WHAT CONSTITUTES CREDIBLE EVIDENCE OF EFFECTIVENESS?

Professor Sir Michael Rawlins

19th Annual Lecture 2012

OHE Office of Health Economics 50 YEARS
50 years of research and expertise

What Constitutes Credible Evidence of Effectiveness?

19th Annual Lecture

Professor Sir Michael Rawlins

18 June 2012

Royal College of Physicians

London

About the Office of Health Economics

Founded in 1962, the OHE's terms of reference are to:

- Commission and undertake research on the economics of health and health care
- Collect and analyse health and health care data for the UK and other countries
- Disseminate the results of this work and stimulate discussion of them and their policy implications.

The OHE's work is supported by research grants and consultancy revenues from a wide range of UK and international sources.

The views expressed in this publication are those of the author and do not necessarily represent those of OHE.

Office of Health Economics
Southside, 7th Floor
105 Victoria Street
London SW1E 6QT
United Kingdom
www.ohe.org

Printed in the United Kingdom
ISBN 978-1-899040-29-2

Contents

About the author

Professor Sir Michael Rawlins, President of the Royal Society of Medicine, was chairman of the National Institute of Health and Clinical Excellence from its formation in 1999 until 2013. He is Chairman of the UK Biobank, an honorary professor at the London School of Hygiene and Tropical Medicine and the University of London, and an emeritus professor at the University of Newcastle upon Tyne.

From 1973 to 2006, Sir Michael was the Ruth and Lionel Jacobson Professor of Clinical Pharmacology at the University of Newcastle upon Tyne. At the same time, he was consultant physician and consultant clinical pharmacologist to the Newcastle Hospitals NHS Trust. He has been Vice Chairman (1987-92) and Chairman (1993-98) of the Committee on Safety of Medicines and Chairman of the Advisory Council on the Misuse of Drugs (1998-2008).

Sir Michael has won several honours and awards, including the Hutchinson and Galen medals. He has authored numerous articles, books chapters and official publications. Sir Michael joined OHE's Policy Board in 2013.

Deductive Versus Inductive Approaches to Science

For centuries, philosophers of science have debated the nature of the evidence that should form the basis for scientific inference. Nowhere has this discussion been louder and more contentious than when focussed on evidence to support the use of therapeutic interventions.

Early philosophers, such as Robert Hooke and Robert Boyle in the 17th century, believed science could be understood only through experimentation, preferably based on explicit hypotheses—the so-called "deductive approach" to scientific inference. The 20th century philosopher Karl Popper went further, asserting science always required experiments designed and conducted to prove a premise or a hypothesis false. If a hypothesis or a premise was incapable of falsification, he believed, it was not scientific. His approach, known as "deductive falsification", attracted a large and lasting following.

In contrast, men such as Francis Bacon, René Descartes and, especially, Thomas Hobbes, regarded observation as the appropriate approach. This inductive approach to science produced the formulation of a hypothesis or premise as a result of observation. The idea that induction should, or even could, form the basis of scientific inference was challenged by the 18th century philosopher David Hume who enunciated what has become known as the "problem of induction". As he famously put it, "Probability is founded on the presumption of a resemblance betwixt those objects of which we have experience, and those of which we have not; and therefore 'tis impossible this assumption can arise from probability". Or, put in different terms, "Just because I have seen a hundred white swans, I cannot assume that all swans are white".

It takes little reflection to appreciate that the dispute between the merits of deduction and induction are absurd. Whole swathes of science depend largely or exclusively on inductive inference, including astronomy, geology, palaeontology, evolutionary biology and genetics. To take just one example, our understanding of the genetic basis of Down syndrome, trisome 21, is based solely on the fact that every case that ever has been studied has had this particular chromosomal abnormality. The sterility of these philosophical debates led Richard Feynman, who was awarded the Nobel Prize for Physics in 1965 for his work on quantum mechanics, purportedly to say that: "Philosophers of science are as useful to science as are ornithologists to birds".

The dispute about the relative merits of deductive and inductive approaches to scientific inference has its apotheosis in the so-called "hierarchies of evidence" that bedevil the evaluation of therapeutic interventions. Figure 1 shows just one of literally dozens of hierarchies that have been produced, all of which place randomised controlled trials at the summit and observational studies in the foothills. Such hierarchies are used widely by public health bodies, health technology assessment agencies, clinical practice guideline developers, and others in assessing the quality of the available evidence about a health technology and informing the strength of any recommendations.

Evidence has but one purpose: to inform decision makers, whether decisions affect an individual patient or an entire health care system. What is important is not the method itself, but whether the particular method is fit for purpose. In drawing conclusions, decision makers must exercise judgement; hierarchies are no substitute.

The evidence for effectiveness, then, falls into two broad categories: the deductive approach, epitomised by randomised controlled trials, and the inductive approach, based on observational study design.

Figure 1. Hierarchies of evidence

Level	Description
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 randomized controlled trials or RCTs with a high risk of bias
2++	High quality systematic reviews of case-control or cohort studies with a low risk of confounding, bias or chance and a high probability of causality
2+	Well-conducted case-control or cohort studies with a low risk of confounding, bias or chance and a significant chance that the relationship is not causal
2-	Case-control or cohort studies with a high risk of confounding, bias or chance and a significant risk that the relationship is not causal
3	Non-analytical studies (for example case records, case series)
4	Expert opinion, formal consensus

Source: Harbour and Milles (2001)

Evidence has but one purpose: to inform decision makers, whether decisions affect an individual patient or an entire health care system. What is important is not the method itself, but whether the particular method is fit for purpose. In drawing conclusions, decision makers must exercise judgement; hierarchies are no substitute.

The evidence for effectiveness, then, falls into two broad categories: the deductive approach, epitomised by randomised controlled trials, and the inductive approach, based on observational study design.

The deductive approach: randomised clinical trials

The design and analysis of a conventional randomised controlled trial (RCT) is well known: it compares the effects of two, sometimes more, interventions that have been allocated randomly between the groups. The outcomes for the groups are compared at the end of the study. If the difference between them is statistically significant, one treatment is considered to be superior to the other.

The RCT technique has three great advantages: it minimises bias, especially selection bias, because every patient has an equal chance of receiving each of the treatments. It minimises confounding because, as a result of randomisation, this likely will be distributed evenly among the groups. Finally, provided the groups are of an appropriate size, the RCT minimises the play of chance or random error. Moreover, the statistical calculations are relatively simple, the methodology is widely accepted and the criteria for significance are well established. RCTs,

however, have inherent limitations and weaknesses that are discussed too infrequently outside the hallowed portals of the Royal Statistical Society.

In discussing these limitations, I do not wish to be misunderstood. Critics of randomised controlled trials are often caricatured as creationists, believers in intelligent design or founding members of the Flat Earth Society. My position is clear: I fully accept that randomised controlled trials have played a major role in 20th century medicine and I expect them to continue to make equally important contributions in the 21st century. I have myself published the results of 19 RCTs and I do not believe I have entirely wasted my time.

Conventional RCTs, however, do present particular problems. First, they may involve some serious statistical issues. Second, RCTs may not be generalisable to the circumstances beyond the population studied; this applies particularly to the evidence for safety. Third, RCTs have become outrageously expensive, whether they are carried out by academia or industry.

Statistical issues

The conventional approach to the design and analysis of RCTs is based on a blend of the ideas of Sir Ronald Fischer, on the one hand, and Jerzy Neyman and Egon Pearson, on the other, in what has become known as the "frequentist approach" to the analysis of RCTs. Such analyses assume no difference between the treatments. This null hypothesis is formally examined in the language of statistics by estimating the probability—the frequency—of obtaining a result as extreme or more extreme than the one observed were the null hypothesis to be true. "Extreme" is generally set at a frequency of less than 1 in 20—in other words p is less than 0.05.

What are the problems with this approach? First, this quantitative definition of "extreme" is both entirely arbitrary and inconsistently applied. Furthermore, a p-value less than 0.05 does not mean that there is a 95% chance that the hypothesis is correct; when the p-value is greater than 0.05, it does not necessarily mean that the treatment is ineffective. Contrary to what many believe, the p-value does not distinguish truth from falsehood.

Second, the null hypothesis ignores previous studies. For example, at the end of a successful Phase II study in a drug development programme, the developers almost invariably know whether the new drug is effective—but the analysis of Phase III studies starts all over again with the null hypothesis.

Third, the null hypothesis is clumsy with studies that are designed to investigate whether there is "no" difference or "not much" difference between treatment groups. Equivalence, non-inferiority and the ineptly named "futility design" are used, but all require some serious intellectual acrobatics to even begin to work; most notably, and to my mind absurdly, they involve reversing the null hypothesis.

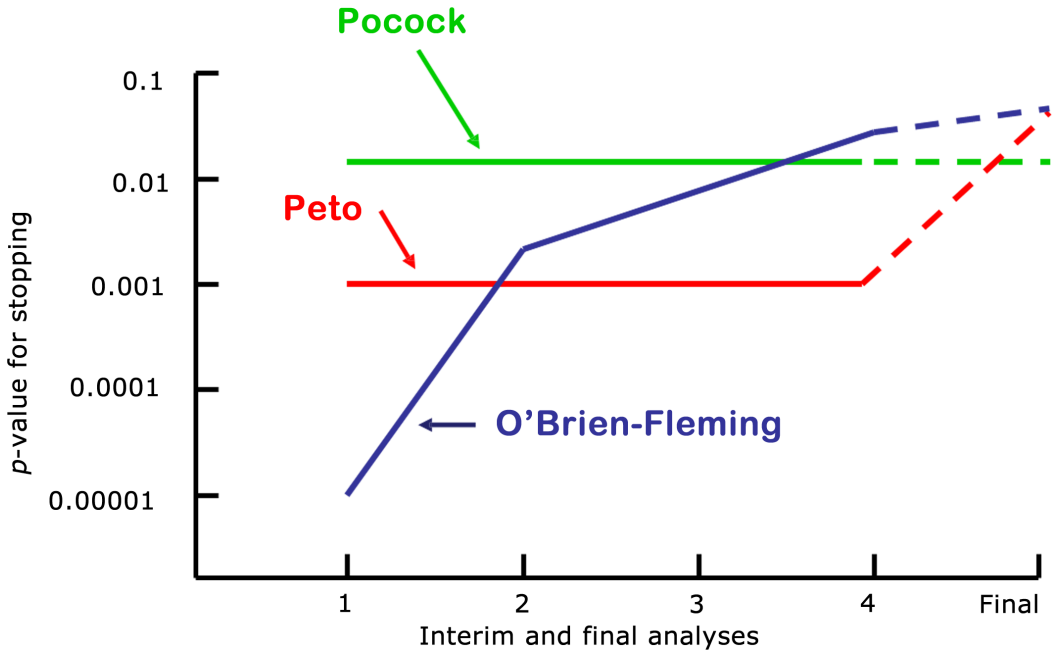
These difficulties become much greater, and in many ways more serious, when dealing with what the statisticians call "multiplicity". If data are subjected to repeated comparisons, whether these be interim analyses during a trial or sub-group or safety analyses at the end of a trial, it is increasingly likely that one will reach conventional levels of statistical significance. If ten different comparisons are made, 40% will produce a p-value of less than 0.05 purely by chance.

The problem of multiplicity is most troubling when analyses are repeated during the course of a trial, known as "interim analyses". Good reasons exist for examining the results of a study still in

progress: if the treatment is showing unacceptable toxicity or dramatic benefits, terminating the study early may be wise. The problem is this: if the results of a trial are examined often enough, getting a result where *p* is less than 0.05 becomes almost a certainty. Statistician do no agree how to distinguish a false positive result from a true positive during interim analyses.

Several approaches for addressing the multiplicity issue have been proposed. Figure 2 shows the various *p*-values that might be reason for stopping a trial early assuming four interim analyses before the fifth, final analysis. Stuart Pocock's scheme gives a constant *p*-value throughout, in this case 0.016, but his final analysis also uses the same, more stringent, *p*-value and therefore may produce a false negative conclusion.

Figure 2. Statistical issues: interim and final analyses



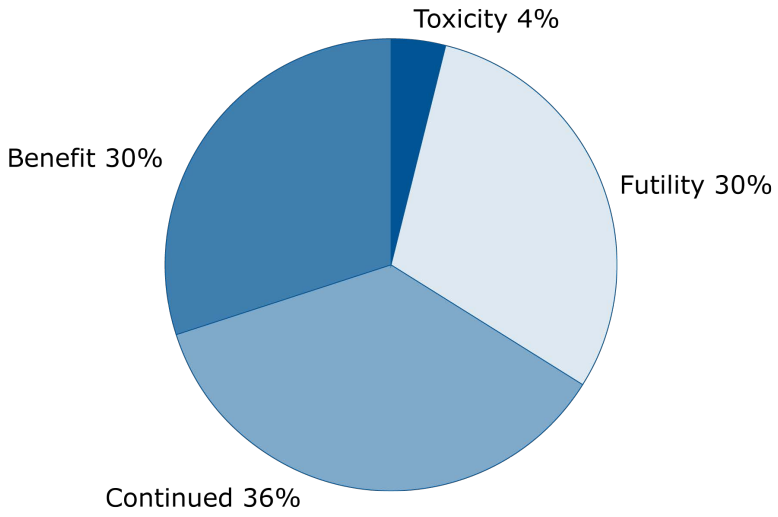
Richard Peto's arrangement proposes a more rigorous, but constant, *p*-value for the interim analyses and a conventional *p*-value for the final analysis. This gets around one of the objections to Pocock's scheme, but provides, for some, too severe a test during the later interim analysis. Under the O'Brien-Fleming scheme, the *p*-value changes substantially with each successive analysis to avoid stopping the study too early.

All this may seem like arcane squabbles among statisticians, but it has increasingly serious implications for clinical care. Clear evidence, especially in oncology, shows that trials are stopped early more often than in the past. Figure 3 is based on a relatively recent review of oncology trials by Trotta, et al (2008). They found that only about a third of trials continued to the end as planned; two-thirds were stopped early. A small number were stopped early for toxicity; about a third were terminated early for futility, meaning that the treatment was showing little or no benefit; and a third were stopped early for benefit.

In oncology, however, the benefits often are relatively modest, at best. Early termination for benefit sometimes may represent what is known in the trade as a "random high", a false-positive result, rather than a true positive one. Even more uncertainty will occur when the benefit has been expressed using a surrogate such as response rate or progression-free survival. A reliable and consistent approach to distinguishing false positives from true positives and true negatives in

interim analyses during clinical trials is needed urgently. Special studies also are needed to check whether the results promised from interim analyses of clinical trials are being achieved in routine care—what some call “real world” data.

Figure 3. Oncology trials: interim analysis (n=93)



Source: After Trotta, et al (2008)

Generalisability

The second major difficulty with RCTs is generalisability. Can the benefits and harms seen in the selected group of patients involved in the clinical trial be extrapolated to the wider population of patients that will use the intervention in the real world? The concern is genuine. First, clinical trials involve a few hundred, sometimes a few thousand individuals, but many millions may use the intervention once it reaches the market. Second, individuals who take part in clinical trials generally are a very homogenous group. Certain groups often are underrepresented—the young, especially children and adolescents; the elderly, particularly the very elderly; ethnic minorities, particularly in trials carried out in Europe; and people who have diseases in addition to the one for which the treatment is being tested, known as “co-morbidities”.

RCTs usually last a few months, occasionally a year or two, but patients may take the medicine for the rest of their lives. Can we really generalise the results of RCTs in the assessment of benefits and harm? It generally is assumed that extrapolating an intervention’s benefits is reasonable, even though the evidence base for that assumption is extraordinarily weak.

RCTs are not necessarily reliable when it comes to harm. Experience over the last 50 years has shown that RCTs generally fail to recognise both less common adverse reactions and those with a long latency, such as cancer. The reasons are largely statistical: to have a 95% chance of seeing an adverse reaction just once, three times that number of patients must be studied. So if a reaction occurs at a rate of one in a hundred, 300 patients will need to be needed to reach a 95% chance of seeing just one case. No adverse reaction, though, is uniquely iatrogenic. This means that to provide reassurance that a possible reaction is not just the play of chance, ten or twenty times that number may need to be studied to avoid accepting a false positive signal.

In summary, RCTs have merit, but they also present problems. To regard the RCT as the “gold standard”, as it often is called, is unsustainable. Austin Bradford Hill, the architect of the modern

randomised controlled trial, put it this way 40 years ago: “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” (Hill, 1966).

The inductive approach: the Bayes-Laplace Theorem

What other approaches are available? How might the problems of the null hypothesis, the *p*-value and multiplicity be resolved? An increasing number of statisticians believe that greater use of Bayesian approaches might offer a solution. Thomas Bayes was an 18th century, non-conformist minister in Tunbridge Wells, England, whose essay on probability was published posthumously.

The frequentist approach has dominated the design and analysis of clinical trials for 60 years. It examines the probability of specific data, such as the result of an RCT, conditional on a particular *hypothesis*, usually the null hypothesis. The Bayesian approach, on the other hand, looks at the probability of a particular hypothesis conditional on specific *data*, which may be either experimental or observational. The reverse of the frequentist approach, it does not involve a null hypothesis or a *p*-value. An inductive, rather than a deductive, approach to scientific inference, it often has been called either the “probability of causes” or “inverse probability”.

Although this approach is associated with Thomas Bayes, many historians believe that Laplace deserves equal credit. Famous for his work in mathematics, astronomy, physics and statistics, Laplace was aware of Bayes’s approach to probability and was the first to express Bayes’s theorem in mathematical terms. Some therefore speak and write of the “Bayes-Laplace theorem”. The association of the theorem with Laplace, however, probably is one of the reasons why interest in Bayes’s theorem declined during the 19th century. Laplace was a very public supporter and friend of Napoleon, who made him a marquis. His association with Napoleon resulted in the rejection of much of his work by many 18th and 19th century mathematicians. During the latter part of the 20th century, however, Bayes’s theorem has experienced a substantial resurgence.

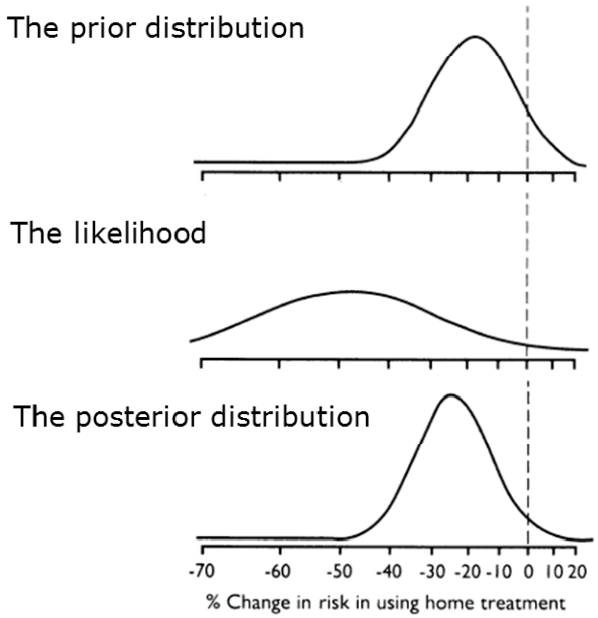
Bayes’s theorem, or the Bayes-Laplace theorem, relates the probabilities about what is already know—the “prior”—to what new experimental or observation evidence shows can now be inferred—the “posterior”. Bayes’s theorem is widely used outside medicine. Spam filters in email systems, for example, rely on it to distinguish between genuine and unwanted e-mail messages. Bookmakers are instinctively Bayesian—a horse’s form provides the result of its outing last week, the prior odds; the 2:30 at Haydock Park corresponds to the new experiment. The odds offered for the 5:20 at Cheltenham, then, are the posterior odds, and bookies expect to make a 10% profit on turnover from each race meeting.

One example of the application of Bayesian statistics in the analysis of a clinical trial is the Grampian region early anistreplase trial, better known as the GREAT study. By the mid 1980s, it was well known that the earlier thrombolytic therapy was given to treat myocardial infarction, the better the outlook for the patient. The GREAT trial was designed to test whether early treatment of acute myocardial infarction with thrombolytic therapy in the patient’s own home would improve survival compared to later treatment after the patient reached hospital. This was the design: patients with a suspected myocardial infarction were seen by their GP at home. If the ECG confirmed the diagnosis, the patient was randomised to either immediate treatment at home or treatment after reaching the hospital. The mortality at three months was 8% for those treated at home, compared to over 15% for those treated once they reached hospital. The *p*-value was 0.04, suggesting that domiciliary thrombolysis might reduce mortality by up to 50% compared to later treatment in hospital. If this was true, the implications would be profound for the NHS and health

care generally. Organising domiciliary thrombolysis treatment on a national scale would have major implications.

Stuart Pocock and David Spiegelhalter, two of the UK's most distinguished statisticians, believed that the results might be overestimating the benefits of domiciliary thrombolysis treatment and undertook a Bayesian reanalysis of the data. The results of are shown in Figure 4 as the percentage reduction in mortality with home treatment. Pocock and Spiegelhalter (1992) derived a prior based on the results of previous trials of hospital thrombolysis. This was that a 15–20% reduction in mortality was highly plausible, but the extremes of no reduction and a 40% reduction were unlikely. The "likelihood" in Figure 4 is the result of the GREAT trial, again showing a probability distribution. The posterior represents the results of revising the power in the light of the GREAT trial. This now provides the most reasonable indication of the likely benefits of domiciliary thrombolysis treatment.

Figure 4. The GREAT trial: a Bayesian re-analysis



Source: Pocock and Spiegelhalter (1992)

Three points are important here. First, the prior, the likelihood, and the posterior all are represented as probability distributions. Each curve shows the probability of the reductions in mortality. Because each curve covers the whole probability distribution, the area under each curve is 1.0, with the maximum at the peak corresponding to the most likely reduction in mortality. Second, the results of the GREAT trial show a very flat probability distribution. This reflects the degree of uncertainty about the magnitude of the claimed benefits because of the relatively small numbers. Third, the posterior distribution shows that domiciliary thrombolytic therapy is most likely to produce a 24% reduction in mortality compared to hospital treatment. The results of the GREAT trial were too good to be true; less advantage would be expected in reality from domiciliary thrombolysis therapy. A massive restructuring of health care delivery would have had much less impact than might have been predicted from the original GREAT trial—and all this with no null hypothesis and no *p*-value.

Bayesian approaches hold out great promise—no null hypothesis, no *p*-value, no prior calculation and a potential resolution of the problem of multiplicity. Why, then, are they not used more widely? First, some find the subjective interpretation of probability distasteful. They prefer instead the apparent, but entirely illusory, security of a *p*-value to distinguish between an extreme and a non-extreme result. Second, some controversy has existed around the derivation of the prior, although this seems to be waning. Third, Bayesian analyses are computationally complex and would have been impossible without modern computing capacity. Fourth, some statisticians are not equipped to do Bayesian analyses. Finally, regulatory authorities have been reluctant in the past to accept Bayesian analysis, although this now appears to be changing.

Observational studies

A number of observational techniques have been used in the assessment of the effectiveness of interventions, each with its own particular advantages and disadvantages. Two of these are historical controlled trials and case-control studies.

Observational studies in general have three potential strengths: in some cases, they can provide an alternative to RCTs in the assessment of benefit; they play a critical role in the assessment of harms; and they provide valuable data about generalisability. Their two important weaknesses are potential selection bias, because the comparisons between treatments are not random, and confounding by indication. “Selection bias” is a systematic error in creating the intervention groups so that they differ with respect to prognosis. Confounding by indication happens when the effect of an intervention is biased by differences between the groups that predict the outcome.

Historical controlled trials are one kind of observational study. These compare a historical cohort of patients and a group of patients currently treated, usually with a new intervention. The historical cohort may be implicit, that is to say, it may be what is generally and reliably known about the natural history of a condition. Or it may be explicit, based on direct comparison with a specific group of patients in which the prognosis and progress of the condition has been monitored previously. The evidence of benefit comes solely from historical controlled trials for numerous conditions and confidence in effectiveness of the treatment is absolute. Figure 5 provides examples.

These include:

- The use of thyroxin in the treatment of myxoedema, introduced by a physician in the north east of England, Dr George Murray, in 1891, essentially produces a complete cure with the right dose.

- Ganciclovir for cytomegalovirus (CMV) retinitis, is a condition to which patients with HIV-AIDS are particularly prone. In CMV retinitis, the lesions become confluent and cause blindness about a year after the onset of symptoms. In the mid 1980s, a new anti-viral agent, ganciclovir, became available. It was observed that it arrested the progression of the disease and prevented blindness. This was regarded in the UK as adequate evidence for licencing, but some countries required evidence from a randomised, placebo-controlled trial. The results two years later confirmed what was already know.

  The price for requiring an RCT was high: the 200 patients in the placebo arm went blind. Equally catastrophic consequences were experienced by patients with CMV

retinitis who lived in countries where ganciclovir was not licensed while authorities waited for completion of an RCT. In the intervening period, before a licence was granted, many more went blind. Clearly, then, the failure to accept the evidence from historical controlled trials can cause great harm to patients.

- Laser therapy can erase “port wine stain” birthmarks, which are vascular abnormalities the lie just beneath the skin, are present from birth and persist throughout life. No physical harm comes from them, but they can be very unsightly when present on an exposed part of the body, especially the face. Relatively recently, it has been shown that laser treatment results in almost complete disappearance of the lesion over a period of about six months. An RCT is not needed: we know beyond any doubt that without treatment port wine stains stay for life and that they disappear with laser therapy.

Figure 5. Evidence of benefit from historical-controlled trials

Indication	Intervention
Myxedema	Thyroxine (1891)
Tuberculous meningitis	Streptomycin (1948)
Ventricular fibrillation	Defibrillation (1948)
Malignant hypertension	Ganglion blockers (1959)
Laryngeal obstruction	Heimlich manoeuvre (1975)
Paracetamol poisoning	N-acetylcysteine (1979)
CMV retinitis	Ganciclovir (1986)
Gaucher’s disease	Imiglucerase (1990)
Port wine stains	Laser therapy (2000)
Chronic myeloid leukaemia	Imatinib (2002)
Gastrointestinal stromal tumours	Imatinib (2005)

Historical controlled trials are not invariably reliable. Five conditions should be fulfilled before the efficacy results of a historical controlled trial can be accepted as generalisable.

1. It should be a biologically plausible form of treatment
2. No appropriate comparator should be available
3. The condition should have a predictable natural history
4. The adverse effects should not be expected to compromise the potential benefits
5. The effect of the treatment should be substantial.

Sir Iain Chalmers, Paul Glasziou and I have suggested that the signal-to-noise ratio should be perhaps in the order of ten, or more, to one. Effect sizes of this magnitude largely overcome any potential for selection bias and confounding.

Case-control studies are a second type of observational study. These compare the rates of exposure to a drug in patients with a particular condition (the cases) to patients without the condition (the controls). Of course, at the outset some idea of the specific condition or harm is necessary.

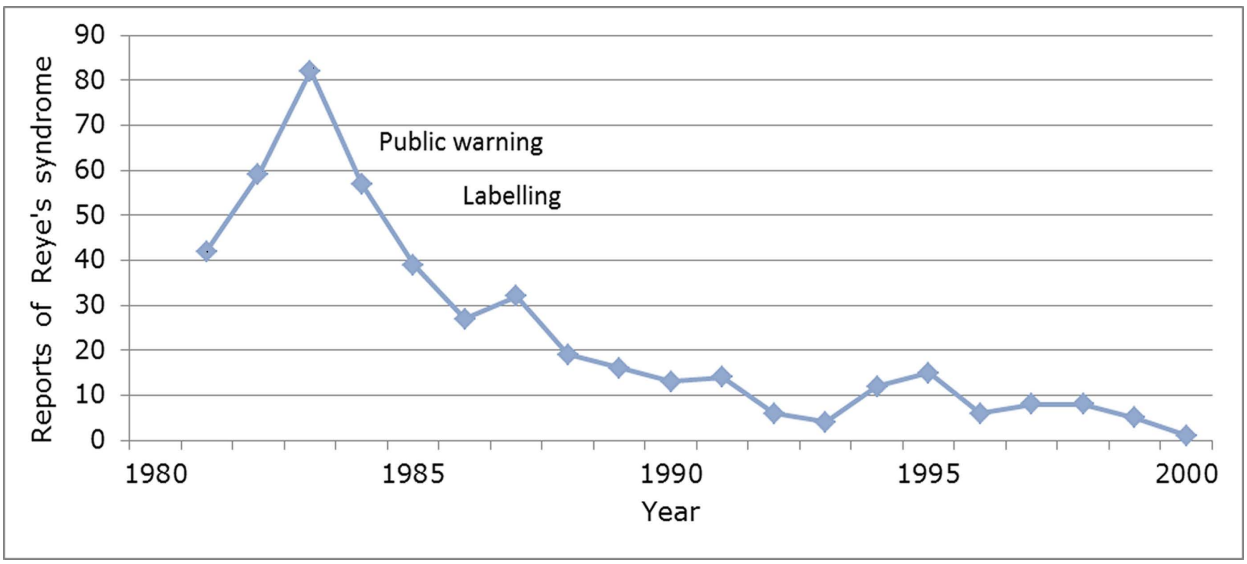
An important example of the use of the case-control method is in understanding the cause of Reye's syndrome. This affects children usually after a mild febrile illness, such as influenza. An acute encephalopathy develops suddenly together with evidence of severe liver damage. The condition has a 30–40% mortality rate and about half of the survivors have permanent neurological damage.

A case-control study examined prior exposure to aspirin in cases of Reye's syndrome compared to controls. It found that a much higher proportion of Reye's syndrome cases (94 out of 97) were children who had received aspirin for treatment of the fever associated with their mild febrile illness. From the ratio of the two ranges, an odds ratio was calculated to produce a measure of the association between the aspirin use and Reye's syndrome. An odds ratio of 13 was calculated, suggesting that a child with prior exposure to aspirin is 13 times more likely to develop Reye's syndrome.

Although an odds ratio of 13 is a very strong association, additional evidence is needed to determine causality. One particularly powerful form of additional evidence is replication. Seven case-control studies that examined the association between exposure to aspirin and the development of Reye's syndrome showed an increased risk. These studies also showed no association between Reye's syndrome and either paracetamol or ibuprofen; the syndrome was specific to aspirin.

Eventually, and not without considerable difficulty, paediatric preparations of aspirin were removed from the market in both the UK and the US and extensive public health campaigns were carried out in each country. The effects have been dramatic. Figure 6 shows the decline of Reye's syndrome in the UK, where it has virtually disappeared. The same thing has happened in Australia and the US.

Figure 6. Reye's syndrome: UK data



Source: RCPCH (2001)

The example of the association between Reye's syndrome and aspirin is just one of many case-control studies that have identified important adverse effects of treatment. Other examples include venous thromboembolism with oral contraceptives, breast cancer with hormone replacement therapy and atypical femoral fractures with biphosphonates. Case-control studies have a crucial and established place in assessing the harms of therapeutic interventions.

Case-control studies also can be useful in showing the *absence* of an association between an intervention and a specific adverse effect. One very important example is the reputed association between measles-mumps-rubella (MMR) immunisation and the subsequent appearance of pervasive developmental disorders, autism and autistic spectrum disorders. A 1998 paper that appeared in *The Lancet* purported to show an association, laying blame specifically on the measles component (Wakefield, et al, 2004). Twelve of the thirteen original authors subsequently retracted the paper, but not before immense damage had been done. The widespread publicity given to this study led to tens of thousands of parents in the UK refusing to allow their children to be immunised with the MMR vaccine. The consequences were catastrophic and predictable: as MMR immunisation rates fell, the number of cases of measles rose.

To confirm or refute the association between MMR immunisation and pervasive developmental disorders, several case-control studies were carried out. The most notable was conducted by Liam Smeeth and colleagues (2004). Using the General Practice Research Database, this study identified almost 1,300 children with pervasive developmental disorders, reviewed their immunisation history and compared this with nearly 4,500 controls. Not the slightest suggestion of an increased risk was shown in the adjusted odds ratios and the associated 95% competencies. None of the other studies examining the same issue indicated an increased risk.

The case-control methodology has been used to assess not just harm, but also potential benefits. The first intimation of the preventive effects of aspirin in acute myocardial infarction came from a case-control study, subsequently confirmed by a number of RCTs.

Case-control studies also established the benefit of putting babies to sleep on their backs, rather than their tummies, to prevent sudden infant death syndrome. To have attempted to undertake an RCT would have been completely impractical. The fall in the number of sudden infant deaths as a result of widespread publicity and education based on the care-controlled trial has been profound.

Case-control studies also offered the first indications that neural tube defects, such as spina bifida, could be reduced by increasing folate intake at about the time of conception. This later was confirmed by an RCT in the UK. As a consequence, flour is fortified with folate in both the US and Canada, producing a 30–40% reduction in the incidence of these seriously disabling conditions. The UK has not yet take similar action, even 20 years later.

Because of the potential for bias and confounding, considerable care must be taken in interpreting the results of case-control studies, especially when studying benefits. To take an example, during the late 1980s a number of case-control studies suggested that hormone replacement therapy protected against cardiovascular disease, especially acute myocardial infarction. The pooled estimate suggested that the reduction in risk might be as great as 50%. Many of us at the time thought that the data were likely to be badly confounded: women who took hormone replacement therapy were likely to be a highly selected population, reasonably well-off and probably with a lower risk of cardiovascular disease anyway. Neither Britain nor the US licenced hormone replacement therapy for this purpose because of these doubts, but it was available in both jurisdictions for the treatment of menopausal symptoms. Claims that hormone replacement therapy prevented heart disease nevertheless were extensively discussed in medical journals; many doctors in Britain and other countries prescribed it for that purpose. For a time, hormone replacement therapy was one of the most widely prescribed group of drugs.

Fortunately, several case-control trials were started in the US to address this question, with findings reported in the early 21st century. Not one of these studies showed any significant benefit—just the opposite. The pooled estimate indicated, within the narrowest of confidence intervals, that hormone replacement therapy is ineffective in preventing cardiovascular disease. The original data indeed were confounded; selection bias had probably been responsible for the earlier enthusiastic, but flawed, findings.

Five criteria need to be considered before drawing conclusions about the reliability of evidence drawn from case-control studies.

1. The biological plausibility of the association should help inform evaluation. It is not necessary, or sufficient, but certainly provides supportive evidence.
2. The magnitude of the effect sizer is similarly informative in concluding a causal association. Some have suggested that the odds ratio from a case-control study should be at least 3:1 to inform or infer a causal relationship. This is a crude approach to distinguishing truth from falsehood, albeit the greater the odds ratio the less the opportunity for bias confounding.
3. Observational studies in general, and case-control studies in particular, are susceptible to both bias and confounding. Serious efforts must be made to adjust for these.
4. A dose-response relationship is neither necessary nor sufficient, but can provide useful information.
5. Replication—observing similar results in different studies—provides further evidence of causality.

None of these factors on its own provides a final answer about a causal association. Taken together, however, they can inform a decision-maker's judgement.

An Irrational Ordering of Evidence

In the face of all this, what is the place of hierarchies of evidence such as the one in Figure 7. This one uses a 13-point scale, is even more elaborate than the one in Figure 1 and gives special prominence to systematic reviews. Since their appearance 30 years ago, over 60 different hierarchies have been published. Although they differ in detail, all place RCTs at the top and observational studies near the bottom. The ordering of evidence in this way is entirely irrational. My confidence in the benefits of penicillin in the treatment of lumbar pneumonia, which might fall somewhere in level 2, is no less secure than my confidence in the benefits of trastuzimab (Herceptin®) in the treatment of early breast cancer. Apart from the irrational ordering of evidence, hierarchies are incapable of accommodating studies that combine the results of RCTs with observational data, a technique widely used in decision analytic modelling, and they cannot incorporate the results of qualitative studies.

Evidence hierarchies attempt to replace judgement with an overly simplistic pseudo-quantitative approach to the assessment of the totality of the available evidence. Decision makers, however, must always incorporate judgement as part of their appraisal of the evidence in reaching their conclusions. Such judgements importantly include the extent to which each component of the evidence base is fit for purpose—whether it is reliable and likely to be generalisable, whether the benefits outweigh the harms, and so on. As William Blake put it, "God forbid that truth should be confined to mathematical demonstration".

Experiments and observations, individually and collectively, have a crucial role in providing the evidential base for modern therapeutics. Arguments about the relative importance of each are an unnecessary distraction. Hierarchies should be replaced by embracing a pluralistic approach to evidence. This is neither a plea to abandon RCTs and replace them with observational studies, nor is it a claim that Bayesian approaches to the design of both experimental and observational data should supplant other statistical methods. Rather, it is a plea to investigators to continue to develop and improve their methodologies, for decision-makers to avoid adopting entrenched positions about the nature of evidence, and for both to accept that the interpretation of evidence requires judgement.

Figure 7. Hierarchies of evidence: 13 levels

Level	Description
1a	Systematic review of randomised controlled trials with homogeneity
1a-	Systematic review of randomised controlled trials with some heterogeneity
1b	Individual randomised controlled trial with narrow confidence interval
1c	All or none effects
2a	Systematic review of cohort studies with homogeneity
2a-	Systematic review of cohort studies with some heterogeneity
2b	Individual cohort study including randomised controlled trials with <80% follow-up
2c	Outcomes research or ecological studies
3a	Systematic review of case-control studies with homogeneity
3a-	Systematic review of case-control studies with heterogeneity
3b	Individual case-control studies
4	Case series and poor quality cohort or case control studies
5	Expert opinion without explicit critical appraisal; or based on physiology or “first principles”

I am aware that those who develop and use hierarchies as a replacement for judgement believe that hierarchies are a more reliable and robust approach to assessing evidence. They are dangerously wrong. It is judgement, conditioned by the totality of the evidence base, that lies at the heart of decision-making.

Questions from the audience

QUESTION: How confident are you that regulators are starting to see beyond the RCT?

Rawlins: I was discouraged not long ago at a meeting of the European Medicines Agency where a member of the CPMP said she could not imagine a drug being put on the market without an RCT. Actually, at that meeting, representatives of the FDA were much more sensible on the subject of assessing the effectiveness of drugs for people who are known to be developing, but do not yet have, florid symptoms of diseases such as Alzheimer's or Huntington's. Treatment for these patients must start ten years before such symptoms appear. The FDA understood that an RCT is not the right approach in such cases.

QUESTION: Many diseases—cancer, for example—often are treated with combinations of products that may change over time. How can such therapies be evaluated?

Rawlins: This is a serious problem, particularly for new products. Do we study each in turn, then start studying the combinations in some factorial design? This show the divergence between the rigid methodology of RCTs and the current development paradigm. It is important to rethink this.

QUESTION: Has NICE taken non-RCT evidence into account in its technology assessment?

Rawlins: Yes, when we have it. An example is imatinib (Glivec®), a drug used to treat chronic myeloid leukaemia (CML). This was licensed without an RCT; the decision was based a comparison with a historical control: placebo-treated patients in previous studies of treatment for CML were used as a historical control. The study demonstrated remarkable survival rates. The use of non-RCT evidence also is particularly important in developing clinical practice guidelines.

QUESTION: Can anyone definitely say that Bayesian analyses always are "better" than frequentist analyses?

Rawlins: The methodology needs much more work—for example, RCTs subject to both a frequentist and a Bayesian analysis done in parallel by separate teams. Of course, pharmaceutical companies will hesitate to do this because two different findings could result, and that would seriously complicate decisions by the regulatory authorities. We do need to move away from the idea that we know it all; we do not.

QUESTION: A critical issue has less to do with the approach than with the process. RCTs and observational studies are begun without first understanding what already is known. This is due in part to the fact that over half of RCTs are not reported in publications. Also, no-one updates the systematic review that existed before the RCT with the new data once the RCT is completed. These are fundamental problems within science.

Rawlins: I agree about the lack of openness in science. The Royal Society is publishing a report this week about these issues for science generally, with very specific comments on medicine[1]. Enormous enthusiasm has developed among the scientific community generally about sharing data. This applies to the biological sciences and the physical sciences.

[1]Editor's note: see Royal Society (2012).

References

Harbour, R. and Miller, J. (2001) A new system for grading recommendations in evidence based guidelines. *British Medical Journal*. 323(7308), 334-336.

Hill, A.B. (1966) Reflections on the controlled trial. *Annals of the Rheumatic Diseases.* 25(2), 107-113.

Pocock, S.J. and Spiegelhalter, D.J. (1992) Domiciliary thrombolysis by general practitioners. Letter to the editor. *British Medical Journal*. 305(6860), 1015.

RCPCH (Royal College of Paediatrics and Child Health). (2001) *British Paediatric Surveillance Unit annual report 2000-2001*. London: RCPCH.

Royal Society. (2012) *Science as an open enterprise: Open data for open science.* Report 02/12. London: The Royal Society Science Policy Centre. Available at: http://royalsociety.org/policy/projects/science-public-enterprise/report/ [Accessed 3 June 2013].

Smeeth, L., Cook, C., Fombonne, E., Heavey, L., Rodrigues, L.C., Smith, P.G. and Hall, A.J. (2004) MMR vaccination and pervasive developmental disorders: A case-control study. *The Lancet*. 364 (9438), 963-699.

Trotta, F., Apolone, G., Garrattini, S. and Tafuri, G. (2008) Stopping a trial early in oncology: For patients or for industry? *Annals of Oncology*. 19(7), 1347-1353.

Wakefield, A.J., Murch, S.H., Anthony, A., Linnell, J., Casson, D.M., Malik, M., Berelowitz, A.P., Dhillon, A.P., Thomson, M.A., Harvey, P., Valentine, A., Davies, S.E. and Walker-Smith, J.A. (1998) Ileal lymphoid nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children [retracted]. *The Lancet.* 351(9103), 637-641.

Office of Health Economics • Southside, 7th Floor

105 Victoria Street • London SW1E 6QT

+44 20 7747 8850 • www.ohe.org