.

174 Rawlins MD. Spontaneous reporting of adverse drug reactions. I. Uses. *Br J Clin Pharmacol* 1988;26:8–13.

175 Davies S, King B, Raine JM. Spontaneous reporting - UK. In: Mann RD, Andrews EB (eds), *Pharmacovigilance*, 2nd edn. Chichester: John Wiley & Sons, 2007.

176 Bateman DN, Rawlins MD, Simpson JM. Extrapyramidal reactions with metoclopramide. *BMJ* 1985;291:930–2.

177 Rawlins MD. Spontaneous reporting of adverse drug reactions. I. The data. *Br J Clin Pharmacol* 1988;26:1–7.

178 Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Saf* 2001;10:483–6.

179 Bate A, Lundqvist M, Edwards IR *et al.* A Bayesian neural network method for adverse reaction signal generation. *Eur J Clin Pharmacol* 1998;54:315–21.

180 DuMouchel W. Bayesian data mining in large frequency tables. With an application to the FDA spontaneous reporting system. *American Statistician* 1999;53:177–90.

181 Whitaker HJ, Farrington CP, Spiessens B, Musonda P. Tutorial in biostatistics: the self-controlled case series method. *Stat Med* 2006;25:1768–97.

182 Farrington CP, Nash J, Miller E. Case series analysis of adverse reactions to vaccines. *Am J Epidemiol* 1996;143:1165–73.

183 Musonda P, Farrington CP, Whitaker H. Sample sizes for self-controlled case series. Tutorial in biostatistics: the self-controlled case series method. *Stat Med* 2006;25:2618–31.

184 Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P. Qualitative research methods in health technology assessment: a review of the literature. *Health Technol Assess* 1998;2(16):1–274.

185 Khan KS, Popay J, Kleijn J. Planning the review. Development of a review protocol. In: Kahn KS, ter Riet G, Glanville J, Snowden AJ, Kleijnen J (eds), *Undertaking systematic reviews on effectiveness: guidance for those carrying out or commissioning reviews.* York: NHS Centre for Reviews and Dissemination, 2001.

186 Dixon-Woods M, Fitzpatrick R. Qualitative research in systematic reviews. *BMJ* 2001;323: 765–6.

187 Carnes D, Anwer Y, Underwood M, Harding G, Parsons S. Influences on older people's decision making regarding choice of topical or oral NSAIDs for knee pain: qualitative analysis. *BMJ* 2008;doi:10.1136/bmj.39401.699603BE

188 Silvestri G, Pritchard R, Welch HG. Preferences for chemotherapy in patients with advanced non-small cell lung cancer: descriptive study based on scripted interviews. *BMJ* 1998;317: 771–5.

189 Koops L, Lindley R. Thrombolysis for acute stroke: consumer involvement in design of new randomised controlled trial. *BMJ* 2002;325:415–8.

190 Rawlins MD. Pharmacopolitics and deliberative democracy. *Clin Med* 2005;5:471–5.

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

192 Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* 1989;95:2–4.

193 Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations the use of antithrombotic agents. *Chest* 1992;102;305–11.

194 Wilson MC, Hayward RSA, Tunis SR, Bass EB, Guyatt G. Users' guides to the medical literature: VIII How to use clinical practice guidelines. B. What are the recommendations and will they help you in caring for your patients. *JAMA* 1995;274:1630–2.

195 Hadorn DC, Baker D, Hodges JS, Hicks N. Rating the quality of evidence for clinical practice guidelines. *J Clin Epidemiol* 1996;49:749–54.

196 US Preventive Services Task Force. *The guide to clinical preventive services*, 2nd edn. Washington DC: US Preventive Services Task Force, 1996.

197 Shekelle P, Woolf SH, Eccles M, Grimshaw J. Developing guidelines. *BMJ* 1999;318:593–6.

198 Guyatt GH, Haynes RB, Jaeschke RZ *et al.* Users' guides to the medical literature: XXV. Evidence-based medicine: principles for applying the users' guides to patient care. *JAMA* 2000;284:1290–6.

199 Harris RP, Helfand M, Woolf SH *et al.* Current methods of the US Preventive Services Task Force: A review of the process. *Am J Prev Med* 2001;20:21–35.

200 Ebell MH, Siwek J, Weiss BD *et al.* A strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. *J Am Board Fam Prac* 2004;17:59–67.

201 GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ* 2004;328:1482–90.

202 Atkins D, Best D, Briss PA *et al.* GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ* 2004;328:1490.

203 Atkins D, Briss PA, Eccles M *et al.* Systems for grading the quality of evidence and the strength of recommendations II: pilot study of the new system. *BMC Health Serv Res* 2005; 5:25.

204 US Preventive Services Task Force. *The guide to clinical preventive services 2005.* Washington DC: Agency for Healthcare Research and Quality, 2005.

205 Oxman AD, Schünemann HJ, Fretheim A. Improving the use of research evidence in guideline development: 7. Deciding what evidence to include. *Health Res Policy Syst* 2006;4: 19–26.

206 US Preventive Services Task Force. *The guide to clinical preventive services 2007.* Washington DC: Agency for Healthcare Research and Quality, 2007.

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

208 Venning GR. Identification of adverse reactions to new drugs. II: How were 18 important adverse reactions discovered and with what delays? *BMJ* 1983;286:289–92.

209 Freemantle N, Irs A. Observational evidence for determining drug safety. *BMJ* 2008;338: 627–8.

210 Wald NJ, Morris JK. Teleoanalysis: combining data from different types of study. *BMJ* 2003; 327:616–8.

211 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.

212 Agency for Healthcare Research and Quality. *Systems to rate the strength of scientific evidence.* Rockville, MD: AHRQ, 2002.

213 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.

214 Blake W. Notes on Reynold's discourses, (circa 1808). Quoted in: Fripp M, Fripp D, Fripp J (eds). *Speaking of science.* London: Newnes, 2000.

215 Mendel G. Experiments in plant hybridization, 1865. www.mendelweb.org/Mendel.Experiment.txt.

216 Darwin C. *The origin of species.* Ware: Wordsworth, 1998

217 Einstein A. *Relativity.* Abingdon: Routledge, 1993

218 Harvey W. *On the motion of the heart and blood in animals.* Trans Willis R, Bowie A. Montana: Kessinger, 1628.

References (tables)

1 Murray GM. Note on the treatment of myxoedema by hypodermic injections of an extract of the thyroid gland of a sheep. *BMJ* 1891;ii:796–7.

2 Banting FG, Best CH, Collip JB, Campbell WR, Fletcher AA. Pancreatic extracts in the treatment of diabetes mellitus. *CMAJ* 1922;12:141–6.

3 Minot GR, Murphy WP. Treatment of pernicious anaemia by a special diet. *JAMA* 1926;87: 470–6.

4 Walker MB. Treatment of myasthenia gravis with physostigmine. *Lancet* 1934;i:1200–1.

5 Colebrook L, Purdie AW. Treatment of 106 cases of puerperal fever by sulphanilamide. *Lancet* 1937;ii:1291–4.

6 Abraham EP, Chain E, Fletcher CM *et al.* Further observations on penicillin. *Lancet* 1941;ii: 177–90.

7 Medical Research Council. Streptomycin treatment of tuberculous meningitis. Lancet 1948;i:582–96.

8 Harington M, Kincaid-Smith P, McMichael J. Results of treatment in malignant hypertension: a seven year experience in 94 cases. *BMJ* 1959;ii:969–80.

9 Einhorn LH, Donohue JP. Cis-diamminedichlorplatinum, vincristine and bleomycin combination chemotherapy in disseminated testicular cancer. *Ann Intern Med* 1977;87:293–8.

10 Prescott LF, Illingworth RN, Critchley JAJH *et al.* Intravenous N-acetylcysteine: the treatment of choice for paracetamol poisoning. *BMJ* 1979;2:1097–1100.

11 Collaborative DHPG Study Group. Treatment of serious cytomegalovirus infections with 9-(1,3-dihydroxy-2-propoxymethyl)guanine in patients with AIDS and other immunodeficiencies. *N Engl J Med* 1986;314:801–5.

12 Barton NW, Furbish FS, Murray GJ, Garfield M, Brady RO. Therapeutic response to intravenous infusions of glucocerebrosidase in a patient with Gaucher disease. *Proceedings of the National Academy of Sciences* 1990;87:1913–6.

13 Garside R, Round A, Dalziel K, Stein K, Royle P. The effectiveness and cost-effectiveness of imatinib in chronic myeloid leukaemia. *Health Technol Assess* 2002;6(33).

14 Wilson J, Connock M, Song F *et al.* Imatinib for the treatment of patients with unresectable and/or metastatic gastrointestinal stromal tumours: systematic review and economic evaluation. *Health Technol Assess* 2005;9(25).

15 Morgan C. Tracheostomy. *E-medicine.* www.emedicine.com/ent/topic356.htm

16 Blundell J. *The history of transfusion medicine.* www.bloodbook.com/trans-history.html

17 Beck CS, Pritchard WH, Feil HS. Ventricullar fibrillation of long duration abolished by electric shock. *JAMA* 1948;135:985–6.

18 Heimlich HJ. A life-saving manoeuvre to prevent food choking. *JAMA* 1975;234:398–401.

19 Cushieri A, Hunter J, Wolfe B, Swanstrom LL, Hutson W. Multicentre prospective evaluation of laparoscopic antireflux surgery. Preliminary report. *Surg Endosc* 1993;7:505–10.

20 Botma M, Bader R, Kubba H. A parent's kiss: evaluating an unusual method for removing nasal foreign bodies in children. *J Laryngol Otol* 2000;114:598–600.

21 Goh CL. Flashlamp-pumped pulse dye laser (585 nm) for the treatment of port wine stains: a study of treatment outcome in 94 Asian patients in Singapore. *Singapore Med J* 2000;41:24–8.

22 Medical Research Council. Risk of thromboembolic disease in women taking oral contraceptives. *BMJ* 1967;ii:355–9.

23 Herbst AL, Kurman RJ, Scully RE, Poskanzer DC. Clear-cell adenocarcinoma of the genital tract in young females. *N Engl J Med* 1972;287:878–81.

24 Hurwitz ES, Barrett MJ, Bregman D *et al.* Public Health Service study on Reye's syndrome and medications. Report of the pilot phase. *N Engl J Med* 1985;313:849–57.

25 Eidson M, Philen RM, Sewell CM, Voorhees R, Kilbourne EM. L-tryptophan and eosinophilia-myalgia. *Lancet* 1990;335;645–8.

26 Langman MJ, Weil J, Wainwright P *et al.* Risk of bleeding peptic ulcer associated with individual non-steroidal anti0inflammatory drugs. *Lancet* 1994;343:1075–8.

27 Daly E, Vessey MP, Hawkins MM *et al.* Risk of thromboembolism in users of hormone replacement therapy. *Lancet* 1996;348:977–80.

28 Jick H, Derby LE, Myers MW, Vasilakis C, Newton KM. Risk of hospital admission for idiopathic venous among users of postmenopausal oestrogens. *Lancet* 1996;348:981–3.

29 Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108 411 women without breast cancer. *Lancet* 1997; 350:1047–59.

30 de Abajo FJ, Rodríguez FAG, Montero D. Association between selective serotonin reuptake inhibitors and upper gastrointestinal bleeding: population based case control study. *BMJ* 1999;319:106–1109.

31 Razny B, Correia O, Kelly JP *et al.* Risk of Stevens-Johnson syndrome and toxic epidermal necrolysis during first weeks of antiepileptic therapy: a case-control study. *Lancet* 1999;353:2190.

32 Koro CE, Fedder DO, L'Italien GJ *et al.* Assessment of independent effect of olanzapine and risperidone on risk of diabetes among patients with schizophrenia: population based nested case-control study. *BMJ* 2002;325:243–7.

33 van der Linden PD, Sturkenboom MCJM, Herings RMC, Leufkens HGM, Stricker BHCh. Fluoroquinolones and risk of Achilles tendon disorders: case-control study. *BMJ* 2002;324: 1306–7.

RECENT HARVEIAN ORATIONS

2007	Mark Pepys *Science and serendipity*
2006	Sir Michael Marmot *Health in an unequal world*
2005	Colin Blakemore *In celebration of cerebration*
2004	Sir Keith Peters *Exceptional matters*
2003	Sir Paul Nurse *The great ideas of biology*
2002	Sir Cyril Chantler *The second greatest benefit to mankind?*
2001	David Warrell *To search and studdy out the secrett of tropical diseases by way of experiment*
2000	Lord Turnberg *Science, society and the perplexed physician*
1999	Sir Brian Jarman *The quality of care in hospitals*
1998	Sir Donald Acheson *Equality of health: dream or reality?*
1997	Sir John Grimley Evans *A correct compassion: the medical response to an ageing society*
1996	Sir Walter Bodmer *The somatic evolution of cancer*
1995	John D Swales *The growth of medical science – the lessons of Malthus*
1994	Dame Margaret Turner-Warwick *The marvel of the lung and human responsibility – A great contempt of God's good gifts?*
1993	Sir Colin Dollery *Medicine and the pharmacological revolution*
1992	Sir David Weatherall *The role of nature and nurture in common diseases: Garrod's legacy*