

[Home](#)[Contents](#)jameslindlibrary.org

Fair tests of treatments in health care

The James Lind Library has been created to improve general understanding of fair tests of treatments in health care, and how these have evolved over time.

Misleading claims about the effects of treatments are common, so all of us need to be equipped to judge whether claims about the effects of treatments are valid. Without this knowledge, we risk concluding that useless treatments are helpful, or that helpful treatments are useless.

Fair tests of treatment are tests that take steps to obtain reliable information about treatment effects by reducing the misleading influences of [biases](#) and the [play of chance](#). When the need for fair tests of treatments is ignored, people suffer and die unnecessarily.

The explanatory essays in *The James Lind Library* have been written to promote wider understanding of why fair tests of treatments are needed, and what they have come to consist of. You can access each essay by clicking on the underlined words, below, or by selecting them from the [Contents](#) screen. If you want to download all of the essays, so that they can be printed out together for reading off screen, [click here](#).

[Fair tests are needed](#) because there are many examples of people being inadvertently harmed when treatment decisions have not taken account of reliable evidence.

The [principles of fair tests](#) have been evolving over at least a millennium - and they continue to evolve today.

[Comparisons](#) are essential to address [genuine uncertainties](#) about treatment effects. Fair treatment comparisons must avoid [biases](#), whether from [differences between the people compared](#) or [differences in the way treatment outcomes are assessed](#). Reliable identification of [unexpected effects of treatments](#) poses particular challenges.

[Interpreting unbiased comparisons](#) is often not straightforward. [Differences between treatments intended and treatments received](#) can mean that real effects are overlooked. The [play of chance](#) can be misleading too.

Fair tests of treatments must take account of all the relevant evidence. Preparing systematic reviews of all the relevant evidence entails minimising the impact of [biased reporting](#) and [biased selection from the available evidence](#). A statistical process called [meta-analysis](#) may help avoid being misled by the [play of chance](#) in systematic reviews.

[Up-to-date, systematic reviews](#) of all relevant, reliable evidence are needed for fair tests of treatments in health care. Even with up-to-date systematic reviews, however, it's important to be on the lookout for biases and 'spin'. These can result in separate reviews, although supposedly addressing the same question, reaching conflicting conclusions.

In summary *The James Lind Library* contains the following essays:

[Why fair tests are needed](#)

[Why comparisons are essential](#)

[Why comparisons must address genuine uncertainties](#)

[Avoiding biased comparisons](#)

[Differences between the people compared](#)

[Differences in the way treatment outcomes are assessed](#)

[Interpreting unbiased comparisons](#)

[Differences between intended treatments and treatments actually received](#)

[Taking account of the play of chance](#)

[Identifying unanticipated effects of treatments](#)

Systematic reviews of all the relevant evidence

Dealing with biased reporting of the available evidence

Avoiding biased selection from the available evidence

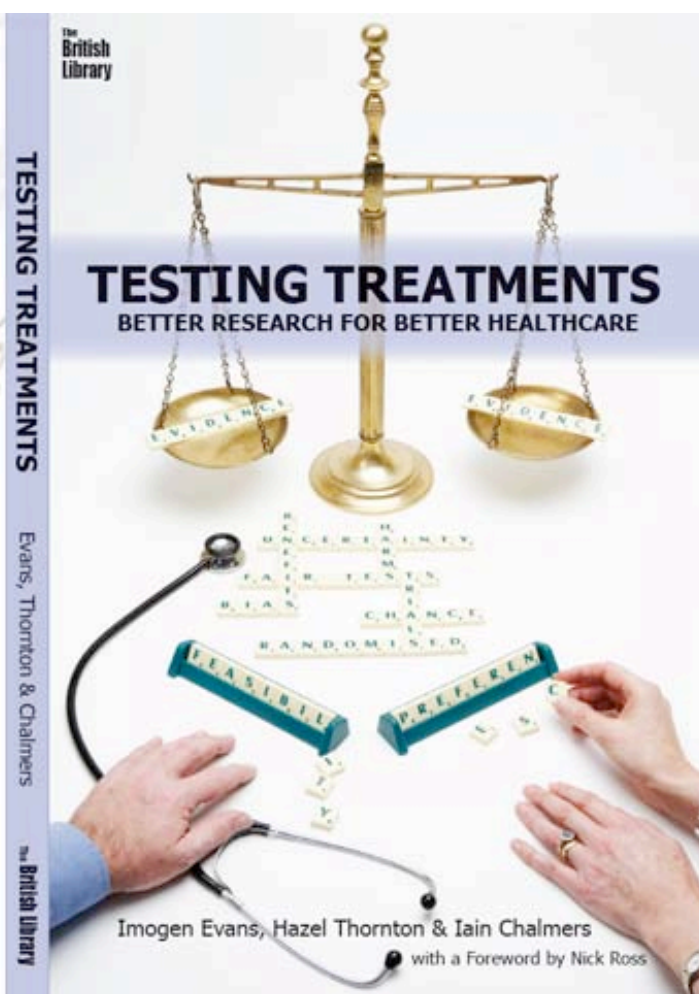
Reducing the play of chance using meta-analysis

Up-to-date, systematic reviews of all relevant, reliable evidence

These explanatory essays draw on a wealth of illustrative material in the *James Lind Library*. This can be accessed by clicking on the underlined links or images in the essays.

The text in these essays may be copied and used for non-commercial purposes on condition that explicit acknowledgement is made to '**The James Lind Library (www.jameslindlibrary.org)**'.

The material in the essays has also been incorporated in Evans, Thornton and Chalmers, '**Testing Treatments: better research for better health care**' - a 100-page book published in 2006 by the British Library. ISBN 0-7123-4909-X.



[Home](#)

[Contents](#)

[Comments welcome](#)

[Home](#)[Contents](#)jameslindlibrary.org

Why fair tests are needed

Why do we need fair tests of treatments in health care? Have not doctors, for centuries, 'done their best' for their patients? Sadly, there are many examples of doctors and other health professionals harming their patients because treatment decisions were not informed by what we consider now to be reliable evidence about the effects of treatments. With hindsight, health professionals in most if not all spheres of health care have harmed their patients inadvertently, sometimes on a very wide scale ([click here for examples](#)). Indeed, patients themselves have sometimes harmed other patients when, on the basis of untested theories and limited personal experiences, they have encouraged the use of treatments that have turned out to be harmful. The question is not whether we must blame these people, but whether the harmful effects of inadequately tested treatments can be reduced. They can, to a great extent.

[Records dealing with principles of testing](#)

Acknowledging that treatments can sometimes do more harm than good is a prerequisite for reducing unintended harm (Gregory 1772; [Haygarth 1800](#); Fordyce 1802; [Behring 1893](#)). We then need to be more ready to admit uncertainties about treatment effects, and to promote fair tests of treatments to reduce uncertainties.

Why theories about the effects treatments must be tested in practice

People have often been harmed because treatments have been based only on theories about how disease should be treated, without testing the theories in practice. For example, for centuries people believed the theory that illnesses were caused by 'humoral imbalances'. So patients were bled and purged, made to vomit and take snuff, in the belief that this would end the supposed imbalances. As long ago as the 17th century, a Flemish doctor was impertinent enough to challenge the medical authorities of the time to assess the validity of their theories in a fair test of the results of their unpleasant treatments ([Van Helmont 1662](#)).



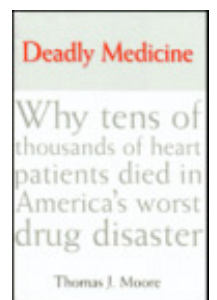
By the beginning of the 19th century, British military surgeons had begun to show the harmful effects of bloodletting (Robertson 1804; [Hamilton 1816](#)). A few decades later, the practice was also challenged by a Parisian physician ([Louis 1835](#)). Yet at the beginning of the 20th century orthodox practitioners in Boston, USA, who were not using bloodletting to treat pneumonia were still being judged negligent (Silverman 1980). Indeed, without citing supporting evidence, Sir William Osler, one of the most influential medical authorities in the world, advised his readers that: "during the last decades we have certainly bled too little. Pneumonia is one of the diseases in which a timely venesection [bleeding] may save life. To be of service it should be done early. In a full-blooded, healthy man with a high fever and bounding pulse the abstraction of from twenty to thirty ounces of blood is in every way beneficial" (Osler 1892).



Although the need to test the validity of theories in practice has been recognized for at least a millennium ([Ibn Hindu 10th-11th century](#)), this important principle is still too often ignored. For instance, based on untested theory, Benjamin Spock, the influential American child health expert, informed the readers of his best selling book '*Baby and Child Care*' that a disadvantage of babies sleeping on their backs was that, if they vomited, they would be more likely to choke. Dr Spock therefore advised his millions of readers to encourage babies to sleep on their tummies (Spock 1966). We now know that this advice, apparently rational in theory, led to the cot deaths of tens of thousands of infants (Gilbert et al. 2004).



The use of drugs to prevent heart rhythm abnormalities in people having heart attacks provides another example of the dangers of applying untested theory in practice. Because heart rhythm abnormalities are associated with an increased risk of early death after heart attack, the theory was that these drugs would reduce such early deaths. Just because a theory seems reasonable doesn't mean that it is necessarily right, however. Years after the drugs had been licensed and adopted in practice, it was discovered that they actually increase the risk of sudden death after heart attack. Indeed, it has been estimated that, at the peak of their use in the late 1980s, they may have been killing as many as 70,000 people every year in the United States alone (Moore 1995) – many more than the total number of Americans who died in the Vietnam War.



Misplaced confidence in the validity of theory as a guide to practice has also resulted in some



treatments being rejected inappropriately because researchers did not believe that they could work. Theories based on the results of animal research, for example, sometimes correctly predict the results of treatment tests in humans, but this is not always the case. Based on the results of experiments in rats, some researchers became convinced that there was no point in giving clot-dissolving drugs to patients who had experienced heart attacks more than six hours previously. Had not such patients participated in some of the fair tests of these drugs we would not know that they can benefit from treatment (Fibrinolytic Therapy Trialists' Collaborative Group 1994).

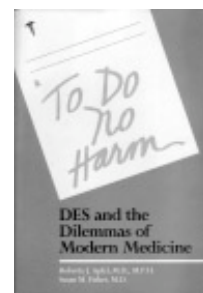


Observations in clinical practice or in laboratory and animal research may suggest that particular treatments will or will not benefit patients; but as these and many other examples make clear, it is essential to use fair tests to find out whether, in practice, these treatments do more good than harm, or vice versa.

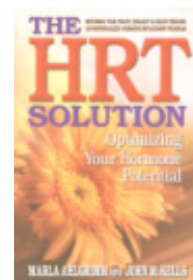
Why tests of medical treatments must be fair tests

Failure to test theories about treatments in practice is not the only preventable cause of treatment tragedies. These have also occurred because the tests used to assess the effects of treatments have been unreliable and misleading. Fair tests entail taking steps to reduce the likelihood that we will be misled by the effects of [biases](#) and the [play of chance](#).

For example, theory suggested that giving a synthetic sex hormone, diethylstilboestrol (DES), to pregnant women who had previously had miscarriages and stillbirths would increase the likelihood of a successful outcome of later pregnancies. Tests done at the time in which [biases](#) had not been adequately controlled suggested that the theory was correct, and that the drug reduced miscarriages and stillbirths. Although fair tests had suggested that DES was useless, the unreliable evidence, together with aggressive marketing, led to DES being prescribed to millions of pregnant women over the next few decades. The consequences were disastrous: some of the daughters of women who had been prescribed DES developed cancers of the vagina, and other children had other health problems, including malformations of their reproductive organs and infertility (Apfel and Fisher 1984).



Problems resulting from inadequate tests of treatments continue to occur. Again, as a result of unreliable evidence and aggressive marketing, millions of women were persuaded to use hormone replacement therapy (HRT), not only because it could reduce unpleasant menopausal symptoms, but also because it was claimed that it would reduce their chances of having heart attacks and strokes. When these claims were assessed in fair tests the results showed that, far from reducing the risks of heart attacks and strokes, HRT increases the risks of these life-threatening conditions, as well as having other undesirable effects (McPherson 2004).



These examples of the need for fair tests of treatments are a few of many hundreds that illustrate how treatments can do more harm than good. Improved general knowledge about fair tests of treatments is needed so that - laced with a healthy dose of scepticism - we can all assess claims about the effects of treatments more critically. That way, we will all become more able to judge which treatments are likely to do more good than harm.

The principles of fair tests of treatments have been evolving for centuries - and they continue to evolve. If you would like to see some examples, [click here](#).

References

Apfel RJ, Fisher SM (1984). To do no harm: DES and the dilemmas of modern medicine. New Haven, Ct: Yale University Press.

Behring, Boer, Kossel H (1893). Zur Behandlung diphtheriekranker Menschen mit Diphtherieheilserum. Deutsche Medicinische Wochenschrift 17:389-393.

Fibrinolytic Therapy Trialists' Collaborative Group (1994). Indications for fibrinolytic therapy in suspected acute myocardial infarction: collaborative overview of early mortality and major morbidity results from all randomised trials of more than 1000 patients. Lancet 1994;343:311-322.

Fordyce G (1802). A second dissertation on fever. London: J Johnson

Gilbert R, Salanti G, Harden M, See S (2005). Infant sleeping position and the sudden infant death syndrome:

systematic review of observational studies and historical review of recommendations from 1940 to 2002. *International Journal of Epidemiology* 34:874-87.

Gregory J (1772). *Lectures on the duties and qualifications of a physician*. London: Strahan and Cadell.

Hamilton AL (1816). *Dissertatio Medica Inauguralis De Synocho Castrensi (Inaugural medical dissertation on camp fever)*. Edinburgh: J Ballantyne.

Haygarth J (1800). *Of the imagination, as a cause and as a cure of disorders of the body: exemplified by fictitious tractors, and epidemical convulsions*. Bath: R. Crutwell.

Ibn Hindu (10th-11th century CE; 4th-5th century AH). *Miftah al-tibb wa-minhaj al-tullab [The key to the science of medicine and the students' guide]*.

Louis PCA (1835). *Recherches sur les effets de la saignée dans quelques maladies inflammatoires et sur l'action de l'émétique et des vésicatoires dans la pneumonie*. Paris: Librairie de l'Académie royale de médecine.

McPherson K (2004). Where are we now with hormone replacement therapy? *BMJ* 328:357-358.

Moore TJ (1995). *Deadly Medicine*. New York: Simon and Schuster.

Osler W (1892). *Principles and Practice of Medicine*. London: Appleton, p 530.

Robertson R (1804). *Observations on the diseases incident to seamen*, 2nd edn. Vol. 1, London: for the author.

Silverman W (1980). In: Chalmers I, McIlwaine G (eds). *Perinatal Audit and Surveillance*. London: Royal College of Obstetricians and Gynaecologists, 1980:110.

Spock B (1966). *Baby and Child Care*. 165th printing. New York: Pocket Books, pp 163-164.

van Helmont JB (1662). *Oriatrike, or physick refined: the common errors therein refuted and the whole are reformed and rectified [translated by J Chandler]*. Lodowick-Loyd: London, p 526.

[Home](#)

[Contents](#)

[Comments welcome](#)

[Home](#)[Contents](#)jameslindlibrary.org

Why comparisons are essential

Is a treatment better than nature and time?

Patients and healthcare professionals hope that treatments will be helpful. These optimistic expectations can have a very positive effect on everybody's satisfaction with health care, as the British doctor Richard Asher noted in one of his essays for doctors:

"If you can believe fervently in your treatment, even though controlled tests show that it is quite useless, then your results are much better, your patients are much better, and your income is much better too. I believe this accounts for the remarkable success of some of the less gifted, but more credulous members of our profession, and also for the violent dislike of statistics and controlled tests which fashionable and successful doctors are accustomed to display." (Asher 1972)

People often recover from illness without any specific treatment: nature and time are great healers. As Oliver Wendell Holmes suggested in the 19th century when there were very few useful treatments ([Holmes 1861](#)), "I firmly believe that if the whole materia medica, as now used, could be sunk to the bottom of the sea, it would be all the better for mankind - and all the worse for the fishes."

The progress and outcome of illness if left untreated must obviously be taken into account when treatments are being tested: treatment may improve or it may worsen outcomes. Writers over the centuries have drawn attention to the need to be sceptical about claims that the effects of treatments can improve on the effects of nature: or as one wag has put it, "If you leave a dose of 'flu to nature, you'll probably get over it in a week; but if you go to the doctor, you'll recover in a mere seven days."

In the knowledge that much illness is self-limiting, doctors sometimes prescribe inert treatments in the hope that their patients will derive psychological benefit while nature takes its course - the so-called placebo effect. Patients who believe that a treatment will help to relieve their symptoms - even though the treatment, in fact, has no physical effects - may well feel better.

Doctors have recognized the importance of using placebos for centuries. For example, William Cullen referred to his use of a placebo as long ago as 1772 (Cullen 1772), and references to placebos increased during the 19th century (Cummings 1805; [Ministry of Internal Affairs 1832](#); [Forbes 1846](#)). Because Austin Flint believed that orthodox drug treatment was usurping the credit due to 'nature', he gave thirteen patients with rheumatism a 'placeboic remedy' consisting of a highly dilute extract of the bark of the quassia tree. The result was that "the favourable progress of the cases was such as to secure for the remedy generally the entire confidence of the patients" ([Flint 1863](#)). At Guy's Hospital in London, William Withey Gull came to similar conclusions after treating 21 rheumatic fever patients "for the most part with mint water" ([Sutton 1865](#)). At the beginning of the 20th century William Rivers discussed psychologically-mediated effects of treatments in detail ([Rivers 1908](#)).



The need for comparisons

Just as the healing power of nature and placebo have been recognized for centuries, so also has the need for comparisons to assess effects treatments over and above natural and psychologically-mediated effects. Sometimes treatment comparisons are made in people's minds: they have an impression that they or others are responding differently to a new treatment compared with previous responses to treatments. For example, Ambroise Paré, a French military surgeon, concluded that treatment of battle wounds with boiling oil (as was common practice) was likely to be harmful. He concluded this when the supply of oil ran out and his patients recovered more quickly than usual ([Paré 1575](#)).



Impressions like this need to be followed up by formal investigations, perhaps initially by analysis of healthcare records. Such impressions may then lead to carefully conducted comparisons. The danger arises when impressions alone are used as a guide to treatment recommendations and decisions.

Dramatic effects and moderate effects of treatments

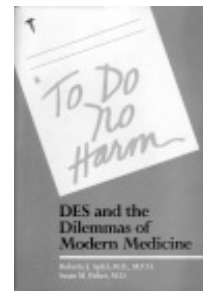
Treatment comparisons based on impressions, or relatively restricted analyses, only provide reliable information

in the rare circumstances when treatment effects are dramatic ([click here to list relevant records](#)). Examples include opium for pain relief ([Tibi 2005](#)), hygiene for preventing tetanus (lockjaw) in newborn babies (Schleisner 1849), chloroform for anaesthesia, insulin for diabetes ([Banting et al. 1922](#)), liver diet for pernicious anaemia ([Minot and Murphy 1926](#)), sulpha drugs for infection after childbirth ([Colebrook and Purdie 1937](#)), streptomycin for tuberculous meningitis ([MRC 1948](#)), adrenaline for life-threatening allergic reactions (McLean-Tooke et al. 2003), and genetically-designed drugs for some rare forms of leukaemia (Druker et al. 2001). Most medical treatments don't have such dramatic effects as these, however, and unless care is taken to avoid biased comparisons, dangerously mistaken conclusions about the effects of treatment may result.

Comparing treatments given today with treatments given in the past

It was partly because of reliance on biased comparisons with past experience that doctors and women believed that the drug diethylstilboestrol (DES) would reduce the risk of miscarriages and stillbirths. There was never any evidence from fair (unbiased) tests that DES could do this, and it was later shown that it caused cancer in the daughters of some of the pregnant women for whom it had been prescribed. A treatment that has not been reliably shown to be useful should not be promoted.

Comparing treatments given today with treatments given in the past only rarely provides a secure basis for a fair test ([Behring et al. 1893](#); [Roux et al. 1894](#)), because relevant factors other than the treatments themselves change over time. For example, miscarriages and stillbirths are more common in first pregnancies than in later pregnancies. Comparing the frequency of miscarriages and stillbirths in later pregnancies in which DES was prescribed with the outcome of first pregnancies in which the drug wasn't used is thus likely to be a seriously misleading basis for assessing its effects. If possible, therefore, comparisons should involve giving different treatments at more or less the same time.



Comparing treatments in crossover tests in individual patients

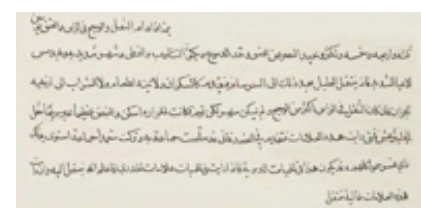
Sometimes giving different treatments at more or less the same time may involve giving a patient different treatments one after the other – a so-called crossover test ([Martini 1932](#); [click here to list relevant records](#)). An early example of a crossover test was reported in 1786 by Dr Caleb Parry in Bath, England. He wanted to find out whether there was any reason to pay for expensive, imported Turkish rhubarb as a purgative for treating his patients, rather than using rhubarb grown locally in England. So he 'crossed-over' the type of rhubarb given to each individual patient at different times and then compared the symptoms each patient experienced while eating each type of rhubarb ([Parry 1786](#)).



Although treatment comparisons within individual patients have their place, there are many circumstances in which they can't be applied. For example, it is usually impossible to compare different surgical operations in this way, or treatments given for progressive conditions.

Comparing groups of patients given different treatments concurrently

Treatments are usually tested by comparing groups of people who receive different treatments. A comparison of two treatments will be unfair if relatively well people have received one treatment and relatively ill people have received the other, so the experiences of similar groups of people who receive different treatments over the same period of time must be compared. Al-Razi recognized this more than a thousand years ago when, wishing to reach a conclusion about how to treat patients with signs of early meningitis, he treated one group of patients and intentionally withheld treatment from a comparison group ([al-Razi 9th century](#)).



Comparisons with nature or with other treatments are needed for fair tests of treatments, and if these comparisons are to be fair, they must [address genuine uncertainties](#), avoid [biases](#) and the [play of chance](#), and [be interpreted carefully](#).

References

al-Razi (10th century CE; 4th Century AH). *Kitab al-Hawi fi al-tibb* [The comprehensive book of medicine].

Asher R (1972). *Talking sense*. London: Pitman Medical.

Banting FG, Best CH, Collip JB, Campbell WR, Fletcher AA (1922). Pancreatic extracts in the treatment of diabetes mellitus. *Canadian Medical Association Journal* 12:141-146.

Behring, Boer, Kossel H (1893). Zur Behandlung diphtheriekranker Menschen mit Diphtherieheilserum. Deutsche Medicinische Wochenschrift 17:389-393.

Colebrook L, Purdie AW (1937). Treatment of 106 cases of puerperal fever by sulphanilamide. Lancet 2:1237-1242 & 1291-1294.

Cullen W (1772). Clinical lectures. Edinburgh, February-April, 218-9.

Cummings R (1805). Medical and Physical Journal, page 6.

Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, Lydon NB, Kantarjian H, Capdeville R, Ohno-Jones S, Sawyers CL (2001). Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. New England Journal of Medicine 344:1031-1037.

Flint A (1863). A contribution toward the natural history of articular rheumatism; consisting of a report of thirteen cases treated solely with palliative measures. American Journal of the Medical Sciences 46:17-36.

Forbes J (1846). Homeopathy, allopathy and 'young physic.' British and Foreign Medical Review 21:225-265.

Holmes OW (1861). Currents and countercurrents in medical science. In: Works, 1861 Vol ix, p 185.

Martini P (1932). Methodenlehre der Therapeutischen Untersuchung. Berlin: Springer.

McLean-Tooke APC, Bethune CA, Fay AC, Spickett GP (2003). Adrenaline in the treatment of anaphylaxis: what is the evidence? BMJ 327:1332-1335.

Medical Research Council (1948). Streptomycin treatment of tuberculous meningitis. Lancet 1:582-596.

Ministry of Internal Affairs (1823). [Conclusion of the Medical Council regarding homeopathic treatment]. Zhurnal Ministerstva Vnutrennih del, 3:49-63.

Minot GR, Murphy WP (1926). Treatment of pernicious anaemia by a special diet. JAMA 87:470-476.

Paré A (1575). Les oeuvres de M. Ambroise Paré conseiller, et premier chirurgien du Roy avec les figures & portraits tant de l'Anatomie que des instruments de Chirurgie, & de plusieurs Monstres. Paris: Gabriel Buon.

Parry CH (1786). Experiments relative to the medical effects of Turkey Rhubarb, and of the English Rhubarbs, No. I and No. II made on patients of the Pauper Charity. Letters and Papers of the Bath Society III: 431-453.

Rivers WHR (1908). The influence of alcohol and other drugs on fatigue. London:Edward Arnold.

Roux E, Martin L, Chaillou A (1894). Trois cent cas de diphthérie traité par le serum antidiphthérique. Annales de l'Institut Pasteur 8:640-661.

Schleisner PA (1849). Island fra et lægevidenskabeligt Symspunkt. København: Boghandler Iversen.

Sutton HG (1865). Cases of rheumatic fever, treated for the most part by mint water. Collected from the clinical books of Dr Gull, with some remarks on the natural history of that disease. Guy's Hospital Report 11:392-428.

Tibi S (2005). The medicinal use of opium in ninth-century Baghdad. Leiden: Brill.

[Home](#)

[Contents](#)

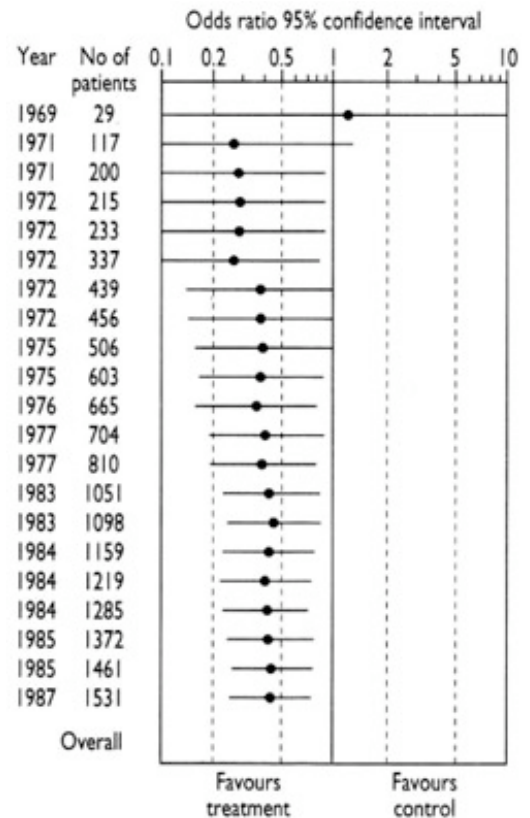
[Comments welcome](#)

[Home](#)[Contents](#)jameslindlibrary.org

Why comparisons must address genuine uncertainties

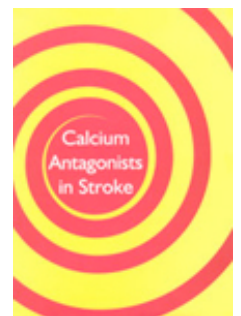
A good deal of research is done even when there are no genuine uncertainties. Researchers who fail to conduct systematic reviews of past tests of treatments before embarking on further studies sometimes don't recognise (or choose to ignore the fact) that uncertainties about treatment effects have already been convincingly addressed. This means that people participating in research are sometimes denied treatment that could help them, or given treatment likely to harm them.

The diagram that accompanies this and the following paragraph shows the accumulation of evidence from fair tests done to assess whether antibiotics (compared with inactive placebos) reduce the risk of post-operative death in people having bowel surgery (Lau et al. 1995). The first fair test was reported in 1969. The results of this small study left uncertainty about whether antibiotics were useful – the horizontal line representing the results spans the vertical line that separates favourable from unfavourable effects of antibiotics. Quite properly, this uncertainty was addressed in further tests in the early 1970s.



As the evidence accumulated, however, it became clear by the mid-1970s that antibiotics reduce the risk of death after surgery (the horizontal line falls clearly on the side of the vertical line favouring treatment). Yet researchers continued to do studies through to the late 1980s. Half the patients who received placebos in these later studies were thus denied a form of care which had been shown to reduce their risk of dying after their operations. How could this have happened? It was probably because researchers continued to embark on research without reviewing existing evidence systematically. This behaviour remains all too common in the research community, partly because some of the incentives in the world of research – commercial and academic – do not put the interests of patients first (Chalmers 2000).

Patients and participants in research can also suffer because researchers have not systematically reviewed relevant evidence from animal research before beginning to test treatments in humans. A Dutch team reviewed the experience of over 7000 patients who had participated in tests of a new calcium-blocking drug given to people experiencing a stroke. They found no evidence to support its increasing use in practice (Horn and Limburg 2001). This made them wonder about the quality and findings of the animal research that had led to the research on patients. Their review of the animal studies revealed that these had never suggested that the drug would be useful in humans (Horn et al. 2001).



The most common reason that research does not address genuine uncertainties is that researchers simply have not been sufficiently disciplined to review relevant existing evidence systematically before embarking on new studies. Sometimes there are more sinister reasons, however. Researchers may be aware of existing evidence, but they want to design studies to ensure that their own research will yield favourable results for particular treatments. Usually, but not always, this is for commercial reasons (Djulbegovic et al. 2000; Sackett and Oxman

2003). These studies are deliberately designed to be unfair tests of treatments. This can be done by withholding a comparison treatment known to help patients (as in the example given above), or giving comparison treatments in inappropriately low doses (so that they don't work so well), or in inappropriately high doses (so that they have more unwanted side effects) ([see commentary by Mann and Djulbegovic](#)).

It is probably surprising to many readers of this essay that the research ethics committees established to ensure that research is ethical have done so little to influence this research malpractice. These committees have let down the people they should have been protecting because they have not required researchers and sponsors seeking approval for new tests to have reviewed existing evidence systematically (Savulescu et al. 1996; Chalmers 2002). The failure of research ethics committees to protect patients and the public efficiently emphasises the importance of improving general knowledge about the characteristics of fair tests of medical treatments.

References

Chalmers I. Current Controlled Trials: an opportunity to help improve the quality of clinical research. *Current Controlled Trials in Cardiovascular Medicine* 2000;1:3-8.
Available: <http://cvm.controlled-trials.com/content/1/1/3>

Chalmers I (2002). Lessons for research ethics committees. *Lancet* 359:174.

Djulbegovic B, Lacey M, Cantor A, Fields KK, Bennett CL, Adams JR, Kuderer NM, Lyman GH (2000). The uncertainty principle and industry-sponsored research. *Lancet* 356:635-638.

Horn J, Limburg M (2001). Calcium antagonists for acute ischemic stroke (Cochrane Review). In: *The Cochrane Library*, Issue 3, Oxford: Update Software.

Horn J, de Haan RJ, Vermeulen M, Luiten PGM, Limburg M (2001). Nimodipine in animal model experiments of focal cerebral ischaemia: a systematic review. *Stroke* 32:2433-38.

Lau J, Schmid CH, Chalmers TC (1995). Cumulative meta-analysis of clinical trials builds evidence for exemplary clinical practice. *Journal of Clinical Epidemiology* 48:45-57.

Mann H, Djulbegovic B. Why comparisons must address genuine uncertainties. James Lind Library (www.jameslindlibrary.org).

Sackett DL, Oxman AD (2003). HARLOT plc: an amalgamation of the world's two oldest professions. *BMJ* 2003;327:1442-1445.

Savulescu J, Chalmers I, Blunt J (1996). Are research ethics committees behaving unethically? Some suggestions for improving performance and accountability. *BMJ* 313:1390-1393.

[Home](#)

[Contents](#)

[Comments welcome](#)

[Home](#)[Contents](#)jameslindlibrary.org

Avoiding biased comparisons

Sometimes treatments have dramatic effects ([click here to list relevant records](#)). These may be unintended and specific, for example, when a person has an allergic reaction to an antibiotic drug. Treatments can also have striking beneficial effects, like adrenaline for life-threatening allergic reactions (McLean-Tooke et al. 2003). Such striking effects are rare, however. Usually, treatment effects are more modest, but nevertheless well worth knowing about.

[Records dealing with dramatic effects](#)

For example, aspirin doesn't prevent all premature deaths after a heart attack, but it does reduce the likelihood of death by about twenty per cent, which is important in such a common condition. If these moderate but important effects of most treatments are to be detected reliably, care must be taken to ensure that biased comparisons don't lead us to believe that treatments are useful when they are useless or harmful, or useless when they can actually be helpful.

Biases in tests of treatment are those influences and factors that can lead to conclusions about treatment effects that are systematically different from the truth. Although many kinds of biases can distort the results of health research (Sackett 1979), we have concentrated in *The James Lind Library* on those biases that must be minimised in fair tests of treatments. These are:

- [biases due to differences in people compared](#);
- [biases due to differences in the way treatment outcomes are assessed](#);
- [biased reporting of the available evidence](#); and
- [biased selection from the available evidence](#).

Ignoring these biases (or sometimes unscrupulously taking advantage of them), may lead people to believe that a new treatment is better than an existing treatment, when it is not. This could result from basing conclusions on:

- studies that compare the progress of relatively well people given a new treatment with the progress of relatively ill people given a standard treatment ([allocation bias](#)).
- studies in which assessments of treatment outcomes are likely to be biased in favour of a new treatment, for example, by comparing the opinions of people who know that they have used an expensive new treatment with the opinions of those who may be disappointed that they were continuing to use an unexciting standard treatment ([observer or measurement bias](#)).
- studies that show a new treatment in a favourable light, and not those that suggest that it may be harmful, which are often not reported ([reporting bias](#)).
- biased selection from and interpretation of the available evidence to support a particular viewpoint ([reviewer bias](#)).

Usually, the unfair tests of treatment resulting from these biases are not recognised for what they are. However, people with vested interests sometimes exploit these biases so that treatments are presented as if they are better than they really are (Sackett and Oxman 2003).

Whether biases are inadvertent or deliberate, the consequences are the same: unless tests of treatment are fair, some useless or harmful treatments will seem to be useful, while some useful treatments will seem useless or harmful.

References

McLean-Tooke APC, Bethune CA, Fay AC, Spickett GP (2003). Adrenaline in the treatment of anaphylaxis: what is the evidence? *BMJ* 327:1332-1335.

Sackett DL (1979). Bias in analytic research. *Journal of Chronic Diseases* 32:51-63.

Sackett DL, Oxman AD (2003). HARLOT plc: an amalgamation of the world's two oldest professions. *BMJ* 2003;327:1442-1445.

Comments welcome

[Home](#)[Contents](#)jameslindlibrary.org

Avoiding biased comparisons

Differences between the people compared

Comparing different treatments given to groups of people

Treatment comparisons usually entail comparing the experiences of groups of people who have received different treatments. If these comparisons are to be fair, the composition of the groups must be similar

– so that like will be compared with like. If those who receive one treatment are more likely anyway to do well (or badly) than those receiving an alternative treatment, this allocation bias makes it impossible to be confident that outcomes reflect differential effects of the treatments, rather than the effects of nature and the passage of time.

[Records dealing with allocation bias](#)[Records dealing with crossover test](#)

The 18th century surgeon William Cheselden was aware of this problem of dissimilar groups when surgeons were comparing their respective mortality rates after operations to remove bladder stones. Cheselden pointed out that it was important to take account of the ages of the people treated by different surgeons. He drew attention to the fact that mortality rates varied with the patients' ages ([Cheselden 1740](#)) - older patients were more likely than younger patients to die. This meant that, if one wished to compare the frequency of deaths in groups of patients who had undergone different types of operation, one had to take account of differences in the ages of the patients in the comparison groups.



Comparing the experiences and outcomes of patients who happened to have received different treatments in the past is still used today as a way of trying to assess the effects of treatments. The challenge is to know whether the comparison groups were sufficiently alike before receiving treatment. This is illustrated by attempts to assess the effects of hormone replacement therapy (HRT) by comparing the illness experiences of women who had used HRT with those of other women who had not used it. As subsequent analysis of fair tests of HRT showed, trying to assess the effects of treatments in retrospect in this way can sometimes be dangerously misleading (McPherson 2004).

It is rarely possible to be completely confident that comparison groups selected from people who have been given one treatment in the past are comparable in all the respects that matter with people who have more recently received an alternative treatment. This is the case even if some information about the patients who have received different treatments is available (such as their ages, or their past history of illness). Other information that may be of great importance (such as the likelihood of spontaneous recovery) may simply not be available.

A better approach is to plan the treatment comparisons before starting treatment. For example, before beginning his comparison of six treatments for scurvy on board *HMS Salisbury* in 1747, James Lind took care to select patients who were at a similar stage of this often fatal disease. He also ensured that they had the same basic diet and were accommodated in similar conditions. These were factors, other than treatment, that might have influenced their likelihood of recovering ([Lind 1753](#)). Comparable efforts must be taken today to try to ensure that treatment comparison groups will be composed of similar people.



Unbiased assembly of treatment comparison groups using alternation or randomisation

Although Lind took care to ensure that the sailors in his six comparison groups were alike, he didn't describe how he decided which sailors would receive which of the six treatments. There is only one way to ensure that treatment comparison groups are similar in all the ways that matter, known and unknown. This is by using some form of chance process to assemble treatment comparison groups to avoid biased selection for different treatments before starting treatment.

One hundred years after Lind, an army doctor, Graham Balfour, illustrated how this could be done in a test to see whether belladonna prevented scarlet fever in children. In the military orphanage for which he had responsibility, he used alternation - "to prevent the imputation of selection" - to decide which boys would receive and which would not receive belladonna ([Balfour 1854](#)). Alternation is one of several unbiased methods for assembling similar treatment comparison groups before giving the treatments. During the first half of the 20th century, there are many examples of treatment comparison groups being assembled using



alternation or rotation (for example [Hamilton 1816](#); [MRC 1944](#)), or by drawing lots ([Colebrook 1929](#)) – for example, using dice ([Doull et al. 1931](#)), coloured beads ([Theobald 1937](#)), or random sampling numbers ([Bell 1941](#); [MRC 1948](#); [MRC 1950](#); [MRC 1951](#)). This 'random allocation' is the sole, but crucially important, feature of the category of fair tests referred to as 'randomized'.



As illustrated in the essay available by clicking [here](#), casting or drawing lots is a time-honoured way of making fair decisions. These methods help to ensure that comparison groups are composed of people who are similar, not just in respect of known and measured factors of importance, like age, but also unmeasured factors that may influence recovery from illness, such as diet, occupation, and anxiety. If you would like to see how random allocation generates similar groups of people ([click here for a demonstration](#)).

As experience of using alternation and random allocation for unbiased assembly of groups of patients for comparing different treatments became more widespread, it became clear that strict adherence to allocation schedules was required to avoid biased creation of treatment comparison groups ([MRC 1934](#)). The risk of biased allocation can be abolished if treatment allocation schedules are concealed from those making decisions about participation in treatment comparisons – in brief, to prevent them cheating, and thus biasing the comparisons ([MRC 1944](#); [MRC 1948](#); [MRC 1950](#); [MRC 1951](#)).



Avoiding biased losses from treatment comparison groups

After taking the trouble to ensure that treatment comparison groups are assembled in ways that ensure that like will be compared with like, it is important to avoid bias being introduced as a result of selective withdrawal of patients from the comparison groups. As far as possible, group similarity should be maintained by ensuring that all the people allocated to the treatment comparison groups are followed up and included in the main analysis of the test results – a so-called 'intention-to-treat' analysis ([Bell 1941](#)).

Failure to do this can result in unfair tests of treatments. Take, for example, a test to assess whether an operation to unblock a blood vessel supplying the brain can reduce strokes in people experiencing dizzy spells as a result of the narrowed vessel. If the frequency of strokes was recorded only among patients who had survived the immediate effects of the operation, the test would miss the important fact that the operation itself can cause stroke and death. This would be an unfair test of the effects of the operation. People want unbiased information about the overall effects of treatments. This means taking account of the experiences of all the patients who are assigned to treatment comparison groups, without exceptions.

References

- Balfour TG (1854). Quoted in West C. Lectures on the Diseases of Infancy and Childhood. London, Longman, Brown, Green and Longmans, p 600.
- Bell JA (1941). Pertussis prophylaxis with two doses of alum-precipitated vaccine. Public Health Reports 56:1535-1546.
- Cheselden W (1740). The anatomy of the human body. 5th edition. London: William Bowyer.
- Colebrook D (1929). Irradiation and health. Medical Research Council Special Report Series No.131.
- Doull JA, Hardy M, Clark JH, Herman NB (1931). The effect of irradiation with ultra-violet light on the frequency of attacks of upper respiratory disease (common colds). American Journal of Hygiene 13:460-77.
- Hamilton AL (1816). Dissertatio Medica Inauguralis De Synocho Castrensi (Inaugural medical dissertation on camp fever). Edinburgh: J Ballantyne.
- Lind J (1753). A treatise of the scurvy. In three parts. Containing an inquiry into the nature, causes and cure, of that disease. Together with a critical and chronological view of what has been published on the subject. Edinburgh: Printed by Sands, Murray and Cochran for A Kincaid and A Donaldson.
- McPherson K (2004). Where are we now with hormone replacement therapy? BMJ 328:357-358.
- Medical Research Council Therapeutic Trials Committee (1934). The serum treatment of lobar pneumonia. BMJ 1:241-245.

Medical Research Council (1944). Clinical trial of patulin in the common cold. *Lancet* 2:373-5.

Medical Research Council (1948). Streptomycin treatment of pulmonary tuberculosis: a Medical Research Council investigation. *BMJ* 2:769-782.

Medical Research Council (1950). Clinical trials of antihistaminic drugs in the prevention and treatment of the common cold. *BMJ* 2:425-431.

Medical Research Council (1951). The prevention of whooping-cough by vaccination. *BMJ* 1:1463-1471

Parry CH (1786). Experiments relative to the medical effects of Turkey Rhubarb, and of the English Rhubarbs, No. I and No. II made on patients of the Pauper Charity. *Letters and Papers of the Bath Society* III:407-422.

Silverman WA, Chalmers I. Casting and drawing lots. The James Lind Library (www.jameslindlibrary.org).

Theobald GW (1937). Effect of calcium and vitamin A and D on incidence of pregnancy toxæmia. *Lancet* 2:1397-1399.

[Home](#)

[Contents](#)

[Comments welcome](#)

[Home](#)[Contents](#)jameslindlibrary.org

Avoiding biased comparisons

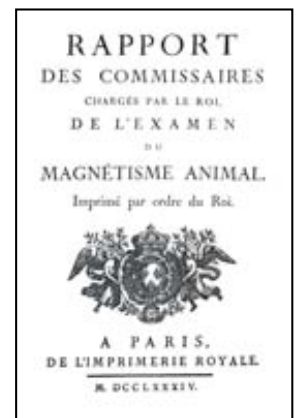
Differences in the way treatment outcomes are assessed

Using blinding to reduce bias in assessing treatment outcomes

[Records dealing with control of observer bias](#)

For some outcomes by which treatment is assessed – survival, for example - biased assessment is very unlikely because there is little room for opinion. This was the case in some of the 18th century tests of surgical procedures, where survival was the main measure of treatment success or failure. The assessment of most other outcomes, however, either always involves subjectivity (as with patients' symptoms), or may involve it. All of us are aware that we can sometimes mislead ourselves into thinking that we have detected, seen or experienced something that actually isn't there. The biases that lead to these misperceptions are termed observer biases. They cause a particular problem when people believe that they already 'know' the effect of a treatment, or when they may have particular reasons for preferring one of the treatments being compared. When measures are not taken to reduce biased outcome assessments in treatment comparisons, treatment effects tend to be overestimated (Schulz et al. 1995). The greater the element of subjectivity in assessing outcomes, the greater the need to reduce these observer biases to ensure fair tests of treatments.

In these common circumstances, 'blinding' of patients and doctors is a desirable element of fair tests. The earliest blinded (masked) assessment of a treatment appears to have been performed by a commission of inquiry appointed by Louis XVI in 1784 to investigate Anton Mesmer's claims of the effects of 'animal magnetism' ([Commission Royale 1784](#)). The commission assessed whether the purported effects of this new healing method were due to any 'real' force, or due to the 'illusions of the mind'. Blindfolded people were told that they were receiving or not receiving magnetism when in fact, at times, the reverse was happening. The people being studied felt the effects of 'animal magnetism' only when they were told they were receiving the treatment, but not otherwise (Kaptchuk 1998; Schulz et al. 2002).



Using placebos to achieve blinding

A few years after the tests of the effects of animal magnetism, John Haygarth conducted a single blind experiment using a sham device (a placebo) to achieve blinding ([Haygarth 1800](#)). The cartoon that accompanies this paragraph shows a doctor treating a wealthy client with a device patented and marketed by Elisha Perkins. Perkins claimed that his 'tractors' – small metal rods - cured a variety of ailments through 'electrophysical force'. In a pamphlet entitled '*Of the imagination as a cause and as a cure of disorders of the body: exemplified by fictitious tractors*', John Haygarth reported how he put Perkins' claims to a fair test. In a series of patients who were unaware of the details of his evaluation, he used a cross-over study to compare the patented, metal tractors (which were meant to work through 'electrophysical force') with identical sham 'tractors' that he had made out of wood ('placebo tractors'). He was unable to detect any benefit of the metal tractors ([Haygarth 1800](#)).



John Haygarth's fair test of Perkins' tractors is an early example of the use of placebos to achieve blinding to reduce biases in assessing the outcome of treatments. Placebos became a research tool in the debates on homeopathy, the nineteenth century's other major form of unconventional healing. Homeopaths often used blind assessment and placebo controls for their "provings", which tested the effects of their remedies on healthy volunteers ([Löhner 1835](#); Kaptchuk 1998). One of the most sophisticated placebo-controlled tests took place under the Milwaukee Academy of Medicine in 1879-1880. This trial was 'double-blind': both patients and experimenters were kept unaware as to whether the treatment was a genuine homeopathic remedy or a sugar pill ([Storke et al. 1880](#)).

It was not until much later that a more skeptical attitude in mainstream medicine led to a recognition that there was a need to adopt blinded assessment and placebos to assess the validity of its own claims. Inspired principally by pharmacologists, German researchers gradually adopted masked assessment. For example, in 1918, Adolf Bingel reported that he had tried to be "as objective as possible" when comparing two different treatments for diphtheria ([Bingel 1918](#)). He assessed whether he or his



colleagues could guess which patients had received which treatment: "I have not relied on my own judgment alone, but have sought the views of the assistant physicians of the diphtheria ward, without informing them about the nature of the serum under test. Their judgment was thus completely without prejudice. I am keen to see my observations checked independently, and most warmly recommend this 'blind' method for the purpose." ([Bingel 1918](#)). In fact, no difference was detected between the two treatments. A strong tradition of blind assessment developed in Germany, and this was codified by the clinical pharmacologist Paul Martini ([Martini 1932](#)).



Blind assessment in the modern anglophone world first began when pharmacologists were influenced by the German tradition, as well as by an indigenous 'quackbuster' movement that used masked assessment (Kaptchuk 1998). By the 1930's, they had taken the lead in using placebo controls in clinical experiments. For example, two of the UK Medical Research Council's earliest fair tests were of treatments for the common cold. It would have been very difficult to interpret their results had 'double blinding' not been used to prevent patients and doctors knowing which patients had received the new drugs and which had received placebos ([MRC 1944](#); [MRC 1950](#)). Harry Gold's strenuous advocacy of the importance of blinded assessment appears to have had a particularly important influence in the United States ([Conference on Therapy 1954](#)).

Blinding observers when it is impossible to blind patients and clinicians

Sometimes it is simply impossible to blind patients and doctors to the identity of the treatments being compared, for example, when surgical treatments are compared with drug treatments. Even in these circumstances, however, steps can be taken to reduce biased assessment of treatment outcomes. Independent observers can be kept unaware of which treatments have been received by which patients. For example, in the early 1940s a test compared patients with pulmonary tuberculosis receiving the then standard treatment - bed rest - with other patients who received, in addition, injections of the drug streptomycin. The researchers felt that it would be unethical to inject inactive placebos in patients allocated to bed rest alone simply to achieve 'blinding' of the patients and doctors treating them ([MRC 1948](#)), but they took alternative precautions to reduce biased assessment of outcomes. Although there was little danger of biased assessment of the principal outcome (survival), subjectivity could have biased the assessment of the chest X-rays. Accordingly, X-rays were assessed by doctors who were kept unaware of whether they were evaluating outcome in a patient who had been treated with streptomycin or one treated with bed rest alone.

Together with randomization, masked assessment, when possible using placebos, has now become one of the crucial methodological components of fair tests of treatments.

References

Bingel A (1918). Über Behandlung der Diphtherie mit gewöhnlichem Pferdeserum. *Deutsches Archiv für Klinische Medizin* 125:284-332.

Commission Royale (1784). *Rapport des commissaires chargés par le roi du magnetisme animal*. Paris: Imprimerie royale.

Conference on Therapy (1954). How to evaluate a new drug. *American Journal of Medicine* 17:722-727.

Haygarth J (1800). *Of the imagination, as a cause and as a cure of disorders of the body: exemplified by fictitious tractors, and epidemical convulsions*. Bath: R. Crutwell.

Kaptchuk TJ (1998). Intentional ignorance: a history of blind assessment and placebo controls in medicine. *Bulletin of the History of Medicine* 72:389-433.

Löhner G (1835), on behalf of a Society of truth-loving men. *Die Homoöopathischen Kochsalzversuche zu Nürnberg [The homeopathic salt trials in Nuremberg]*.

Martini P (1932). *Methodenlehre der Therapeutischen Untersuchung*. Berlin:Springer.

Medical Research Council (1944). Clinical trial of patulin in the common cold. *Lancet* 2:373-375.

Medical Research Council (1948). Streptomycin treatment of pulmonary tuberculosis: a Medical Research Council investigation. *BMJ* 2:769-782.

Medical Research Council (1950). Clinical trials of antihistaminic drugs in the prevention and treatment of the common cold. *BMJ* 2:425-431.

Schulz KF, Chalmers I, Hayes RJ, Altman DG (1995). Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 273:408-412.

Schulz KF, Chalmers I, Altman D (2002). The landscape and lexicon of blinding. *Annals of Internal Medicine* 136:254-259.

Storke EF, Martin R, Rosenkrans EM, Ford J, Schloemilch A, McDermott GC, Carlson OW (1880). Final report of the Milwaukee test of the thirtieth dilution. *Homeopathic Times: A Monthly Journal of Medicine, Surgery and the Collateral Sciences* 7:280-281.

[Home](#)

[Contents](#)

[Comments welcome](#)

[Home](#)[Contents](#)jameslindlibrary.org

Interpreting unbiased comparisons

A fair treatment comparison is one that [avoids biased comparisons](#). This entails taking steps to minimise [biases due to differences between the patients compared](#), and [biases due to differences in the way treatment outcomes are assessed](#).

Even if these biases have been avoided, however, interpreting unbiased comparisons is often not straightforward. For example, have any [differences between treatments intended and treatments received](#) been taken into account, and has account been taken of [the play of chance](#)?

Sometimes, a new study provides very strong evidence of the effects of a treatment. For example, tens of thousands of people participated in a remarkable study that showed that an aspirin tablet could substantially reduce the risk of death among people who are experiencing heart attacks. It is only very rarely, however, that a single study provides such strong evidence, so it's important when reading reports of most studies to ask whether the new evidence has been integrated in [systematic reviews of all other relevant evidence](#). If so, have steps been taken during that process of synthesis to minimise the impact of [biased reporting of the available evidence](#) and [biased selection from the available evidence](#)? Has the potential for [reducing the play of chance using meta-analysis](#) been considered?

[Home](#)[Contents](#)

[Comments welcome](#)

[Home](#)[Contents](#)jameslindlibrary.org

Interpreting unbiased comparisons

Differences between intended treatments and treatments actually received

Fair tests of medical treatments have to be planned carefully. The documents setting out these plans are referred to as protocols, and, among other things, they specify details about the treatments that will be compared. The best laid plans don't always work out quite as intended, however. The treatments actually received by patients in tests sometimes differ from those it was intended they should have received. These departures from intention need to be taken into account in interpreting the results of treatment comparisons.

For example, two groups of people might be created using an unbiased method (like random allocation) to compare a new drug with a standard drug. If the new drug has an unexpected, rather nasty side-effect leading some people to abandon it, it would be misleading to exclude those people from the analysis. Comparing only people who could tolerate the new drug with all those who took the standard drug would clearly not be fair. This would result in bias and a misleadingly optimistic picture of the new drug: like would not be being compared with like. The principal comparison must be based as far as possible on all the people assigned to receive each of the treatments compared, in the groups to which they were assigned.

One of the reasons that placebos were introduced in the evolution of fair tests of medical treatments was to reduce departures from intended treatments (Kaptchuk 1998). But things may go astray even in placebo controlled trials. During the 2nd World War, people suffering from colds were given a solution of drug called patulin and compared with other people given only the fluid in which the drug had been dissolved ([MRC 1944](#)). Analysis of the results failed to reveal any beneficial effects of the drug, but then a concern emerged that the liquid used to dissolve the drug might have inactivated it. In other words, over 1000 patients might have participated in a comparison of two inactive treatments! Fortunately, tests confirmed that the patulin used in the trial had indeed been active, although it had no detectable effects on colds (Chalmers and Clarke 2004)!

Treatments received may differ from treatments intended for a variety of reasons. For example, doctors may decide that the treatment to which some of their patients have been allocated in a formal treatment comparison should not be offered to them; patients may reject the treatments allocated to them, or not take them as intended; doses of the treatment different from those intended may be given; or the supply of one of the treatments may run out.

For example, when differences emerged in the results of apparently identical treatments for leukaemia in British and American children, investigation revealed that the worse results in Britain reflected unwillingness among British clinicians to persist with chemotherapy when nasty toxic effects of treatment developed.

For these reasons, interpretations of fair tests must consider the possibility that treatments received were not those intended. If discrepancies between intention and practice have occurred, it is important to consider the implications for interpreting the evidence.

References

Chalmers I, Clarke M (2004). The 1944 Patulin Trial: the first properly controlled multicentre trial conducted under the aegis of the British Medical Research Council. *International Journal of Epidemiology* 32:253-260.

Kaptchuk TJ (1998). Intentional ignorance: a history of blind assessment and placebo controls in medicine. *Bulletin of the History of Medicine* 72:389-433.

Medical Research Council (1944). Clinical trial of patulin in the common cold. *Lancet* 2:373-375.

[Home](#)[Contents](#)

[Comments welcome](#)

[Home](#)[Contents](#)jameslindlibrary.org

Interpreting unbiased comparisons

Taking account of the play of chance

When two treatments are compared, any differences in outcome may simply be caused by the play of chance. For example, take a comparison of a new treatment with a standard treatment in which 4 people improved with the former and 6 people improved with the latter. It would clearly be wrong to conclude confidently that the new treatment was worse than the standard treatment: these results might simply reflect the play of chance. If the comparison was repeated, the numbers of patients who improved might be reversed (6 against 4), or come out the same (5 against 5), or in some other ratio.

[Records dealing with the play of chance](#)

If, however, 40 people improved with the new treatment and 60 with the standard treatment, chance becomes a less likely explanation for the difference. And if 400 people improved with the new treatment and 600 with the standard treatment, it would be clear that the new treatment was indeed very likely to be worse than the standard. The way to reduce the likelihood of being misled by the play of chance in treatment comparisons is thus to ensure that fair tests include sufficiently large numbers of people who experience the outcomes one hopes to influence, such as improvement or deterioration.

In some circumstances very large numbers of people – thousands and sometimes tens of thousands - need to participate in fair tests to obtain reliable estimates of treatment effects. Large numbers of participants are necessary, for example, if the treatment outcomes of interest are rare – for example, heart attacks and strokes among apparently healthy middle-aged women using hormone replacement therapy (HRT). Large numbers are also needed if moderate but important effects of treatments are to be detected reliably – for example, a reduction by 20 per cent in the risk of early death among people having heart attacks.

To assess the role that chance may have played in the results of fair tests, researchers use 'tests of statistical significance'. When statisticians and others refer to 'significant differences' between treatments, they are usually referring to statistical significance. Statistically significant differences between treatments are not necessarily significant in the usual sense of the word. But tests of statistical significance are important nevertheless because they help us to avoid mistaken conclusions that real differences in treatments exist when they don't - sometimes referred to as Type I errors.

It is also important to take account of a sufficiently large number of outcomes of treatment to avoid a far more common danger – concluding that there are no differences between treatments when in fact there are. These mistakes are sometimes referred to as Type II errors. Thomas Graham Balfour was aware of this latter danger when he interpreted the results of his test of claims that belladonna could prevent the orphans under his care developing scarlet fever ([Balfour 1854](#)). Two out of 76 boys allocated to receive belladonna developed scarlet fever compared with 2 out of 75 boys who did not receive the drug. Balfour noted that "the numbers are too small to justify deductions as to the prophylactic power of belladonna". If more of the boys had developed scarlet fever, Balfour might have been able to reach a more confident conclusion about the possible effects of belladonna. Instead, he simply noted that 4 cases of scarlet fever among 151 boys was too small a number to reach a confident conclusion.

One approach that reduces the likelihood that we will be misled by chance effects involves estimating a range of treatment differences within which the real differences are likely to lie ([Gavarret 1840](#); Huth 2006). These range estimates are known as confidence intervals. As illustrated in the opening paragraph of this essay, repeating a treatment comparison is likely to yield varying estimates of the differential effects of treatments on outcomes, particularly if the estimates are based on small numbers of outcomes. Confidence intervals take account of this variation.

Statistical tests and confidence intervals - whether for analysis of individual studies, or in [meta-analysis](#) of a number of separate but similar studies - help us to take account of the play of chance and avoid concluding that treatment effects and differences exist when they don't, and don't exist when they do.

Reference

Balfour TG (1854). Quoted in West C. Lectures on the Diseases of Infancy and Childhood. London, Longman,

Brown, Green and Longmans, p 600.

Gavarret LDJ (1840). *Principes généraux de statistique médicale: ou développement des règles qui doivent présider à son emploi*. Paris: Bechet jeune & Labé.

Huth EJ (2006). Jules Gavarret's *Principes Généraux de Statistique Médicale*: a pioneering text on the statistical analysis of the results of treatments.

[Home](#)

[Contents](#)

[Comments welcome](#)

[Home](#)[Contents](#)jameslindlibrary.org

Identifying unanticipated effects of treatments

It is only to be expected that unanticipated effects of treatments will emerge when new treatments are introduced more widely. Initial tests - for example, those required to license new drugs - cover at most a few hundred or a few thousand people treated for a few months. Only relatively frequent and short-term unanticipated effects are likely to be picked up at this stage.

[Records dealing with unexpected effects](#)

Rare treatment effects, or those that take some time to develop, will not be discovered until there has been more widespread use of treatments. Moreover, new treatments will often be used in people who may differ in important ways from those who participated in the original tests. They may be older or younger, of a different sex, more or less ill, living in different circumstances, or suffering from other health problems in addition to the condition at which the treatment is targeted. These differences may modify treatment effects, and new, unanticipated effects may emerge.

Detection and verification of unanticipated effects, whether [adverse](#) or [beneficial](#), usually occur rather differently from the methods used to assess hoped-for effects of new treatments. Unanticipated effects of treatments are sometimes suspected initially by health professionals or patients. Identifying which among these initial hunches reflect real effects of treatments poses a challenge which will have become familiar to readers of previous essays in this series, namely - to avoid being misled by [biases](#) and the [play of chance](#).

If the unanticipated effect of a treatment is very striking and occurs quite often after the treatment has been used, it may be noticed by health professionals or patients. For example, babies born without limbs are almost unheard of, so when a sudden increase in their numbers occurred in the 1960s it naturally raised concerns. All mothers of such babies had used a newly marketed anti-nausea drug - [thalidomide](#) - prescribed during early pregnancy, so this was likely to be the cause and little further assessment was necessary. Unanticipated beneficial effects of drugs are often detected in similar ways, for example, when it was found that a drug to treat psychosis also lowered cholesterol (Goodwin 1991).

When such striking relationships are noticed, they often turn out to be confirmed as real unanticipated effects of treatment ([Venning 1982](#)). However, a lot of hunches about unanticipated effects of treatment are based on far less convincing evidence. So, as with tests designed to detect hoped-for effects of treatments, planning tests to confirm or dismiss less striking suspected unanticipated effects involves [avoiding biased comparisons](#).

Studies to test whether suspected unanticipated effects of treatment are real must observe the principle of comparing 'like with like'. Random allocation to treatments is the ideal way to accomplish this. Only rarely, however, can suspected treatment effects can be investigated by further analysis of follow-up of people who were randomly allocated to treatments before the were given (Hemminki and McPherson 1997). The challenge is therefore to assemble unbiased comparison groups **after** treatment decisions have been taken in other ways, often using information collected routinely during health care.

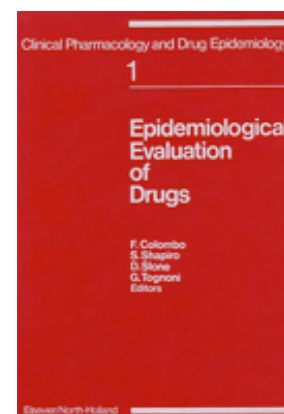
In these studies, it actually helps that the suspect effects were not anticipated at the time that treatment decisions were taken. This is because it means that no account could have been taken of the risk of the suspect condition at the time people were being selected for treatment: the unanticipated effect is usually a different condition or disease from the condition or disease for which the treatment was prescribed (Vandenbroucke 2004a).

For example, when hormone replacement therapy (HRT) was introduced for treating menopausal symptoms, a woman's risk of developing venous thrombosis was unlikely to have been taken into account because most doctors and women thought it was irrelevant. There was therefore no reason to expect that women who were prescribed HRT differed in their risk of developing venous thrombosis from those who did not receive the drug. The basis was thus established for fair tests, and these showed that HRT increases the risk of venous thrombosis.

When a suspected unanticipated effect relates to a treatment for a common health problem (such as heart attack) but does not occur very often with the new treatment (or is not completely relieved by it), large-scale surveillance of people receiving the treatment is needed to detect the unanticipated effect. For example, although some people thought that [aspirin](#) might reduce the risk of heart attack and began fair tests of this theory in patients in the late 1960s ([Elwood et al. 1974](#)), most people would have thought that the theory was highly

implausible. The breakthrough came when a large study was done to detect unanticipated adverse effects of drugs: researchers noticed that people admitted to hospital with heart attacks were less likely to have recently taken aspirin than apparently similar patients ([Boston Collaborative Drug Surveillance Group 1974](#)). These findings were consistent with those of a fair test, in which people had been allocated at random to receive or not receive aspirin after heart attack. The two reports were published back-to-back in the same issue of the British Medical Journal ([BMJ 1974](#)).

The ground rules for detecting and investigating unanticipated effects of treatments were first set out clearly in the late 1970s ([Jick 1977](#); Colombo et al. 1977), drawing on the collective experience of investigating unanticipated effects which had accumulated following the [thalidomide](#) disaster. The requirements for one important type of research, case-control studies of possible adverse effects of treatment, were laid down in a paper based on the experiences of researchers in Boston and Oxford ([Jick and Vessey 1978](#)). With many powerful treatments introduced since that time, this aspect of fair tests of treatments remains just as challenging and important today as it did then (Vandenbroucke 2004b).



As emphasized in previous essays in this series, it is important to recognise that individual reports suggesting or dismissing suspicions about unanticipated effects of treatments can be misleading. As with all other fair tests of treatment, possible unanticipated effects of treatment must be investigated using [systematic reviews](#) of all the relevant evidence, such as those that confirmed the relationship between HRT and heart disease, stroke and breast cancer (Hemminki and McPherson 1997; Collaborative Group on Hormonal Factors in Breast Cancer 1997).

References

- Boston Collaborative Drug Surveillance Group (1974). Regular aspirin intake and acute myocardial infarction. *BMJ* 1:440-443.
- Collaborative Group on Hormonal Factors in Breast Cancer (1997). 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* 350:1047-1059 .
- Colombo F, Shapiro S, Slone D, Tognoni G, eds (1977). *Epidemiological Evaluation of Drugs*. Amsterdam: Elsevier/North Holland Biomedical Press, 1977.
- Elwood PC, Cochrane AL, Burr ML, Sweetnam PM, Williams G, Welsby E, Hughes SJ, Renton R (1974). A randomised controlled trial of acetyl salicylic acid in the secondary prevention of mortality from myocardial infarction. *BMJ* 1:436-440.
- Goodwin JS (1991). The empirical basis for the discovery of new therapies. *Perspectives in Biology and Medicine* 35:20-36.
- Hemminki E, McPherson K (1997). Impact of postmenopausal hormone therapy on cardiovascular events and cancer: pooled data from clinical trials. *BMJ*;315:149-153.
- Jick H (1977). The discovery of drug-induced illness. *New England Journal of Medicine* 296:481-485.
- Jick H, Vessey M (1978). Case-control studies in the evaluation of drug-induced illness. *American Journal of Epidemiology* 107:1-7.
- Vandenbroucke JP (2004a). When are observational studies as credible as randomised trials? *Lancet* 363:1728-1731.
- Vandenbroucke JP (2004b). Benefits and harms of drug treatments. *BMJ* 329:2-3.
- Venning GR (1982). Validity of anecdotal reports of suspected adverse drug reactions: the problem of false alarms. *BMJ* 284:249-254.

Comments welcome

[Home](#)[Contents](#)jameslindlibrary.org

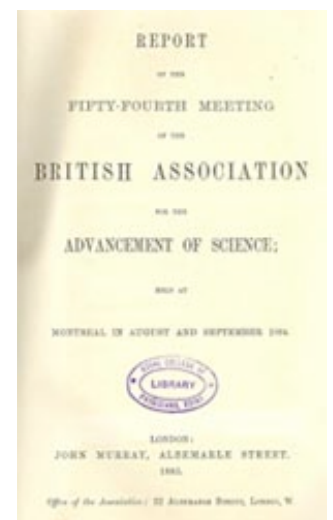
Systematic reviews of all the relevant evidence

One of the twentieth century pioneers of fair tests of treatments, [Austin Bradford Hill](#), noted that readers of reports of research want answers to four questions: 'Why did you start?', 'What did you do?', 'What did you find?', and 'What does it mean anyway?' (Hill 1965). The quality of the answer to Hill's last question is particularly important because this is the element of a research report which is most likely to influence actual choices and decisions about treatments.

Records dealing with systematic review

Only very rarely will a single fair test of a treatment yield sufficiently strong evidence to provide a confident answer to the question 'What does it mean?'. A fair test of a treatment is usually one of a number of tests addressing the same question. For a reliable answer to the question 'What does it mean?', then, it is important to interpret the evidence from a particular fair test in the context of a careful assessment of all the evidence from fair tests that have addressed the question concerned.

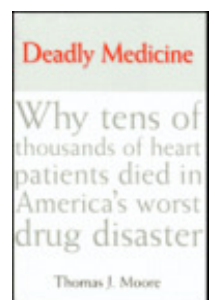
The president of the British Association for the Advancement of Science expressed the need to observe this principle more than a century ago:



"If, as is sometimes supposed, science consisted in nothing but the laborious accumulation of facts, it would soon come to a standstill, crushed, as it were, under its own weight.... Two processes are thus at work side by side, the reception of new material and the digestion and assimilation of the old...The work which deserves, but I am afraid does not always receive, the most credit is that in which discovery and explanation go hand in hand, in which not only are new facts presented, but their relation to old ones is pointed out." ([Rayleigh 1885](#))

Perhaps it is because application of this principle in practice still attracts little credit within academia that very few reports of fair tests of treatments discuss their results in the context of a systematic assessment of all the other relevant evidence (Clarke et al. 2002). As a result, it is usually difficult for readers to obtain a reliable answer to the question 'What does it mean?' from reports of new research.

As noted in an earlier explanatory essay, embarking on new tests of medical treatments without first reviewing systematically what can be learnt from existing research is dangerous, wasteful and unethical (see [Why comparisons must address genuine uncertainties](#)). Reporting the results of new tests without interpreting new evidence in the light of systematic assessments of other relevant evidence is also dangerous because it results in delays in the identification of both useful and harmful treatments (Antman et al. 1992). For example, between the 1960s and the early 1990s, over 50 fair tests of drugs to reduce heart rhythm abnormalities in people having heart attacks were done before it was realised that these drugs were killing people. Had each report assessed the results of new tests in the context of all the relevant evidence, the lethal effects of the drugs could have been identified a decade earlier, and many unnecessarily premature deaths could have been avoided.



In an age of electronic publishing it should be possible to deal with these limitations of most reports of new research (Chalmers and Altman 1999; Smith and Chalmers 2001). However, users of research evidence are increasingly turning for reliable information to [up-to-date, systematic reviews](#) of all relevant, reliable evidence, because these are increasingly recognised as providing the best basis for conclusions about the effects of medical

treatments.

Just as it is important to take steps to avoid being misled by [biases](#) and the [play of chance](#) in planning, conducting, analysing and interpreting individual fair tests of treatments, so also must similar steps be taken in planning, conducting, analysing and interpreting systematic reviews. This entails:

- specifying the question to be addressed by the systematic review
- defining eligibility criteria for studies to be included
- identifying (all) potentially eligible studies
- applying eligibility criteria in ways that limit bias
- assembling as high a proportion as possible of the relevant information from the studies
- analysing this information, if appropriate and possible, using meta-analysis and a variety of analyses
- preparing a structured report

One manifestation of the increasing recognition of the crucial importance of systematic reviews for assessing the effects of treatments is the rapid evolution of methods to improve the reliability of reviews themselves. The first edition of a book entitled *Systematic Reviews* was less than 100 pages long ([Chalmers and Altman 1995](#)): only six years later, the second edition weighed in at nearly 500 pages (Egger et al. 2001).

There are currently important developments in the methods used for preparing systematic reviews, including those needed to identify unexpected effects of treatments (Glasziou et al. 2004) and for incorporating the results of research describing and analysing the experiences of people giving and receiving treatments (Thomas 2004). Relevant material will be added to *The James Lind Library* as it emerges.

References

Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. *JAMA* 268:240-48.

Chalmers I, Altman DG (1995). *Systematic Reviews*. London: BMJ Publications.

Chalmers I, Altman DG (1999). How can medical journals help prevent poor medical research? Some opportunities presented by electronic publishing. *Lancet* 353:490-493.

Egger M, Davey Smith G, Altman D (2001). *Systematic Reviews in Health Care: meta-analysis in context*. 2nd Edition of *Systematic Reviews*. London: BMJ Books.

Glasziou P, Vandenbroucke J, Chalmers I (2004). Assessing the quality of research *BMJ* 328:39-41.

Hill AB (1965). Cited in 'The reasons for writing'. *BMJ* 4:870.

Rayleigh (1885). Address by the Rt. Hon. Lord Rayleigh. In: Report of the fifty-fourth meeting of the British Association for the Advancement of Science; held at Montreal in August and September 1884, London: John Murray.

Smith R, Chalmers I (2001). Britain's gift: a 'Medline' of synthesized evidence. *BMJ* 323:1437-1438.

Thomas J, Harden A, Oakley A, Oliver S, Sutcliffe K, Rees R, Brunton G, Kavanagh J (2004). Integrating qualitative research with trials in systematic reviews *BMJ* 328:1010-1012.

[Home](#)

[Contents](#)

[Comments welcome](#)

[Home](#)[Contents](#)jameslindlibrary.org

Systematic reviews of all the relevant evidence

Dealing with biased reporting of the available evidence

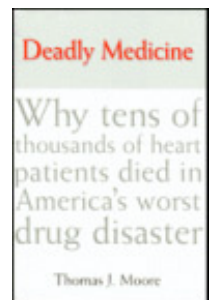
[Avoiding biased comparisons](#) entails identifying and taking account of all the relevant reliable evidence in systematic reviews. This is challenging in many ways, particularly as some pertinent evidence is not published because biased decisions are made about which results of research are submitted and accepted for publication. Studies that have yielded 'disappointing' or 'negative' results are less likely to be reported than others. This is often called 'publication bias' or 'reporting bias'.

[Records dealing with reporting bias](#)

These reporting biases have been recognized for centuries (Dickersin 2004a). In 1792, for example, James Ferriar stressed the importance of recording treatment failures as well as treatment successes (Ferriar 1792). This principle was reiterated in an editorial published in the Boston Medical and Surgical Journal just over a century later ([Editorial 1909](#)).

There is now a large body of evidence confirming that reporting bias is a substantial problem. There is also evidence that reporting bias results principally from researchers not writing up or submitting reports of research for publication, not because of biased rejection of submitted reports by journal editors (Dickersin 2004b). Recent research has also revealed an additional problem: if estimates of treatment effects on some of the outcomes studied don't support the conclusions of researchers, these data sometimes don't get reported either (Chan et al. 2004).

For example, had all the studies of the effects of giving drugs to reduce heart rhythm abnormalities in patients having heart attacks been reported, tens of thousands of deaths from these drugs could have been avoided. In 1993, Dr Cowley and his colleagues pointed out how an unpublished study done 13 years previously might have "provided an early warning of trouble ahead". Nine patients had died among the 49 assigned to the anti-arrhythmic drug (lorcainide) compared with only one patient among a similar number given placebos. "When we carried out our study in 1980", they reported, "we thought that the increased death rate was an effect of chance...The development of lorcainide was abandoned for commercial reasons, and this study was therefore never published; it is now a good example of 'publication bias'" (Cowley et al. 1993).



Reporting biases tend to lead to conclusions that medical treatments are more useful than they are in fact. They can therefore result in unnecessary suffering and death, and in wasted resources spent on ineffective or dangerous treatments (Chalmers 2004). People who agree to researchers' requests that they participate in tests of treatments assume that their participation will lead to an increase in knowledge. This implied contract between researchers and participants in research is breached by researchers who do not make public the results of the research.

Biased under-reporting of research is scientific misconduct and unethical (Chalmers 1990). Selective reporting of studies sponsored by the pharmaceutical industry is a particular problem (Melander et al. 2003). Research ethics committees, medical ethicists and research funders have so far not done enough to protect patients and the public from the adverse effects of reporting biases (Savulescu et al. 1996). Fair testing of treatments – particularly those treatments in which there is commercial interest – will remain compromised just as long as this form of research misconduct is tolerated by governments and others who should be protecting the interests of the public.

We must all support the lead given by the World Health Organisation to reduce reporting biases by requiring registration of all fair tests of treatments at inception (www.who.int/ictpr), and insisting that their results should be published.

References

Chalmers I (1990). Under-reporting research is scientific misconduct. JAMA 263:1405-1408.

Chalmers I (2004). In the dark: drug companies should be forced to publish all the results of clinical trials. New Scientist 181:19.

Chan A-W, Hróbjartsson A, Haahr M, Gøtzsche PC, Altman DG (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to publications. *JAMA* 291:2457-2465.

Cowley AJ, Skene A, Stainer, Hampton JR (1993). The effect of lorcaïnide on arrhythmias and survival in patients with acute myocardial infarction. *International Journal of Cardiology* 40:161-166.

Dickersin K (2004a). Publication bias: recognising the problem, understanding its origins and scope, and preventing harm. In: Rothstein H, Sutton A, Borenstein M, eds. *Handbook of publication bias*. New York: Wiley.

Dickersin K (2004b). How important is publication bias? A synthesis of available data. *AIDS Educ Prev* 1997;9 (1 Suppl):15-21.

Editorial (1909). The reporting of unsuccessful cases. *Boston Medical and Surgical Journal* 161:263-264.

Ferriar J (1792). *Medical histories and reflexions*. Vol 1. London: Cadell and Davies, 1792.

Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B (2003). Evidence b(i)ased medicine - selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. *BMJ* 326:1171-3.

Savulescu J, Chalmers I, Blunt J (1996). Are research ethics committees behaving unethically? Some suggestions for improving performance and accountability. *BMJ* 313:1390-1393.

[Home](#)

[Contents](#)

[Comments welcome](#)

[Home](#)[Contents](#)jameslindlibrary.org

Systematic reviews of all the relevant evidence

Avoiding biased selection from the available evidence

Biases can distort tests of medical treatments and lead to erroneous conclusions. They can also distort reviews of evidence. Plans for systematic reviews should be set out in protocols, such as those published by [The Cochrane Collaboration](#), making clear what measures will be taken to reduce biases.

These include specifying clearly:

- which question about treatments will be addressed in the review;
- the criteria that will make a study eligible for inclusion;
- the strategies that will be used to search for potentially eligible studies; and
- the steps that will be taken to minimise biases in selecting studies and data for inclusion in the review (Berlin 1997).

Different systematic reviews addressing what appears to be the same question about the effects of medical treatments quite often reach different conclusions. Sometimes this is because the questions addressed are subtly different. Sometimes it reflects differences in the materials and methods used by the reviewers, and in these circumstances it is important to judge which reviews are most likely to have reduced allocation biases most successfully.

It is also worth considering whether the reviewers have other interests that might affect the conduct or interpretation of their review. For example, people associated with the manufacturers of evening primrose oil reviewed the drug's effects on eczema (Morse et al. 1989). They reached a far more enthusiastic conclusion about the value of the drug than a review done by investigators with no commercial interest, who included the results of unpublished studies in their assessment (Williams 2003).

It is not only commercial interests that can lead to biased selection from the available evidence for inclusion in reviews. We all have prejudices that can lead to biased selection of evidence, and researchers, health professionals, patients and others assessing the effects of treatments are no exception.

References

Berlin JA (1997). Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group. *Lancet* 350:185-186.

Morse PF, Horrobin DF, Manku MS, Stewart JC, Allen R, Littlewood S, Wright S, Burton J, Gould DJ, Holt PJ, et al (1989). Meta-analysis of placebo-controlled studies of the efficacy of Epogam in the treatment of atopic eczema. Relationship between plasma essential fatty acid changes and clinical response. *British Journal of Dermatology* 121:75-90.

Sackett DL, Oxman AD (2003). HARLOT plc: an amalgamation of the world's two oldest professions *BMJ* 327:1442-1445.

Williams HC (2003). Evening primrose oil for atopic dermatitis. *BMJ* 327:1358-1359.

[Home](#)[Contents](#)

[Comments welcome](#)

[Home](#)[Contents](#)jameslindlibrary.org

Systematic reviews of all the relevant evidence

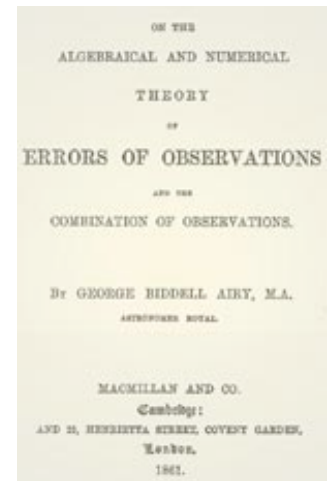
Reducing the play of chance using meta-analysis

[Systematic reviews](#) of all the relevant, reliable evidence are needed for fair tests of medical treatments. To avoid misleading conclusions about the effects of treatments, people preparing systematic reviews must take steps to avoid [biases](#) of various kinds, for example, by [taking account of all the relevant evidence](#) and by [avoiding biased selection from the available evidence](#).

[Records dealing with meta-analysis](#)

Even though care may be taken to minimize biases in reviews, however, misleading conclusions about the effects of treatments may also result from the [play of chance](#). Discussing separate but similar studies one at a time in systematic reviews may also leave a confused impression because of the play of chance. If it is both possible and appropriate, this problem can be reduced by combining the data from all the relevant studies, using a statistical procedure now known as 'meta-analysis'.

Most statistical techniques used today in meta-analysis derive from the work of the German mathematician Karl Gauss and the French mathematician Pierre-Simon Laplace during the first half of the 19th century. One of the fields in which their methods found practical application was astronomy: measuring the position of stars on a number of occasions often resulted in slightly different estimates, so techniques were needed to combine the estimates to produce an average derived from the pooled results. In 1861, the British Astronomer Royal, George Airy, published a 'textbook' for astronomers ([Airy 1861](#)) in which he described the methods used for this process of quantitative synthesis. Just over a century later, an American social scientist, Gene Glass, named the process 'meta-analysis' (Glass 1976).



An early medical example of meta-analysis was published in the British Medical Journal in 1904 by Karl Pearson ([Pearson 1904](#); O'Rourke 2006), who had been asked by the government to review evidence on the effects of a vaccine against typhoid. Although methods for meta-analysis were developed by statisticians over the subsequent 70 years, it was not until the 1970s that they began to be applied more widely, initially by social scientists (Glass 1976), and then by medical researchers (Stjernswärd J 1974; Stjernswärd et al. 1976; Cochran et al. 1977; Chalmers et al. 1977; Chalmers 1979; Editorial 1980).

Meta-analysis can be illustrated using the logo of [The Cochrane Collaboration](#). The logo illustrates a meta-analysis of data from seven fair tests. Each horizontal line represents the results of one test (the shorter the line, the more certain the result); and the diamond represents their combined results. The vertical line indicates the position around which the horizontal lines would cluster if the two treatments compared in the trials had similar effects; if a horizontal line crosses the vertical line, it means that that particular test found no clear ('statistically significant') difference between the treatments. When individual horizontal lines cross the vertical 'no difference' line, it suggests that the treatment might either increase or decrease infant deaths. Taken together, however, the horizontal lines tend to fall on the beneficial (left) side of the 'no difference' line. The diamond represents the combined results of these tests, generated using the statistical process of meta-analysis. The position of the diamond clearly to the left of the 'no difference' line indicates that the treatment is beneficial.



This diagram shows the results of a systematic review of fair tests of a short, inexpensive course of a steroid drug given to women expected to give birth prematurely. The first of these tests was reported in 1972. The diagram summarises the evidence that would have been revealed had the available tests been reviewed systematically a decade later, in 1981: it indicates strongly that steroids reduce the risk of babies dying from the complications of immaturity. By 1991, seven more trials had been reported, and the picture in the logo had become still stronger.

No systematic review of these trials was published until 1989 (Crowley 1989), so most obstetricians, midwives, and pregnant women did not realise that the treatment was so effective. After all, some of the tests had not shown a 'statistically significant' benefit, and maybe only these tests had been noticed. Because no systematic reviews had been done, tens of thousands of premature babies suffered, and died unnecessarily and resources were wasted on unnecessary research. This is just one of many examples of the human costs that can result from failure to assess the effects of treatments in [systematic, up-to-date reviews](#) of fair tests, using meta-analysis to reduce the likelihood that the [play of chance](#) will be misleading.

By the end of the 20th century it had become widely accepted that meta-analysis was an important element of fair tests of treatments, and that it helped to avoid incorrect conclusions that treatments had no effects when they were, in fact, either useful or harmful.

References

- Airy GB (1861). On the algebraical and numerical theory of errors of observations and the combination of observations. London: Macmillan.
- Chalmers I (1979). Randomized controlled trials of fetal monitoring 1973-1977. In: Thalhammer O, Baumgarten K, Pollak A, eds. Perinatal Medicine. Stuttgart: Georg Thieme, 260-265.
- Chalmers TC, Matta RJ, Smith H, Kunzler A-M. (1977). Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *New England Journal of Medicine* 297:1091-1096.
- Crowley P (1989). Promoting pulmonary maturity. In: Chalmers I, Enkin M, Keirse MJNC, eds. *Effective care in pregnancy and childbirth*. Oxford: Oxford University Press, pp 746-762.
- Editorial (1980). Aspirin after myocardialinfarction. *Lancet* 1:1172-3.
- Glass GV (1976). Primary, secondary and meta-analysis of research. *Educational Researcher* 10, 3-8.
- O'Rourke K (2006). An historical perspective on meta-anlysis: dealing quantatively with varying study results. *The James Lind Library*.
- Pearson K (1904). Report on certain enteric fever inoculation statistics. *BMJ* 3:1243-1246.
- Stjernswärd J (1974). Decreased survival related to irradiation postoperatively in early operable breast cancer. *Lancet* 2:1285-1286.
- Stjernswärd J, Muenz LR, von Essen CF (1976). Postoperative radiotherapy and breast cancer. *Lancet* 1:749.

[Home](#)

[Contents](#)

[Comments welcome](#)

[Home](#)[Contents](#)jameslindlibrary.org

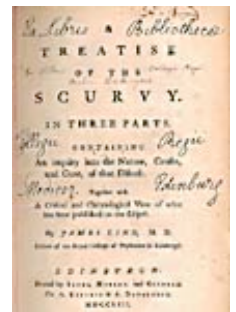
Up-to-date, systematic reviews of all relevant, reliable evidence

Fair tests of treatments in health care

Over the past half century, health care has had a substantial impact on people's chances of living longer, and being free of serious health problems. It has been estimated that health care has been responsible for between a third and a half of the increase in life expectancy and an average of five additional years free of chronic health problems (Bunker et al. 1994). Even so, the public could have obtained - and still could obtain - far better value for the very substantial resources invested in research intended to improve health. Furthermore, some of the treatment disasters of the past could have been prevented, and others could be prevented in future.

The results of individual fair tests of medical treatments are only very rarely set systematically in the context of other similar studies, using methods to reduce [biases](#) and the [play of chance](#). This failure to do systematic reviews of research on the effects of treatments has resulted in a great deal of avoidable suffering. [Fair tests of treatments in health care](#) also entails unbiased preparation of systematic reviews of all the relevant, reliable research studies of the treatments being assessed.

There are some examples of this process going back more than 200 years. In 1753, for example, in his review of the large number of reports about the prevention and treatment of scurvy, James Lind noted:



"As it is no easy matter to root out prejudices,....it became requisite to exhibit a full and impartial view of what had hitherto been published on the scurvy... Indeed, before the subject could be set in a clear and proper light, it was necessary to remove a great deal of rubbish." ([Lind 1753](#))

Systematic reviews of all relevant research addressing questions about the effects of treatments are increasingly seen as providing the most reliable basis for conclusions about treatment effects. Sometimes systematic reviews will show that no reliable evidence exists, and this is one of their most important functions. Similarly, systematic reviews may sometimes confirm that reliable evidence is limited to a single study, and here, too, it is important to make this situation explicit.

The realisation that systematic reviews are needed to provide fair tests of treatments has been reflected in a rapid increase in the numbers of reports of systematic reviews being published on paper and electronically ([DARE](#); [The Cochrane Collaboration](#)). These are being used (i) to inform clinical practice, for example, through the BMJ publication [Clinical Evidence](#) and the [Scottish Intercollegiate Guidelines Network](#); (ii) to assess which medical treatments are cost-effective, for example, by the [National Institute for Health and Clinical Excellence](#); and (iii) to meet the needs of patients for reliable information about the effects of treatments, for example, through [Informed Health Online](#) and the [National Library for Health](#).

Unfinished business

These and similar developments show that the importance of systematic reviews has been accepted by those who are trying to improve access to the evidence needed to inform choices in health care. However, there is still a long way to go: it has been estimated that the Cochrane Collaboration's current output of several thousand systematic reviews will need to be increased to well over 10,000 to cover existing evidence (Mallett and Clarke 2002), and then kept up to date as new evidence emerges. Indeed, one journal editor has suggested that there should be a moratorium on all new research until we've caught up with what existing evidence can tell us (Bausell 1993).

Those responsible for disbursing funds for research must ensure that resources are provided to cope with this backlog, and that new studies are only supported if systematic reviews of existing evidence have shown that additional studies are necessary, and that they have been designed to take account of the lessons from previous research. If journal editors are to serve their readers better, they must follow the lead of *The Lancet* and ensure

that reports of new studies make clear what contribution new evidence has made to an up-to-date systematic review of all the relevant evidence (Young and Horton 2005).

The increased availability of up-to-date, systematic reviews is improving the quality of information about the effects of treatments, but the conclusions of systematic reviews should not be accepted uncritically. Different reviews purportedly addressing the same question about treatments sometimes arrive at different conclusions. Their authors are human and we need to be aware that they may select, analyse and present evidence in ways that support their prejudices and interests. The continuing evolution of reliable methods for preparing and maintaining systematic reviews will help to address this problem, but they cannot be expected to abolish it.

Although growth in the numbers of systematic reviews has increased the availability of the primary fair tests of treatments in health care, these reviews often reveal the poor quality and irrelevance of much research on the effects of treatments. As one editorialist commenting on "the scandal of poor medical research" put it, we need less research, better research and research done for the right reasons (Altman 1994). It seems unlikely that this will be achieved without greater public understanding of the rationale for and characteristics of fair tests of treatments, and greater public influence on and involvement in all phases of fair testing of treatments. Promotion of this agenda depends on uncertainties about the effects of treatments being confronted by new alliances of patients and clinicians (Chalmers 2004; www.duets.nhs.uk; [James Lind Alliance](#)).

The public and health professionals will be well served when they have readier access to up-to-date, systematic reviews of all relevant, reliable evidence addressing important uncertainties about the effects of treatments, and to information about ongoing research addressing these uncertainties (Smith and Chalmers 2001).

References

Altman (1994). The scandal of poor medical research. *BMJ* 308:283-284.

Bausell BB (1993). After the meta-analytic revolution. *Evaluation and the Health Professions* 16:3-12.

Bunker JP, Frazier HS, Mosteller F (1994). Improving health: measuring effects of medical care. *Milbank Quarterly* 72:225-258.

Chalmers I (2004). Well informed uncertainties about the effects of treatments: how should clinicians and patients respond? *BMJ* 328:475-476.

Lind J (1753). A treatise of the scurvy. In three parts. Containing an inquiry into the nature, causes and cure, of that disease. Together with a critical and chronological view of what has been published on the subject. Edinburgh: Printed by Sands, Murray and Cochran for A Kincaid and A Donaldson.

Mallett S, Clarke M (2002). The typical Cochrane Review. *International Journal of Technology Assessment in Health Care* 18:820-823.

Smith R, Chalmers I (2001). Britain's gift: a 'Medline' of synthesized evidence. *BMJ* 323:1437-1438.

Young C, Horton R (2005). Putting clinical trials into context. *Lancet* 366:107-8.

[Home](#)

[Contents](#)

[Comments welcome](#)