The James Lind Library's Introduction to Fair Tests of Treatments

Mike Clarke, Patricia Atkinson, Douglas Badenoch, Iain Chalmers, Paul Glasziou, Scott Podolsky, Ulrich Tröhler

Contents

Introduction

At various times in our lives and to varying levels of intensity, we all use, provide or pay for health and social care. As we decide what to do, take, offer or buy, we need evidence that is reliable, robust and trustworthy about different options. Even before James Lind's experiment comparing possible treatments for scurvy on HMS Salisbury people had recognised that getting this evidence requires strenuous efforts to reduce bias – but that achieving this is often not straightforward. This book of essays from the James Lind Library is our attempt to illustrate some of the challenges encountered and how to overcome them.

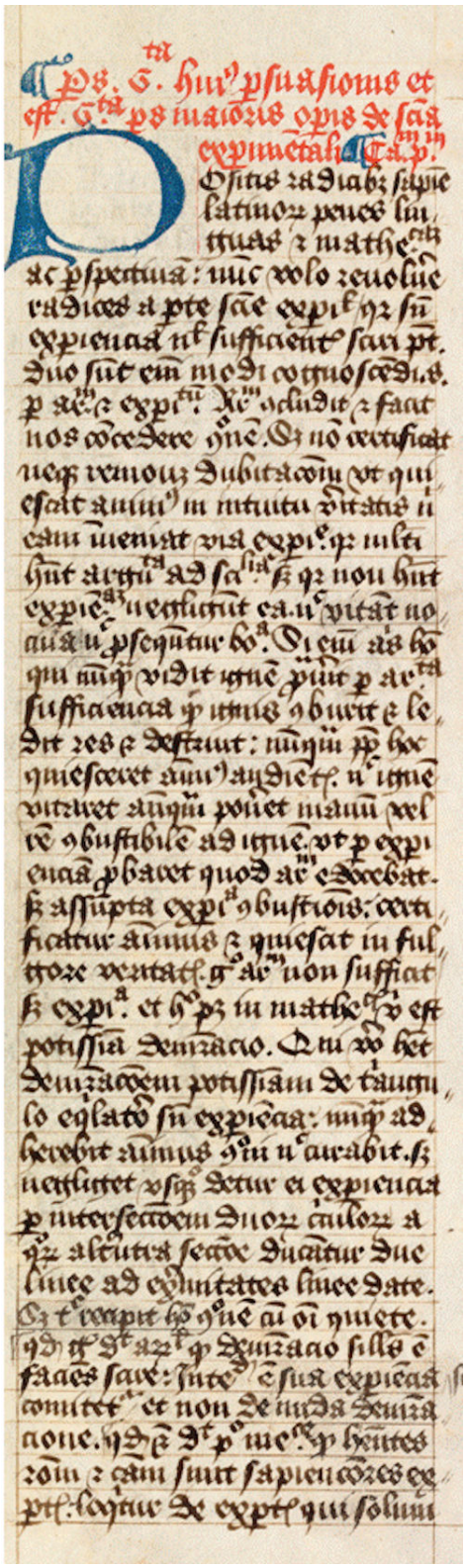
We will take you on a journey through the sometimes stormy waters of why treatments need to be tested, rather than being based on assumptions that "it must work" before the treatment has even been tried, or based on impressions after it has been used a few times, through to the need for fair tests comparing alternative treatment options. We will show why genuine uncertainties must be identified and addressed, and how research to find the most effective and appropriate treatments need to build on research to identify the most effective and appropriate methods for doing that research. We will navigate through the reasons why comparisons need to be fair at the outset, and then kept fair as the treatments being tested are given; outcomes are measured; and results are analysed, reported, and combined in systematic reviews of all the relevant, trustworthy evidence.

We have not cluttered the chapters with references to all the source material on which we have drawn. For that level of detail, please follow the links to the fuller essays on the James Lind Library website (www.jameslindlibrary.org). Instead, where we know of reviews of methodology research which are relevant to a topic, we have listed these at the end of each chapter.

By the end of the book, we hope that you will recognise how, to bring benefits of research to patients and the public, systematic reviews of fair tests are needed to provide key elements of the knowledge needed to inform decisions about health and social care, while taking into account other important factors, such as

values, preferences, needs, resources and priorities. We also hope that, as you finish the book, you will share the sense of enlightenment, education and enjoyment that we have gained from preparing it.

Finally, we dedicate this book to England's National Institute for Health Research. Without the Institute's16-year-long support for the James Lind Initiative, the home of the James Lind Library during that time, neither the Library nor these essays would have been possible. And we also wish to acknowledge the role the Institute plays in recognising the vital contribution of research to the delivery of health and social care that is effective and efficient, and the Institute's leadership in ensuring that the research itself is effective, efficient and reliable, with minimal waste.



"Without experiment nothing can be sufficiently known"

Bacon, Roger (1266)

Opus maius. MS Digby 325, 15th century manuscript. Bodleian Library, Oxford

Section 1: Introduction to JLL Explanatory Essays

Despite acting with the best of intentions, health professionals have sometimes done more harm than good to the patients who have put their trust in them and looked to them for help. Some of this suffering can be reduced by ensuring that fair tests are done to address uncertainties about the effects of treatments.

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 for an average of five additional years free of chronic health problems. Even so, the public could have obtained – and still could obtain – far better value for the very substantial resources it invested in research intended to improve health. Furthermore, some of the treatment disasters of the past could have been avoided, and others could be prevented in future.

Misleading claims about the effects of treatments are common, so all of us should understand how to recognise a valid claim about the effects of treatments and how these are made. Without this knowledge, we risk concluding that useless treatments are helpful, or that helpful treatments are useless. The James Lind Library has been created to improve general understanding of fair tests of treatments in health care, and how they have evolved over time.

These Explanatory Essays provide a brief introduction and overview of the scope of the Library. You can explore in more detail by following the links to in-depth Articles and primary Records in the Library itself.

In more depth

The James Lind Library 1.0 Introduction to JLL Explanatory Essays (http://www.jameslindlibrary.org/essays/1-0-introduction-to-jll-explanatory-essays/)

1.1 Why treatment uncertainties should be addressed

Ignoring uncertainties about the effects of treatments has led to avoidable suffering and deaths. To reduce this suffering and premature mortality, treatment uncertainties must be acknowledged and addressed, first by reviewing systematically what is already known, and then by doing well-designed research to reduce continuing uncertainties.

Trying to do more good than harm

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 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. 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.

A first step is acknowledging that treatments can sometimes do more harm than good. We need to be more willing to admit uncertainties about treatment effects, and to promote tests of treatments to adequately reduce uncertainties. We refer to such tests as 'fair tests'.

In the 17th century, the Flemish physician Jean-Baptiste van Helmont proposed casting lots to decide which patients should be treated by orthodox physicians with bloodletting and purging, and which by him without these treatments.

Van Helmont JB (1648)

Ortus medicinæ: Id est Initia physicæ inaudita. Progressus medicinae novus, in morborum ultionem, ad vitam longam [The dawn of medicine: That is, the beginning of a new Physic. A new advance in medicine, a victory over disease, to (promote) a long life]. Amsterodami: Apud Ludovicum Elzevirium, pp 526-527.

Why theories about the effects of treatments must be tested in practice

People have often been harmed because treatments have been based only on theories about how health problems should be treated, without testing how the theories play out in practice.

For example, for centuries people believed the theory that illnesses were caused by 'humoral imbalances'. As a result, patients were bled and purged, made to vomit and take snuff, all in the belief that this would end the supposed imbalances. Still, a 17th century Flemish doctor, Johannes van Helmont, was impertinent enough to challenge the medical authorities of the time to assess the validity of their theories by proposing a fair test of the results of their unpleasant treatments.

By the end of the 18th century, British army and naval surgeons had begun to show the harmful effects of bloodletting for treating "fevers". A few decades later, the practice was also challenged by a Parisian physician. Yet at the beginning of the 20th century, mainstream doctors in Boston, USA, who were not using bloodletting to treat pneumonia were still being judged negligent. Indeed, Sir William Osler, one of the most influential medical authorities in the world, who was generally cautious about recommending unproven treatments, advised his readers at the end of the 19th century that: "during the last decades we have certainly bled too little."

Although the need to test the validity of theories in practice was recognized by Islamic physicians at least a millennium ago, 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 bestselling 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. We now know that this advice, apparently rational in theory, led to

Dr Spock's bestselling book contained harmful advice on babies' sleeping position.

the cot (crib) deaths of tens of thousands of infants.

Another dramatic example of the dangers of applying untested theory in practice was found in the use of drugs to prevent heart rhythm abnormalities in people having heart attacks. Because heart rhythm abnormalities are associated with an increased risk of early death after heart attack, the theory was that drugs that reduced these abnormalities would also reduce early deaths. Just because a theory seems reasonable doesn't prove that it is necessarily right when applied in practice. Years after the drugs had been licensed, it was discovered that they actually increase the risk of sudden death after heart attack in these patients. 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 – many more than the total number of Americans who died through the whole of the Vietnam War.

Misplaced confidence in theoretical thinking has also resulted in some treatments being rejected inappropriately because researchers did not believe that they could work. 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. Only when such patients participated in fair tests of these drugs did we find out that this treatment could help them survive the heart attack.

Observations in clinical practice or in laboratory and animal research may suggest that 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 genuinely do more good than harm, or vice versa.

In more depth

The James Lind Library 1.1 Why treatment uncertainties should be addressed (http://www.jameslindlibrary.org/essays/1-1-why-treatment-uncertainties-should-be-addressed/)

Systematic reviews of methodology:

- Djulbegovic B, Kumar A, Glasziou PP, Perera R, Reljic T, Dent L, Raftery J, Johansen M, Di Tanna GL, Miladinovic B, Soares HP, Vist GE, Chalmers I (2012). New treatments compared to established treatments in randomized trials. Cochrane Database of Systematic Reviews (10):MR000024
- Perel P, Roberts I, Sena E, Wheble P, Briscoe C, Sandercock S, Macleod M, Mignini LE, Jayaram P, Khan KS (2007). Comparison of treatment effects between animal experiments and clinical trials: systematic review. BMJ 334:197-200
- Price A, Albarqouni L, Kirkpatrick J, Clarke M, Liew SM, Roberts N, Burls A (2018). Patient and public involvement in the design of clinical trials: An overview of systematic reviews. Journal of Evaluation in Clinical Practice 24(1):240-53

1.2 Why treatment comparisons are essential

Treatment comparisons are required to take account of the natural course of health problems and 'placebo effects', and to go beyond impressions about treatment effects. But treatment comparisons need to be fair to avoid untrustworthy and sometimes dangerously incorrect conclusions about the effects of treatments.

The effects of nature and time?

People often recover from illness without any specific treatment: nature and time are great healers. 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.

As the American physician and poet Oliver Wendell Holmes suggested in the 19th century when there were very few useful treatments,

> *"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."*

Put another way, "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."

The progress and outcome of illness if left untreated must obviously be considered when treatments are being tested. We must take care to ensure that we don't mistakenly believe the effects of time and nature are caused by a treatment we happened to be taking. In 1616, James V1 observed that many people made this mistake with tobacco smoking, thinking that their natural recovery from colds was due to their smoking, when in fact they would have got better anyway.

James VI of Scotland and I of England identified, in the case of people believing tobacco could cure illnesses, the logical fallacy of ascribing a person's natural recovery from disease to whatever treatment they happened to have tried.

Stuart, James, King of Great Britaine, France and Ireland (1616).
A counterblaste to tobacco. In: The workes of the most high and mightie prince, James Published by James, Bishop of Winton, and Deane of his Majesties Chappel Royall. London: printed by Robert Barker and John Bill, printers to the Kings most excellent Majestie, pp 214-222.

'Placebo effects'

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 1959:

> *"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."*

In the knowledge that much illness is self-limiting, doctors sometimes prescribe inert treatments in the hope that their patients will derive psychological benefit – 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 effect of using placebos for centuries. William Cullen referred to his use of a placebo as long ago as 1772, and references to placebos increased during the 19th century. The American physician Austin Flint believed that orthodox drug treatment was getting the credit due to 'nature', so he gave thirteen patients with rheumatism a 'placeboic remedy' consisting of a highly dilute extract of the bark of the quassia tree. He found that "the favourable progress of the cases was such as to secure for the remedy generally the entire confidence of the patients". 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". Thus, we must ensure that improvements seen with remedies aren't owing to the charisma of the clinician, or the expectation that the remedies will make them better.

The reported cases were treated throughout the whole course of the disease with only palliative measures. These measures, as will be seen, consisted of opium in some form, given in small or moderate doses, the application generally of dry flannel to the affected joints, and the use of either the soap and opium liniment, camphorated oil, or the tincture of aconite. But to secure the moral effect of a remedy given specially for the disease, the patients were placed on the use of a placebo which consisted, in nearly all of the cases, of the tincture of quassia, very largely diluted. This was given regularly, and became well known in my wards as the *placeboic remedy* for rheumatism. The favourable progress of the cases was such as to secure for the remedy generally the entire confidence of the patients. I may add that all the cases were brought before the medical class in attendance during the winter.

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.

The need for comparisons

Just as the healing power of nature and the placebo effect have been recognized for centuries, so also has the need for comparisons to assess the effects of 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 16th century French military surgeon, resolved that treatment of battle wounds with boiling oil (as was common practice) was harmful. He concluded this when the supply of oil ran out and he observed that his patients recovered much better than those he had previously treated in the usual way.

Most of the time, 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.

The James Lind Library contains examples both of dramatic beneficial effects of treatments – for example, opium for pain relief, insulin for diabetes, liver diet for pernicious anaemia, sulpha drugs for infection after childbirth and streptomycin for tuberculous meningitis – and of dramatic harmful effects, for example, limb reduction deformities caused by thalidomide. Sometimes a treatment - sulphonamide drugs for example - can have a dramatic effect in some diseases, but modest or little effect in others.

Most medical treatments don't have dramatic effects, however, and unless care is taken to avoid biased comparisons, dangerously mistaken conclusions about the effects of treatment may result.

Comparisons should involve groups of people who were given the different treatments at more or less the same time

Comparing treatments given today with treatments given in the past (historical controls) only rarely provides a trustworthy basis for a fair test because relevant factors other than the treatments themselves change over time.

In the 1970s, Stuart Pocock, a British medical statistician, demonstrated this using evidence from a series of fair tests of treatments for cancer in which it was common practice to include the same control treatment in consecutive controlled trials. This meant that it was possible to compare the death rates of two groups of similar patients given the same treatment at different times. One would have expected little difference in the death rates associated with the same treatment given to apparently similar patients at different points in time. In fact, the differences observed ranged from a 46% lower to a 24% higher mortality, and in four of these comparisons, the differences using these historical controls were unlikely to be explained by chance. Presumably, although the patients and the treatments in the comparisons were apparently identical, there must have been subtle unrecognised or unmeasured differences in either the patients or the treatments given to them. In the light of this demonstration of the untrustworthiness of treatment comparisons using historical controls, the author wrote that

> *"Such marked evidence of differences between trials indicates that any comparison of treatments not within a [randomized] control trial must be deemed highly suspect".*

Comparing treatments in crossover tests in individual patients

Sometimes giving different treatments at more or less the same time may involve giving patients different treatments one after the other – a so-called crossover test. Sometimes this is done in a single patient – a so-called N-of-1 trial.

An early example of a crossover test was reported in 1786 by 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. (He didn't find any advantage for the expensive rhubarb!)

Treatment comparisons within individual patients have their place when their condition returns after stopping treatment. There are many circumstances in which this doesn't apply, however. For example, it is usually impossible to compare different surgical operations in this way, or treatments given for conditions that get progressively worse over time.

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. A Persian physician, Abu Bakr Muhammad ibn Zakariyaʾ Al-Razi (known as Rhazes, in Latin), recognized this more than a thousand years ago. 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.

If we are to know whether or not a treatment really does make us better, we need comparisons of the treatment with 'nature' or with other treatments that are fair tests. If these comparisons are to be fair, they must address genuine uncertainties, avoid biases and the play of chance, and be interpreted carefully.

In more depth

The James Lind Library 1.2 Why treatment comparisons are essential (http://www.jameslindlibrary.org/essays/1-2-why-treatment-comparisons-are-essential/)

Systematic reviews of methodology:

- Clarke M, Loudon K (2011). Effects on patients of their healthcare practitioner's or institution's participation in clinical trials: a systematic review. Trials 12:16
- Hróbjartsson A, Gøtzsche PC (2010). Placebo interventions for all clinical conditions. Cochrane Database of Systematic Reviews (1):CD003974
- Vist GE, Bryant D, Somerville L, Birminghem T, Oxman AD (2008). Outcomes of patients who participate in randomized controlled trials compared to similar patients receiving similar interventions who do not participate. Cochrane Database of Systematic Reviews (3):MR000009

In the 10th century, the Persian physician Abu Bakr Muhammad ibn Zakariya Al-Razi (Rhazes) recognized the need for an untreated control group when assessing the effects of a treatment.

al-Razi (10th century CE; 4th century AH)

Kitab al-Hawi fi al-tibb [The comprehensive book of medicine].

1.3 Why treatment comparisons must be fair

Untrustworthy treatment comparisons are those in which biases, or the play of chance, or both result in misleading estimates of the effects of treatments. Fair treatment comparisons avoid biases and reduce the effects of the play of chance.

It is not only failure to test theories about treatments in practice that has caused preventable tragedies. They have also occurred because the tests used to assess the effects of treatments have been unreliable and misleading. In the 1950s, theory and poorly controlled tests yielded unreliable evidence suggesting that diethylstilboestrol (DES) helped pregnant women who had previously had miscarriages and stillbirths. Although fair tests suggested that DES was useless, theory and 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 for the women and their children, who experienced infertility and cancers as a result. The lesson is that a treatment that has not been reliably shown to be useful should not be promoted.

Problems resulting from inadequate tests of treatments continue to occur. Again, because of unreliable evidence and aggressive marketing, millions of women were persuaded to use hormone replacement therapy (HRT). It was claimed that, not only could it reduce unpleasant menopausal symptoms, but also the chances of having heart attacks and strokes. When these claims were assessed in fair tests, the results showed that in women over 60, 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.

These examples of the need for fair tests of treatments are a few of many 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.

In more depth

The James Lind Library 1.3 Why treatment comparisons must be fair (http://www.jameslindlibrary.org/essays/1-3-why-treatment-comparisons-must-be-fair/)

Systematic reviews of methodology:

- Anglemyer A, Horvath HT, Bero L (2014). Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database of Systematic Reviews (4):MR000034
- Dechartres A, Trinquart L, Faber T, Ravaud P (2016). Empirical evaluation of which trial characteristics are associated with treatment effect estimates. Journal of Clinical Epidemiology 77:24-37

Section 2: Avoiding biased treatment comparisons

What are biases? Biases in tests of treatments are those influences and factors that can lead to conclusions about treatment effects that are systematically different from the truth.

Sometimes treatments have dramatic effects. 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. Such striking effects are rare, however. Usually, treatment effects are more modest, but nevertheless well worth knowing about, for example, using aspirin to reduce a person's risks after having a heart attack.

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 such moderate but important effects 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. Many kinds of biases can distort the results of health research. This book considers design bias, allocation bias, co-intervention bias, observer bias, analysis bias, biases in assessing unanticipated effects, reporting biases, biases in systematic reviews, and research biases and fraud.

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.

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.

In more depth

The James Lind Library 2.0 Avoiding biased treatment comparisons (http://www.jameslindlibrary.org/essays/2-0-avoiding-biased-treatment-comparisons/)

2.1 Why comparisons must address genuine uncertainties

The design of treatment research often reflects commercial and academic interests; ignores relevant existing evidence; uses comparison treatments known in advance to be inferior; and ignores the needs of users of research results (patients, health professionals and others).

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 that is likely to harm them.

For example, fair tests have been done to assess whether antibiotics (compared with inactive placebos) reduce the risk of people dying after bowel surgery. The first fair test was reported in 1969, but the results of this small study left uncertainty about whether antibiotics were useful. Quite properly, this uncertainty was addressed in further tests.

As the evidence accumulated, however, it became clear that antibiotics reduce the risk of death after surgery, yet researchers continued to do additional studies. 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.

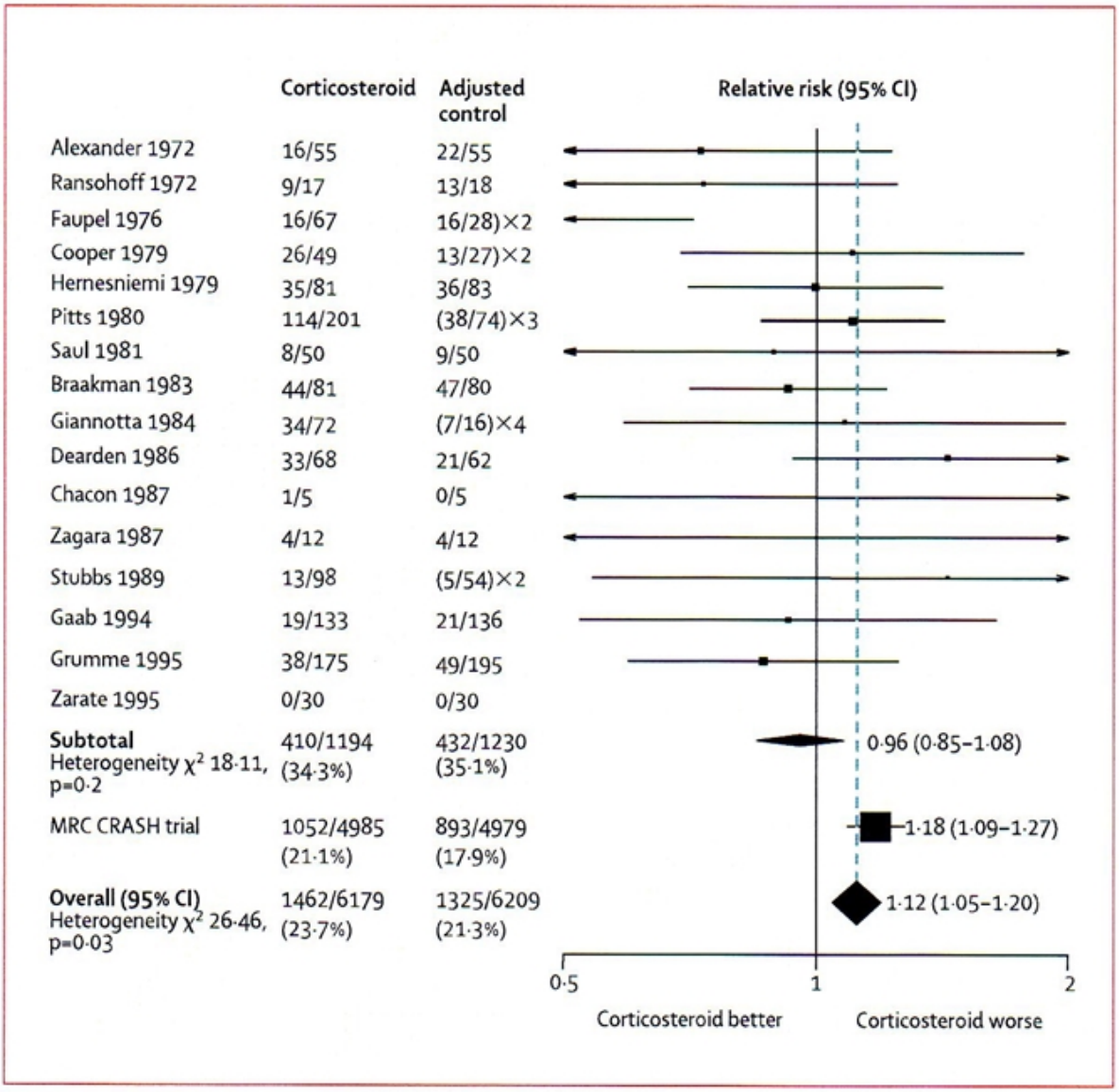


Figure 5: **Updated meta-analysis of effect of corticosteroids on death after head injury**

Adding the data from the large, prospective randomized trial (MRC CRASH) revealed that corticosteroid treatment in preterm infants was harmful, in spite of the large number of small-scale studies that had reported unclear results.

Crowley P, Chalmers I, Keirse MJNC (1990). The effects of corticosteroid administration before preterm delivery: an overview of the evidence from controlled trials. British Journal of Obstetrics and Gynaecology 97:11-25.

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 drug given to people experiencing a stroke. They found no evidence to support its increasing use in practice. 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.

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. These studies are deliberately designed to be unfair tests of treatments. This can be done by 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). It can also result from following up patients for too short a time (and missing delayed effects of treatments), and by using outcome measures ('surrogates') that have little or no correlation with the outcomes that matter to patients.

It may come as a surprise that the research ethics committees established to ensure that research is ethical have done so little to influence this research malpractice. Most such 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. The failure of research ethics committees to protect patients and the public efficiently in this way emphasizes the importance of improving general knowledge about the characteristics of fair tests of medical treatments.

In more depth

The James Lind Library 2.1 Why comparisons must address genuine uncertainties (http://www.jameslindlibrary.org/essays/2-1-why-comparisons-must-address-genuine-uncertainties/)

Systematic reviews of methodology:

- Clarke M, Brice A, Chalmers I (2014). Accumulating research: a systematic account of how cumulative meta-analyses would have provided knowledge, improved health, reduced harm and saved resources. PLoS ONE 9(7):e102670
- Martínez García L, Pardo-Hernandez H, Superchi C, Niño de Guzman E, Ballesteros M, Ibargoyen Roteta N, McFarlane E, Posso M, Roqué I Figuls M, Rotaeche Del Campo R, Sanabria AJ, Selva A, Solà I, Vernooij RWM, Alonso-Coello P (2017). Methodological systematic review identifies major limitations in prioritization processes for updating. Journal of Clinical Epidemiology 86:11-24
- Tricco AC, Tetzlaff J, Sampson M, Fergusson D, Cogo E, Horsley T, Moher D (2008). Few systematic reviews exist documenting the extent of bias: a systematic review. Journal of Clinical Epidemiology 61(5):422-34

2.2 The need to compare like with like in treatment comparisons

Allocation bias results when treatment comparisons fail to ensure that, apart from the treatments being compared, 'like will be compared with like'.

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.

It is rarely possible to be completely confident that groups of people assembled in the past who have been given one treatment are comparable in all the respects that matter with people who have more recently received a treatment. This is the case even if some information about the patients who have received these treatments is available (such as their ages, or their 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.

In 1747, James Lind, a Scottish naval surgeon faced with uncertainty about which of many proposed treatments for scurvy to use, compared six of them in a prospective controlled trial.

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.

The 18th century surgeon William Cheselden was aware of the 'dissimilar groups' problem 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: 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.

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 set up in such a way that they 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.

One hundred years after Lind, the British physician Thomas 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 was responsible, he used alternation to decide which boys would receive and which would not receive belladonna. During the first half of the 20th century, there were many examples of treatment comparison groups being assembled using alternation, or by drawing lots (for example, by using dice, coloured beads, or random sampling numbers). This 'random allocation' is the only, albeit crucially important, feature of the category of fair tests referred to as 'randomized'.

Casting or drawing lots is a time-honoured way of making fair decisions, and when used in tests of treatments, these methods help to ensure that comparison

groups are composed of similar types of people. Known and measured factors of importance, like age, can be checked. Moreover, unmeasured factors that may influence recovery from illness (such as diet, occupation, and anxiety) can be expected to balance out on average.

It's possible to undermine random allocation if the researchers in a study know which treatment a particular patient is going to get when participants are recruited to the trial. For example, this knowledge might influence a doctor's decision on whether or not to offer the patient the chance to participate in the trial.

Strict adherence to unbiased allocation schedules is required to avoid biased creation of treatment comparison groups. 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 assembly of comparison groups.

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 by 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.

Failure to do this can result in unfair tests of treatments. Take, for example, two very different ways of treating people experiencing dizzy spells because of partially blocked blood vessels supplying their brains. Treatment for this condition can be important because people experiencing such dizzy spells are at increased risk of suffering a stroke. One of the treatments involves taking aspirin to stop the blockage from getting worse; the other involves a surgical operation to try to remove the blockage in the blood vessel.

A fair comparison of these two approaches to treating dizzy spells would involve creating two groups of people using an unbiased allocation method (like random allocation), and then treating patients in one group with surgery and patients in the other group with aspirin. The comparison would thus begin by comparing two groups of patients who were alike, and go on to compare their respective frequencies of subsequent strokes. But if the frequency of strokes in the surgically treated group was only recorded among patients who had survived the immediate effects of the operation, the important fact that the operation itself can cause stroke and death would be missed. This would result in an unfair comparison of the two treatments, resulting in a biased and misleadingly optimistic picture of the effects of the operation. 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, without exceptions, and in the groups to which they were originally assigned. If this principle is not observed, people may receive biased information about the overall effects of treatments.

In more depth

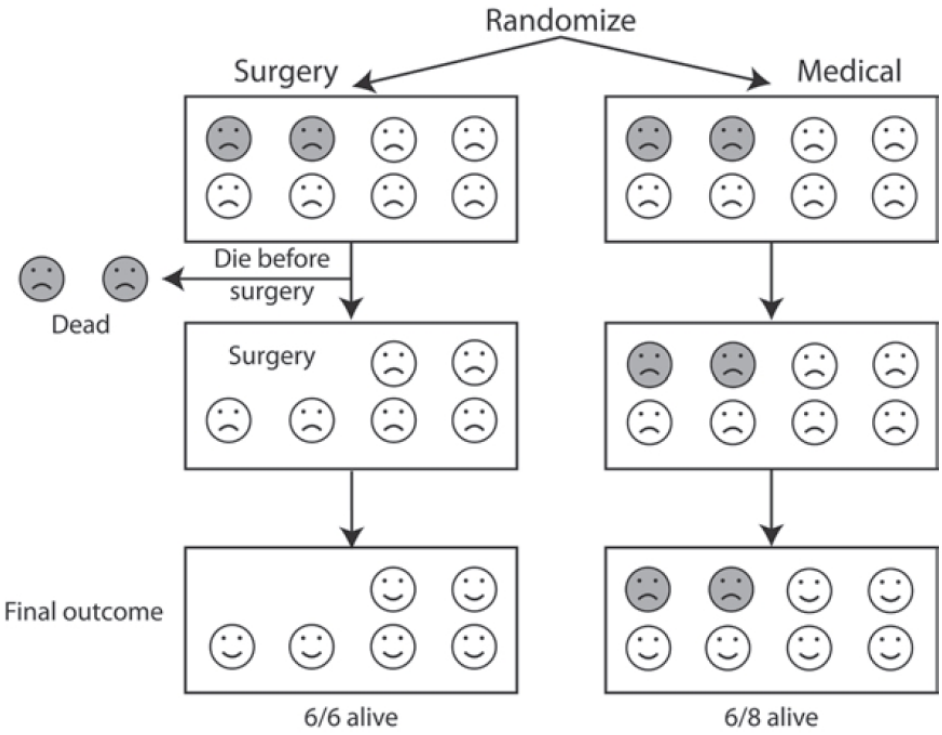
The James Lind Library 2.2 The need to compare like-with-like in treatment comparisons (http://www.jameslindlibrary.org/essays/2-2-the-need-to-compare-like-with-like-in-treatment-comparisons/)

Systematic reviews of methodology:

- Bello S, Wei M, Hilden J, Hróbjartsson A (2016). The matching quality of experimental and control interventions in blinded pharmacological randomised clinical trials: a methodological systematic review. BMC Medical Research Methodology 16:18
- MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AM (2000). A systematic review of comparisons of effect sizes derived

from randomised and non-randomised studies. Health Technology Assessment 4(34):1-154.

- Odgaard-Jensen J, Vist GE, Timmer A, Kunz R, Akl EA, Schünemann H, Briel M, Nordmann AJ, Pregno S, Oxman AD (2011). Randomisation to protect against selection bias in healthcare trials. Cochrane Database of Systematic Reviews (4):MR000012
- Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG, Gluud C, Martin RM, Wood AJ, Sterne JA (2008). Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ 336(7644):601-5.



Any change to the composition of the comparison groups after randomization can bias the results. This is one reason why patients should be analyzed in the groups to which they were originally assigned.

Evans I, Thornton H, Chalmers I & Glasziou P (2011).
Testing Treatments, 2nd Edition. London: Pinter and Martin

2.3 Why avoiding differences between treatments allocated and treatments received is important

Knowledge of which treatments have been received by which study participants can affect adherence to assigned treatments and result in biased use of other treatments (co-interventions). These biases can be reduced by using placebos to conceal the identities of the treatments being compared.

Fair tests of medical treatments must 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 considered when interpreting the results of treatment comparisons. One of the reasons that placebos were introduced in the evolution of fair tests of medical treatments was to reduce departures from the test's intended treatments.

Things may go astray even in placebo-controlled trials, however. During World War II, people suffering from colds were given a solution of a drug called patulin and compared with other people given only the fluid in which the drug had been dissolved. Analysis of the results didn't reveal any beneficial effects of the drug, but a concern then emerged that the liquid used to dissolve the drug might have inactivated the drug. 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!

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

Husband and wife Philip and Ruth d'Arcy Hart, along with colleagues in the UK Medical Research Council (MRC) designed, conducted and reported the MRC's first well controlled multicentre trial, during World War 2. They are pictured here at their home in 2003, 60 years after working together on the patulin trial and the month before Philip's 103rd birthday.

Medical Research Council (1944)

Clinical trial of patulin in the common cold. Lancet 2:373-5.

be offered to them; patients may reject the treatments allocated to them, or not take them as intended; doses of the treatment that differed 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 side effects of treatment developed.

For these reasons, interpretations of fair tests must consider the possibility that treatments received were not those intended, or that additional treatments were given to patients in one treatment comparison group than to those in another. If discrepancies between intention and practice have occurred, it is important to consider the possible implications for interpreting the evidence.

In more depth

The James Lind Library 2.3 Why avoiding differences between treatments allocated and treatments received is important (http://www.jameslindlibrary.org/essays/2-3-why-avoiding-differences-between-treatments-allocated-and-treatments-received-is-important/)

No methodology reviews were identified for this section.

2.4 The need to avoid differences in the way treatment outcomes are assessed

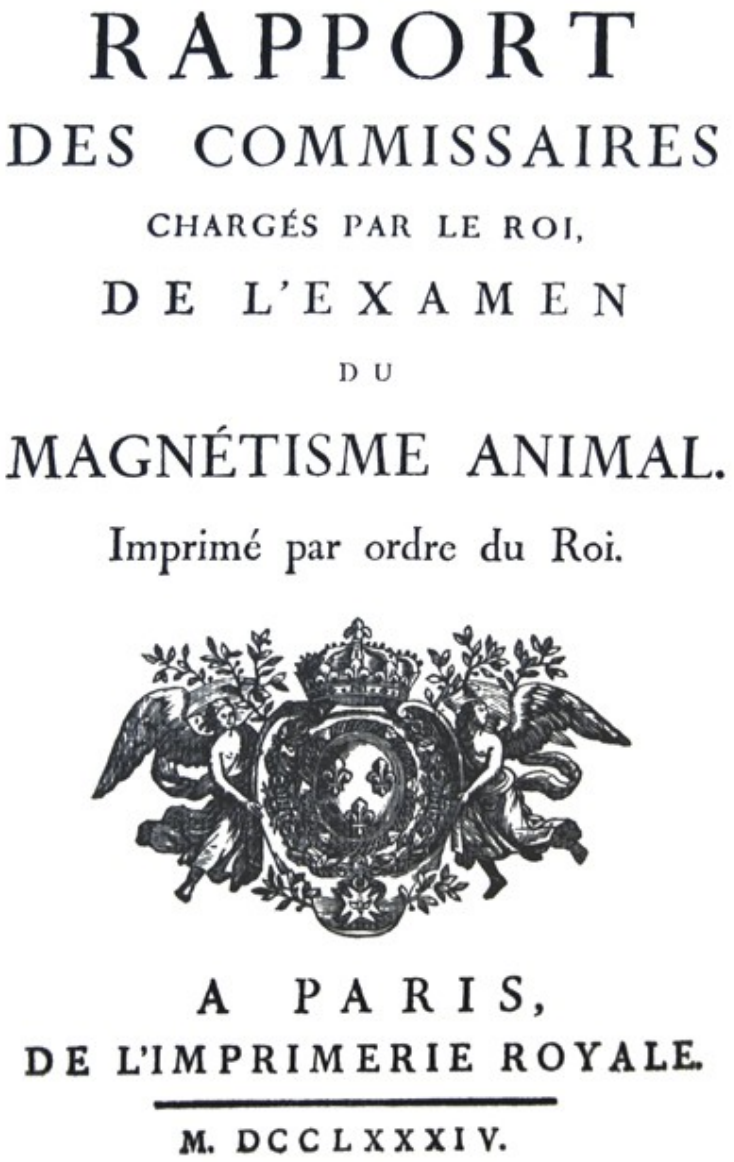
Biased treatment outcome assessment can result if the people receiving or providing care, or those assessing treatment outcomes, know which participants have received which treatments. It is sometimes possible to conceal which treatments have been received by using placebos and in other ways.

Using blinding to reduce bias when assessing treatment outcomes

For some outcomes used to assess treatment – 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, often involves subjectivity (as with patients' symptoms). The biases that lead to these misperceptions of symptom relief or exacerbation are termed observer biases. They cause a problem, particularly when people have special 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. 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' (sometimes called 'masking', especially in tests of treatments involving eyes) of patients and doctors is a desirable element of fair tests. What appears to have been the earliest blinded assessment of a treatment was done by a commission of inquiry appointed by Louis XVI in 1784 to investigate Anton Mesmer's claims about the effects of 'animal magnetism'. The Commission assessed whether the purported effects of this new healing method were due to any 'real' force, or due to the 'illusions of


RAPPORT
DES COMMISSAIRES
CHARGÉS PAR LE ROI,
DE L'EXAMEN
DU
MAGNÉTISME ANIMAL.
Imprimé par ordre du Roi.
A PARIS,
DE L'IMPRIMERIE ROYALE.
M. DCC LXXXIV.

Antoine Lavoisier, Benjamin Franklin and others in Paris assessed the effects of Franz Mesmer's animal magnetism by blindfolding patients to whom it was applied.

Commission Royale. Bailly A (1784)

Rapport des commissaires chargés par le Roi, de l'examen du magnétisme animale [Report of the Commissioners required by the King to examine animal magnetism]. Imprimé par ordre du Roi. Paris: A Paris, de L'Imprimerie Royale.

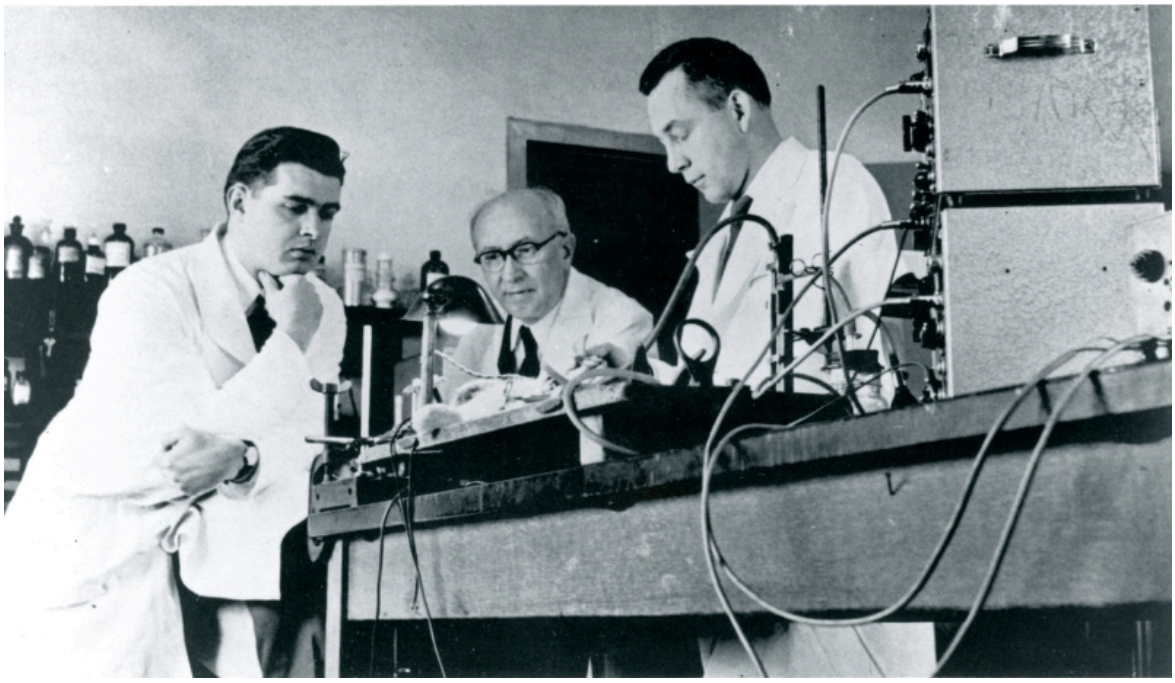
the mind'. Blindfolded people were told that they were receiving or not receiving magnetism. In fact, at times, the reverse was happening. The people being studied felt the effects of 'animal magnetism' only when they were told they were receiving the treatment, but not otherwise.

Using placebos to achieve blinding

A few years later, John Haygarth of Bath conducted an experiment using a sham device (a placebo) to achieve blinding. An American quack – Elisha Perkins - claimed that his small metal rods ('tractors') 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*', Haygarth reported how, in patients who were unaware of the details of his evaluation, he compared the metal tractors with wooden 'tractors' that looked identical to them (what we would consider 'placebo tractors'). Haygarth was unable to detect any benefit of Perkins' metal tractors.

It was not until much later that a more sceptical attitude in mainstream medicine led to a recognition that there was a more general need to adopt blinded assessment and placebos to assess the validity of its claims.

Inspired principally by pharmacologists, German researchers gradually adopted blind assessment. For example, in 1918, Adolf Bingel reported that he had tried to make his comparison of two different treatments for diphtheria "as objective as possible". 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" and noted that "their judgment was thus completely without prejudice", although it may still have been contradictory or inconsistent. He also noted how he used a quantitative analysis because "I am keen to see my observations checked independently, and most warmly recommend this 'blind' method for the purpose"; and, in fact, he did not detect a difference between the two treatments.



Harry Gold (centre, in a photograph from 1955) and Ella Hediger had earlier described a "blind test" to ensure that anaesthetists' judgments about the effectiveness of anaesthesia would not be biased by their knowledge of which type was being used.

Hediger E, Gold H (1935)
USP ether from large drums and ether from small cans labelled 'For Anesthesia'. JAMA 104:2244-2448.

Blind assessment in the modern English-speaking world first began when researchers were influenced by the German tradition, as well as by an indigenous 'quackbuster' movement that used masked treatment outcome assessment.

By the 1930s, anglophone researchers had taken up the use of 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, including the study of patulin mentioned in the last chapter. It would have been very difficult to interpret their results had what would later be termed 'double blinding' not been used to prevent patients and doctors knowing which patients had received the new drugs and which had received placebos. In the 1960s, 'double dummies' were introduced when two very different treatments – an injection and a pill, for example – were being compared. In these studies, injected drugs are compared with injected placebo while a swallowed tablet is compared with an identical-looking placebo tablet.

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, or with no treatment. 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 mid-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, 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 a patient who had been treated with streptomycin or one treated with bed rest alone.

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

In more depth

The James Lind Library 2.4 The need to avoid differences in the way treatment outcomes are assessed (http://www.jameslindlibrary.org/essays/2-4-the-need-to-avoid-differences-in-the-way-treatment-outcomes-are-assessed/)

Systematic reviews of methodology:

- Boutron I, Estellat C, Ravaud P (2005). A review of blinding in randomized controlled trials found results inconsistent and questionable. Journal of Clinical Epidemiology 58(12):1220-6
- Hróbjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S (2014). Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. International Journal of Epidemiology 43(4):1272-83
- Hróbjartsson A, Gøtzsche PC (2010). Placebo interventions for all clinical conditions. Cochrane Database of Systematic Reviews (1):CD003974
- Hróbjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, Ravaud P, Brorson S (2012). Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ 344:e1119
- Hróbjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, Ravaud P, Brorson S (2013). Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors. CMAJ 185(4):E201-11

- Hróbjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Rasmussen JV, Hilden J, Boutron I, Ravaud P, Brorson S (2014). Observer bias in randomized clinical trials with time-to-event outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. International Journal of Epidemiology 43(3):937-48
- Ndounga Diakou LA, Trinquart L, Hróbjartsson A, Barnes C, Yavchitz A, Ravaud P, Boutron I (2016). Comparison of central adjudication of outcomes and onsite outcome assessment on treatment effect estimates. Cochrane Database of Systematic Reviews (3):MR000043

2.5 Bias introduced after looking at study results

Biases can be introduced when knowledge of the results of studies influences analysis and reporting decisions, for example, when studies stop earlier than planned, or if there is bias in the selection of the treatment outcomes analyzed.

Biased analyses before the planned end of a study

Biases can result from informal observations of accumulating data and from doing formal statistical analyses before the full study results are known. As an example of the former, if researchers are collecting or observing outcomes because they are providing treatment for participants in a study, they may have developed a sense of which patients are doing particularly well or badly. This may lead them to alter the planned analyses by changing their views about what constitutes the "most important" outcomes, or by dredging through the data for what might be regarded as significant differences. The risk of these biases can be reduced if the researchers and the practitioners are kept blinded to which treatment was allocated to each participant.

When study results are being analysed more formally, different problems can arise. While a study is still in progress, accumulating results might be examined to see if there is clear evidence of benefit or harm for one of the treatments being compared, and so make it unethical to continue the study. On the other hand, it may become clear that the hoped-for effect is unlikely to be achieved in the study and that it would therefore be better to stop the study for 'futility'. These early stopping decisions can lead to bias when the interim results happen to be 'high' or 'low' simply by chance. The danger of bias is greater if there are vested interests in stopping the study and presenting interim results as if they were final results.

Systematic reviews of the impact of stopping trials earlier than envisaged have shown how early stopping might bias conclusions about the effects of treatments. Interim analyses may have shown implausibly large treatment

Abstract

***Objective*—To determine whether inappropriate subgroup analysis together with chance could change the conclusion of a systematic review of several randomised trials of an ineffective treatment.**

***Design*—44 randomised controlled trials of DICE therapy for stroke were performed (simulated by rolling different coloured dice; two trials per investigator). Each roll of the dice yielded the outcome (death or survival) for that "patient." Publication bias was also simulated. The results were combined in a systematic review.**

***Setting*—Edinburgh.**

***Main outcome measure*—Mortality.**

***Results*—The "hypothesis generating" trial suggested that DICE therapy provided complete protection against death from acute stroke. However, analysis of all the trials suggested a reduction of only 11% (SD 11) in the odds of death. A predefined subgroup analysis by colour of dice suggested that red dice therapy increased the odds by 9% (22). If the analysis excluded red dice trials and those of poor methodological quality the odds decreased by 22% (13, 2P=0·09). Analysis of "published" trials showed a decrease of 23% (13, 2P=0·07) while analysis of only those in which the trialist had become familiar with the intervention showed a decrease of 39% (17, 2P=0·02).**

***Conclusion*—The early benefits of DICE therapy were not confirmed by subsequent trials. A plausible (but inappropriate) subset analysis of the effects of treatment led to the qualitatively different conclusion that DICE therapy reduced mortality, whereas in truth it was ineffective. Chance influences the outcome of clinical trials and systematic reviews of trials much more than many investigators realise, and its effects may lead to incorrect conclusions about the benefits of treatment.**

In the DICE 1 study, Carl Counsell and colleagues showed how some fairly simple, but common, manipulations to the analyses of 44 unbiased randomised trials simulated by rolling dice could end up with a biased conclusion that specific dice rolled by specific people reduced deaths following a stroke by 39%.

Counsell CE, Clarke MJ, Slattery J, Sandercock PAG (1994)

The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis?

BMJ 309:1677-1681.

effects, particularly when the number of events was small. On average, trials that stopped early were found to be more favourable to treatments than those from similar trials that did not stop early.

One way to avoid biases that might arise if the researchers themselves are responsible for these interim decisions is to establish an independent Data Monitoring Committee which looks at the interim analyses in confidence, acting as an oversight group for the study.

Biased analyses after the planned end of a study

Things can get even more problematic after the full study results are known. Biased changes may then be made because analysts know how these changes would favour one or other of the treatments compared. If biased changes occur between the collection of the study data and their eventual reporting, readers of the published results will be unaware of them, and risk being misled.

At the end of a study, changes to the analyses after looking at the results can lead to bias through:

- changes in the designated primary outcome, or in how outcomes are defined or combined in composite outcomes;
- introduction of subgroup analyses, in which different groups of participants are analysed separately, perhaps to highlight the presence or absence of benefit in certain types of person or setting;
- selective reporting of particular outcomes, analyses or treatment comparisons; and
- changes to the statistical techniques, such as the introduction of adjustments for differences in baseline characteristics of the participants which had not been pre-planned or pre-specified.

The potential impact of some of these biases has been studied, and some of these studies have themselves been considered in systematic reviews. These have shown that discrepancies in analyses between publications and other study

documentation are common, but not discussed in the trial reports. For example, prespecified primary outcomes were changed or introduced in about half of the studies analysed by these reviews.

In more depth

The James Lind Library Bias introduced after looking at study results (http://www.jameslindlibrary.org/essays/2-5-bias-introduced-after-looking-at-study-results/)

Systematic reviews of methodology:

- Bassler D, Briel M, Montori VM, Lane M, Glasziou P, Zhou Q, Heels-Ansdell D, Walter SD, Guyatt GH; STOPIT-2 Study Group (2010). Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. JAMA 303:1180-7
- Dwan K, Altman DG, Cresswell L, Blundell M, Gamble CL, Williamson PR (2011). Comparison of protocols and registry entries to published reports for randomised controlled trials. Cochrane Database of Systematic Reviews (1):MR000031
- Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, Lacchetti C, Leung TW, Darling E, Bryant DM, Bucher HC, Schunemann HJ, Meade MO, Cook DJ, Erwin PJ, Sood A, Sood R, Lo B, Thompson CA, Zhou Q, Mills E, Guyatt GH (2005). Randomized trials stopped early for benefit: a systematic review. JAMA 294:2203-9

2.6 Reducing biases in judging unanticipated possible treatment effects

Important unanticipated effects of treatments are often first suspected by people using or prescribing treatments. As with anticipated effects of treatments, steps must be taken to reduce biases and the play of chance in assessing suspected unanticipated effects.

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 for marketing – 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.

Rare treatment effects, or those that take some time to develop, will not be discovered until studies have lasted long enough or 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 targeted by the treatment. These differences may modify treatment effects, and new, unanticipated effects may emerge.

Detection and verification of unanticipated effects of treatments, 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.

If the unanticipated effect of a treatment is very striking and occurs quite often after the treatment has been used, it may be noticed spontaneously by health

THALIDOMIDE AND CONGENITAL ABNORMALITIES

SIR,—Congenital abnormalities are present in approximately 1·5% of babies. In recent months I have observed that the incidence of multiple severe abnormalities in babies delivered of women who were given the drug thalidomide ('Distaval') during pregnancy, as an antiemetic or as a sedative, to be almost 20%.

These abnormalities are present in structures developed from mesenchyme—i.e., the bones and musculature of the gut. Bony development seems to be affected in a very striking manner, resulting in polydactyly, syndactyly, and failure of development of long bones (abnormally short femora and radii).

Have any of your readers seen similar abnormalities in babies delivered of women who have taken this drug during pregnancy?

Hurstville, New South Wales. W. G. MCBRIDE.

*** In our issue of Dec. 2 we included a statement from the Distillers Company (Biochemicals) Ltd. referring to "reports from two overseas sources possibly associating thalidomide ('Distaval') with harmful effects on the fœtus in early pregnancy". Pending further investigation, the company decided to withdraw from the market all its preparations containing thalidomide.—ED.L.

McBride WG (1961)

Thalidomide and congenital abnormalities. Lancet 2:1358.

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 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.

When such striking relationships are noticed, they often turn out to be confirmed to be real, unanticipated effects of treatment. 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 possible treatment effects be investigated by further analysis or follow-up of people who had been randomly allocated to the treatment before it was given to them. The challenge is therefore to assemble unbiased comparison groups in other ways (for example, retrospective comparison), often using information collected routinely during health care.

In these studies, it is helpful that the suspected effect was 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 suspected effect when people were being selected differentially for treatment. This is most likely when the unanticipated effect is a different condition or disease from the condition or disease for which the treatment was prescribed.

For example, when hormone replacement therapy (HRT) was introduced for treating menopausal symptoms in the 1960s a woman's risk of developing venous thrombosis was unlikely to have been considered 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. This created the circumstances for fair tests, and these showed that HRT increases the risk of venous thrombosis.

When a suspected unanticipated effect concerns a treatment and its relationship to a common health problem (such as heart attack) but does not occur very often with the treatment, large-scale surveillance of people receiving the treatment is needed to detect the unanticipated effect. For example, although some people thought in the 1960s that aspirin might reduce the risk of heart attack, most people would have thought that the theory was highly implausible. The breakthrough came when a large study was done in Boston, USA, 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. These findings were consistent with those of a subsequent fair test, in which people were allocated at random to receive or not receive aspirin after heart attack.

The ground rules for detecting and investigating unanticipated effects of treatments were first set out clearly in the late 1970s. They drew on the collective experience of investigating unanticipated effects which had accumulated following the thalidomide disaster. 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.

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 the use of HRT and heart disease, stroke and breast cancer.

In more depth

The James Lind Library 2.6 Reducing biases in judging unanticipated possible treatment effects (http://www.jameslindlibrary.org/essays/2-6-reducing-biases-in-judging-unanticipated-possible-treatment-effects/)

Systematic reviews of methodology:

- Allen EN, Chandler CIR, Mandimika N, Leisegang C, Barnes K (2018). Eliciting adverse effects data from participants in clinical trials. Cochrane Database of Systematic Reviews (1):MR000039
- Golder S, Loke YK (2010). Sources of information on adverse effects: a systematic review. Health Information and Libraries Journal 27(3):176-90.

2.7 Dealing with biased reporting of the available evidence

Biased reporting of research occurs when the direction or statistical significance of results influences whether and how research is reported.

Avoiding biased comparisons entails using systematic reviews to identify and take account of all the relevant, reliable evidence. This is challenging in many ways, particularly because what evidence is available might be influenced by biased decisions 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'.

These reporting biases have been recognized for centuries. In 1792, for example, the Scottish physician John Ferriar stressed the importance of recording treatment failures as well as treatment successes. This principle was reiterated in an editorial published in the *Boston Medical and Surgical Journal* just over a century later.

There is now a large amount 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 their reports by journal editors, among other reasons, because of vested interests. Recent research has also revealed an additional problem: if the observed effects of treatments on some of the outcomes studied don't support the (hoped-for) conclusions of researchers, these data sometimes don't get reported either.

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.

"The whole tribe of diuretics is acknowledged to be uncertain,
and often to disappoint the most rational expectations.
Practitioners are therefore perpetually in search of new remedies belonging to this class, and
are too apt to over-rate the value of such discoveries"

Ferriar J (1792)

Medical histories and reflections. Vol 1. London: Cadell and Davies.

In 1993, Alan 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 a new anti-arrhythmic drug 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'".

Reporting biases tend to lead to conclusions that medical treatments are more useful and freer of side effects than they are in fact. As a consequence, they can result in unnecessary suffering and death, and in wasted resources spent on ineffective or dangerous treatments.

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. 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. Fair testing of treatments will remain compromised as long as this form of research misconduct is tolerated by governments and others who should be protecting the interests of the public.

Among others, the World Health Organization has coordinated solutions to address the problem of unidentifiable research and publication bias. First, it established standards for the registration and exchange of data for the registration of trials. Second, it proposed registration of research protocols in databases that fulfil the above standards, before patient recruitment starts. Finally, it established a freely accessible portal that collates information from national and regional registers, making it easier for people to learn about anticipated, ongoing and finished research protocols.

Using drug research done during the late 1970s, Elina Hemminki, a Finnish health services researcher, showed that studies of new drugs submitted to licensing authorities were less likely to be published subsequently if they had looked for adverse effects.

Hemminki E (1980).

Study of information submitted by drug companies to licensing authorities. BMJ 280:833-6.

Although registration addresses the problem of unidentifiable research by letting people know what research is planned, ongoing or completed, it is only by providing the results of this research that publication bias can be overcome. In recent years, some research registries have started to include study findings, but uptake of this option by researchers remains incomplete and inadequate as a means of ensuring that the findings of all trials are publicly available.

In more depth

The James Lind Library 2.7 Dealing with biased reporting of the available evidence (http://www.jameslindlibrary.org/essays/2-7-dealing-with-biased-reporting-of-the-available-evidence/)

Systematic reviews of methodology:

- Chiu K, Grundy Q, Bero L (2017). 'Spin' in published biomedical literature: A methodological systematic review. PLoS Biology 15(9):e2002173
- Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, Decullier E, Easterbrook PJ, Von Elm E, Gamble C, Ghersi D, Ioannidis JP, Simes J, Williamson PR (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One 3(8):e3081
- Dwan K, Altman DG, Clarke M, Gamble C, Higgins JP, Sterne JA, Williamson PR, Kirkham JJ (2014). Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. PLoS Medicine 11(6):e1001666
- Dwan K, Gamble C, Williamson PR, Kirkham JJ; Reporting Bias Group (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias – an updated review. PLoS One 8(7):e66844
- Hopewell S, Clarke MJ, Stewart L, Tierney J (2007). Time to publication for results of clinical trials. Cochrane Database of Systematic Reviews (2):MR000011

John Simes proposed international registration of all clinical trials after he showed that conclusions about treatments for ovarian cancer differed depending on whether the results of unpublished trials had been taken into account.

Simes RJ (1986)
Publication bias: the case for an international registry of clinical trials. Journal of Clinical Oncology 4:1529-41.

- Hopewell S. Loudon K, Clarke MJ, Oxman AD, Dickersin K (2009). Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database of Systematic Reviews (1):MR000006
- Jones CW, Keil LG, Holland WC, Caughey MC, Platts-Mills TF (2015). Comparison of registered and published outcomes in randomized controlled trials: a systematic review. BMC Medicine 13:282
- Scherer RW, Meerpohl JJ, Pfeifer N, Schmucker C, Schwarzer G, von Elm E (2018). Full publication of results initially presented in abstracts. Cochrane Database of Systematic Reviews (11):MR000005
- Song F, Parekh-Bhurke S, Hooper L, Loke YK, Ryder JJ, Sutton AJ, Hing CB, Harvey I (2009). Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. BMC Medical Research Methodology 9:79.
- Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, Hing C, Kwok CS, Pang C, Harvey I (2010). Dissemination and publication of research findings: an updated review of related biases. Health Technology Assessments 14(8):1-193

2.8 Avoiding biased selection from the available evidence

Because single tests of treatments can be misleading, systematic reviews are used to identify, evaluate and summarize all the evidence relevant to addressing a specific question.

Biases can distort individual 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 Cochrane (formerly, The Cochrane Collaboration), making clear what measures will be taken to reduce biases.

These include specifying clearly:

- which questions 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 use in the review.

Different systematic reviews addressing what appears to be the same question about the effects of 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. In these circumstances it is important to judge which of the reviews are most likely to have been most successful in reducing biases.

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. 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.

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 we should not expect researchers, health professionals, patients and others assessing the effects of treatments to be immune.

In more depth

The James Lind Library 2.8 Avoiding biased selection from the available evidence (http://www.jameslindlibrary.org/essays/2-8-avoiding-biased-selection-from-the-available-evidence/)

Systematic reviews of methodology:

- Thaler K, Kien C, Nussbaumer B, Van Noord MG, Griebler U, Klerings I, Gartlehner G; UNCOVER Project Consortium (2015). Inadequate use and regulation of interventions against publication bias decreases their effectiveness: a systematic review. Journal of Clinical Epidemiology 68(7):792-802.

2.9 Recognizing researcher bias, sponsor bias and fraud

The commercial, academic or other vested interests of researchers and organizations tend to be reflected in the reports of treatment research in which they are involved.

In 1764, a Dr R James published the 6th edition of his book '*A dissertation on fevers and inflammatory distempers'*. In it, he claimed that his secret 'Fever Powder' was successful in treating "smallpox, yellow fever, slow fever and rheumatism". In support of his claims, he cited the testimonies of satisfied patients and a decline in the national mortality rate following the introduction of his miraculous 'cure-all'. 'Snake oil salesmen' like Dr James have probably been a feature of medical practice for as long as patients have looked to doctors and others to help them deal with health problems.

During the 19th century, the ground rules for testing treatment claims began to become clearer. Alternation began to be used to generate comparison groups and so ensure that 'like would be compared with like', and blinding became recognised as a way of reducing observer biases. For example, comparisons of homeopathic with orthodox medical treatments demonstrated not that homeopathy was effective, but that it was safer than the bleeding and purging being offered by mainstream doctors.

By the early years of the 20th century, a pharmaceutical industry had begun to emerge which was profit-driven, and thus tempted to take liberties with claims for its products and the use of data to support these. In 1917, Torald Sollmann, an American pharmacologist, set out the principles to be observed in testing treatments. He noted that "Those who collaborate with [commercial firms] should realize frankly that under present conditions they are collaborating, not so much in determining scientific value, but rather in establishing commercial value". Concerns about these sponsor and researcher biases – and sometimes

"Those who collaborate should realize frankly that under present conditions they are collaborating, not so much in determining the scientific value, but rather in establishing the commercial value of the article."

Sollmann T (1917)

The crucial test of therapeutic evidence. JAMA 69:198-199.

outright fraud – grew throughout the 20th century, fuelled increasingly by evidence going beyond anecdotes. Sponsor and researcher biases make active use of other biases in pursuit of their vested interests. Recognising and reducing research biases and outright fraud remains a substantial challenge.

In more depth

The James Lind Library 2.9 Recognizing researcher/sponsor biases and fraud (http://www.jameslindlibrary.org/essays/2-9-recognizing-researchersponsor-biases-and-fraud/)

Systematic reviews of methodology:

- Lexchin J, Bero LA, Djulbegovic B, Clark O (2003). Pharmaceutical industry sponsorship and research outcome and quality: systematic review. BMJ 326:1167-70
- Lundh A, Lexchin J, Mintzes B, Schroll JB, Bero L (2017). Industry sponsorship and research outcome. Cochrane Database of Systematic Reviews (2):MR000033

Section 3: Taking account of the play of chance

When treatments are compared, any differences in outcome events may simply reflect the play of chance. Increasing the number of events studied in research reduces the likelihood of being misled by 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 in which 4 people improved with a new treatment and 6 people improved with an older treatment. 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.

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 older treatment, it would be clear that the new treatment was indeed very likely to be worse than the older treatment.

The way to reduce the likelihood of being misled by the play of chance is thus to ensure that fair tests include sufficiently large numbers of people who experience the outcomes one hopes to influence.

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.

In more depth

The James Lind Library 3.0 Taking account of the play of chance (http://www.jameslindlibrary.org/essays/3-0-taking-account-of-the-play-of-chance/)

3.1 Recording and interpreting numbers in testing treatments

Numbers are needed to record the results of fair tests of treatments, and tables and graphs are used to describe the characteristics and experience of groups of patients, the treatment they have received, and quantitative estimates of treatment effects.

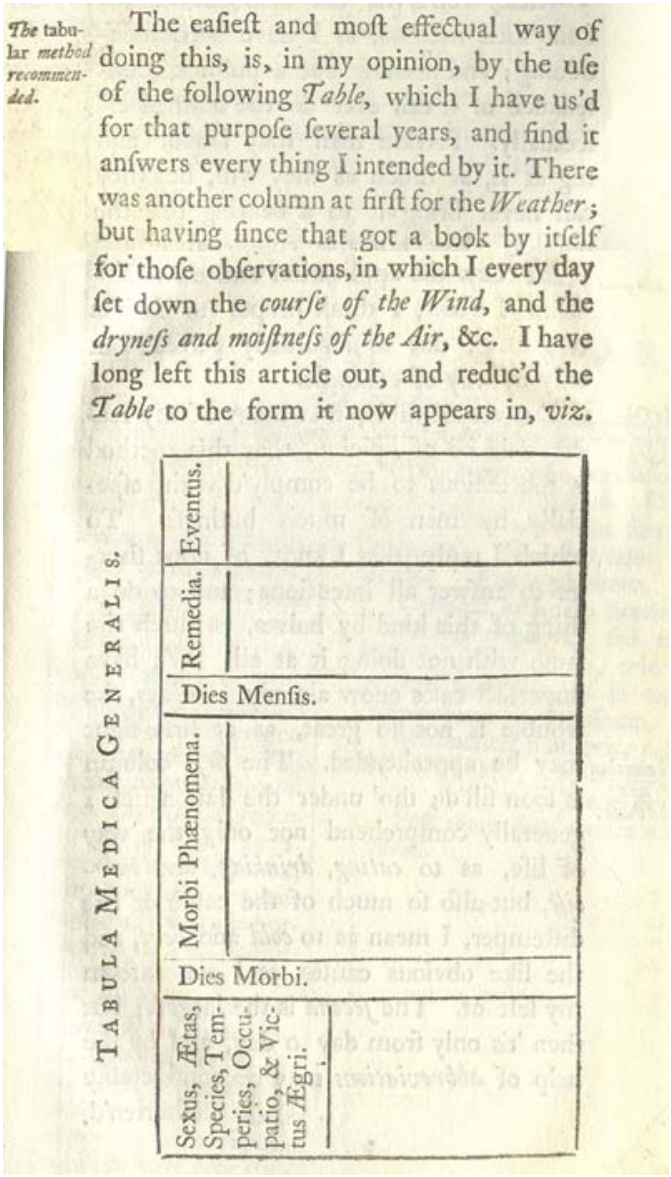
Using quantification in testing treatments

It was not until the early 18th century that numbers began to be used to assess the effects of medical treatments. Quantification began in the 1720s with comparisons of death rates following variolation (inoculation) against smallpox with the death rates associated with the disease itself. In 1732, Francis Clifton of London published a book entitled 'The state of physick, ancient and modern, briefly considered: with a plan for the improvement of it'. He pointed out that, instead of trying to assess the worth of therapies by whether they accorded with theories, physicians needed to base their judgements about the treatment effects they had observed on sufficiently large numbers of their own patients. Tables and graphs were used increasingly to present the numbers and statistics derived from such observations.

Replacing certainties with probabilities

What were the motives for quantifying and tabulating observations? A book by George Fordyce published in 1793 provides an initial answer: Its title was 'An attempt to improve the evidence of medicine', and it was published in the *Transactions of a Society for the Improvement of Medical and Chirurgical Knowledge*. Quantification of experience was aimed at "increasing the certainty of medicine," although he actually meant probability rather than certainty.

Among the issues eagerly debated in 18th century British medicine was that of certainty versus the slowly growing notion of statistical probability. In 1772,



The tabular method recommended. The eaſieſt and moſt effectual way of doing this, is, in my opinion, by the uſe of the following *Table*, which I have us'd for that purpoſe ſeveral years, and find it anſwers every thing I intended by it. There was another column at firſt for the *Weather*; but having ſince that got a book by itſelf for thoſe obſervations, in which I every day ſet down the *courſe of the Wind*, and the *dryneſs and moiſtneſs of the Air*, &c. I have long left this article out, and reduc'd the *Table* to the form it now appears in, *viz.*

TABULA MEDICA GENERALIS.

Sextus, Ætas, Species, Temperies, Occupatio, & Victus Ægri.	Morbi Phænomena.	Dies Morbi.	Remedia.	Eventus.	Dies Menſis.

At the beginning of a century during which British use of quantified, tabulated data became widespread, Francis Clifton designed tables to record illnesses, treatments, and outcomes.

Clifton F (1732)

The state of physick, ancient and modern, briefly considered:
with a plan for the improvement of it.

London, printed by W Bowyer, for John Nourse without Temple-Bar.

James Lind summarized the transition from belief in an absolute authority to reliance on relative statistics, but even these remained partial in his view. More outspokenly, John Haygarth calculated probabilities of escaping infection with 'continuous fever' or smallpox by counting the numbers of people who contracted the disease after contact with a patient. Using results "computed arithmetically by the doctrine of chances, according to the data", Haygarth indicated that immediate isolation of patients with smallpox and fever in specific wards in Chester was required.

Interpreting numbers

How did people judge whether treatment comparisons were trustworthy? For example, during debates about bloodletting for treating fevers around 1800, statistics were widely used on both sides. Besides the issue of honesty, the question of bias was raised – of the need to compare like with like. A writer in the *Edinburgh Medical and Surgical Journal* in 1813 stressed that, if one could assume the data to have been honestly assembled and presented by both sides, the only way out of the maze would be through further "extensive comparative experiments".

During the 19th century there was gradual recognition that it is important to record the extent of uncertainty associated with estimates of treatment differences. Jules Gavarret, a mathematically-inclined Parisian physician, pointed out the need to analyse treatment comparisons of sufficient size and to calculate the 'limits of oscillation' (variation) associated with statistical estimates of treatment differences. However, this practice did not really become widely adopted until the second half of the 20th Century.

In more depth

The James Lind Library 3.1 Recording and interpreting numbers in testing treatments (http://www.jameslindlibrary.org/essays/3-1-recording-and-interpreting-numbers-in-testing-treatments/)

In 1840, the French physician and statistician Jules Gavarret published a book on statistical analysis of treatment tests, stressing the importance of estimating uncertainty and calculating 'limits of oscillation' associated with estimates of treatment effects.

Gavarret LDJ (1840)

Principes généraux de statistique médicale: ou développement des régles qui doivent présider à son emploi [General principles of medical statistics: or the development of rules that must govern their use]. Paris: Bechet jeune & Labé.

Systematic reviews of methodology:

- Akl EA, Oxman AD, Herrin J, Vist GE, Terrenato I, Sperati F, Costiniuk C, Blank D, Schünemann H (2011). Using alternative statistical formats for presenting risks and risk reductions. Cochrane Database of Systematic Reviews(3):CD006776

3.2 Quantifying uncertainty in treatment comparisons

Chance may affect the results of a study if too few outcomes have been observed to yield reliable estimates of treatment effects. Small studies in which few outcome events occur are usually not informative and the results are sometimes seriously misleading.

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, and not necessarily to an 'important difference'. Statistically significant differences between treatments are not necessarily of any practical importance. But, tests of statistical significance are still important because they help us to avoid mistaken conclusions that real differences in treatments exist when they don't. 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.

In an earlier chapter, we mentioned Graham Balfour's unbiased assembly of treatment comparison groups using alternation. He was also aware of the importance of taking account of the play of chance when interpreting the results of his test of claims that belladonna could prevent orphans under his care developing scarlet fever. Two out of 76 boys allocated to receive belladonna developed scarlet fever compared with two 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".

We can reduce the likelihood that we will be misled by chance effects by estimating a range of treatment differences within which the real differences are likely to lie. These range estimates are known as confidence intervals. 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, and so they are more informative than mere tests of statistical significance, and thus more helpful in reducing the likelihood that we will be misled by the play of chance.

Statistical tests and confidence intervals – whether for analysis of individual studies or in 'meta-analysis' of several 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.

In more depth

The James Lind Library 3.2 Quantifying uncertainty in treatment comparisons (http://www.jameslindlibrary.org/essays/3-2-quantifying-uncertainty-in-treatment-comparisons/)

No methodology reviews were identified for this section.

3.3 Reducing the play of chance using systematic reviews and meta-analysis

Combining data from similar studies (systematic reviews and meta-analysis) can help to provide reliable estimates of treatment effects.

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, that is, by avoiding biased selection from the evidence available.

Even though care may have been taken to minimize biases in reviews, 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 estimates derived from all the relevant studies using a statistical procedure now known as 'meta-analysis'.

An early medical example of meta-analysis was published in the British Medical Journal in 1904. 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 (one of whom coined the term meta-analysis), and then by medical researchers.

Meta-analysis can be illustrated using the logo that marked the arrival of The Cochrane Collaboration in 1993. 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

The Cochrane logo is derived from the forest plot of an early systematic review, which has been kept up-to-date since its first publication in 1992.

Roberts D, Brown J, Medley N, Dalziel SR (2017)
Antenatal corticosteroids for accelerating fetal lung maturation for women at risk of preterm birth. Cochrane Database of Systematic Reviews 2017.

similar effects; if a horizontal line crosses the vertical line, it means that that no 'statistically significant' difference had been found between the treatments.

Taken together, however, the horizontal lines tend to fall on the beneficial (left) side of the 'no difference' line. The diamond at the bottom of the picture 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, so most obstetricians, midwives, and pregnant women did not realise that the treatment was so effective. Because no systematic reviews had been done, tens of thousands of premature babies suffered and many died unnecessarily because this effective drug was not used. 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.

In more depth

The James Lind Library 3.3 Reducing the play of chance using meta-analysis (http://www.jameslindlibrary.org/essays/3-3-reducing-the-play-of-chance-using-meta-analysis/)

Systematic reviews of methodology:

- Clarke M, Brice A, Chalmers I (2014). Accumulating research: a systematic account of how cumulative meta-analyses would have provided knowledge, improved health, reduced harm and saved resources. PLoS ONE 9(7):e102670

Section 4: Bringing it all together for the benefit of patients and the public

Improving reports of research and preparing and updating systematic reviews of reliable studies are essential foundations of effective health care.

Fair treatment comparisons avoid biases and reduce as far as possible the likelihood that users of research will be misled by the play of chance. These problems and their potential solutions have been discussed earlier in this book. However, even if the problems have been reduced as far as possible, health professionals, patients, policy makers and the public may often find it difficult to make direct use of reports of research.

Often, this is because both the individual studies and systematic reviews of them are of poor quality. Too often reports fail to provide important details about the design, conduct and analysis of research studies; adequate descriptions of who participated in them; what was done to participants; and what effects treatments had on outcome measures of importance to patients and others.

Very occasionally, a single well conducted and well reported study provides really strong evidence of the beneficial effects of an easily given 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. Another example is a comparison of older with newer treatments to treat eclampsia - convulsions during pregnancy - which showed that the older treatment was more effective. However, only very rarely does a single study provide such strong evidence, so it's important when reading reports of individual studies to ask what other evidence – published and unpublished – is relevant. This is why treatment and policy choices should, as far as possible, be informed by systematic reviews of as high a proportion as possible of the relevant evidence.

In more depth

The James Lind Library 4.0 Bringing it all together for the benefit of patients and the public (http://www.jameslindlibrary.org/essays/4-0-bringing-it-all-together-for-the-benefit-of-patients/

4.1 Improving reports of research

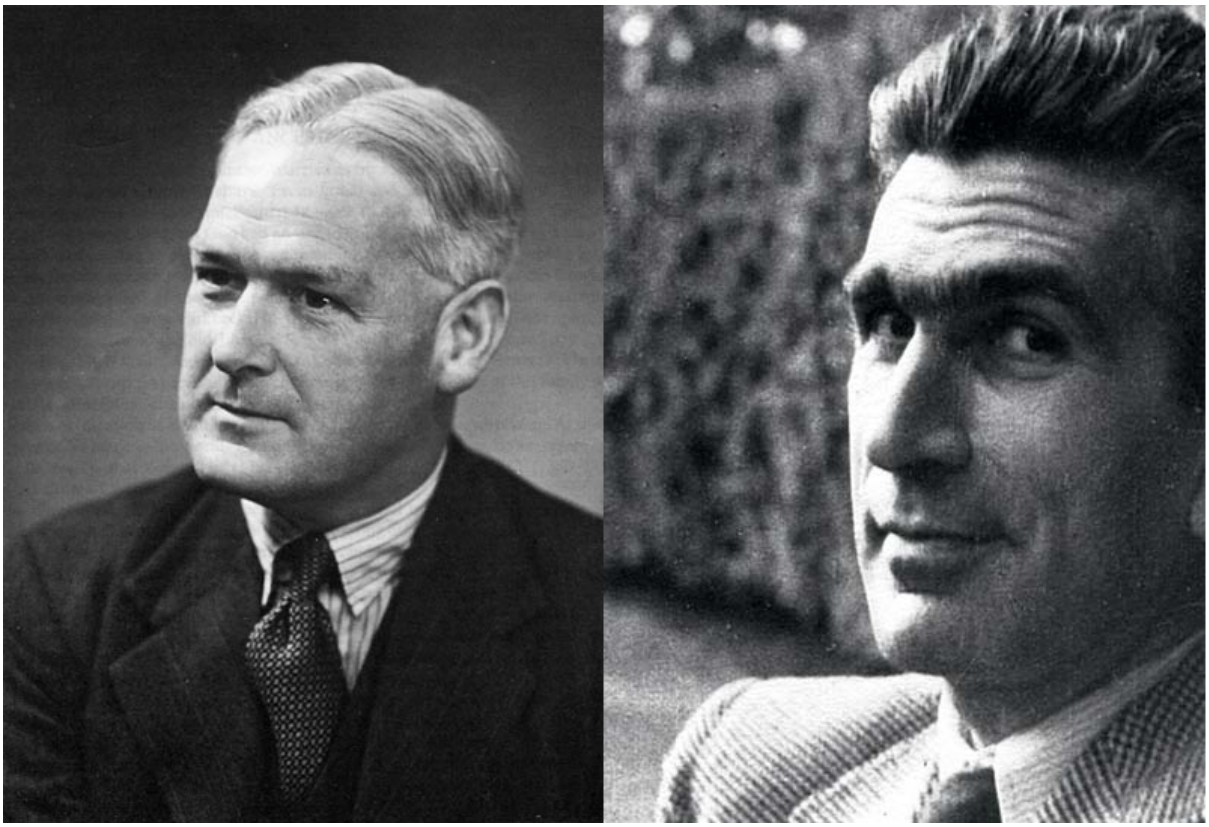
High quality, complete reports of research are needed to provide maximum return on the public's substantial investment in research on the effects of treatments.

The Medical Research Council's randomised trial comparing bed rest alone with bed rest and streptomycin for treating pulmonary tuberculosis (mentioned earlier) is renowned for several reasons. As far as the research methods used are concerned, it introduced secure methods for assuring that the comparison groups would be similar. However, another feature of the study report is that it was exceptionally clearly written. This reflected the care taken by the three members of the research team. One of them, Marc Daniels, went on to publish papers commenting on the inadequacy of many reports of research, and recommending reporting standards. Some years later, Austin Bradford Hill, one of Daniels' two senior colleagues, also offered guidance.

It was not until the 1980s that formal surveys of the quality of reports of research began to reveal just how common deficiencies were. Remedies began to be suggested in proposed reporting standards. The 1990s witnessed concerted international initiatives to improve the quality of reports of research. In a BMJ editorial in 1994, Douglas Altman commented on "the scandal of poor medical research" – "we need less research, better research and research done for the right reasons", he suggested. Since then, he and his colleagues in the Equator Network created a library of guidelines for reporting health research. Promoting adherence to these guidelines by researchers and journals remains a challenge.

In more depth

The James Lind Library 4.1 Improving reports of research (http://www.jameslindlibrary.org/essays/4-1-improving-reports-of-research/)



Austin Bradford-Hill and Marc Daniels

Daniels M, Hill AB (1952)

Chemotherapy of pulmonary tuberculosis in young adults: An analysis of the combined results of three medical research council trials. BMJ 1:1162-1168.

Systematic reviews of methodology:

- Jefferson T, Rudin M, Brodney Folse S, Davidoff F (2007). Editorial peer review for improving the quality of reports of biomedical studies. Cochrane Database of Systematic Reviews (2):MR000016
- Turner L, Shamseer L, Altman DG, Weeks L, Peters J, Kober T, Dias S, Schulz KF, Plint AC, Moher D (2012). Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. Cochrane Database of Systematic Reviews (11):MR000030

4.2 Preparing and maintaining systematic reviews of all the relevant evidence

Unbiased, up-to-date systematic reviews of all the relevant, reliable evidence are needed to provide trustworthy evidence to inform choices in practice and policy.

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?' 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.

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 several 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 fair test in the context of a careful assessment of all the evidence from fair tests that have addressed the question concerned.

Lord Rayleigh - 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."*

In his presidential address to the British Association for the Advancement of Science, Lord Rayleigh, a British physicist, observed that "the work that deserves most credit is that in which not only are new facts presented, but their relation to old ones is pointed out".

Rayleigh, The Lord (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: 3-23.

Very few reports of fair tests of treatments discuss their results in the context of a systematic assessment of all the other relevant evidence. 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 earlier, embarking on new tests of treatments without first reviewing systematically what can be learnt from existing research is dangerous, wasteful and unethical. 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. 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 in which research papers are increasingly made freely available online it should be possible to deal with the limitations found in most reports of new research. Rather than basing conclusions about the treatments on one or a few individual studies, users of research evidence are increasingly turning for reliable information to online, 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 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, similar steps must also 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; and
- preparing a structured report

One manifestation of increasing recognition of the crucial importance of systematic reviews for assessing the effects of treatments has been the rapid evolution of methods to improve the reliability of reviews. The first edition of a book entitled *Systematic Reviews* [1995] was less than 100 pages long: only six years later, the second edition weighed in at nearly 500 pages and included rapidly evolving strategies for increasing the information obtained from research.

There continue to be important developments in the methods used for preparing systematic reviews, including those needed to identify unanticipated effects of treatments and for incorporating the results of research describing and analysing the experiences of people giving and receiving treatments.

In more depth

The James Lind Library 4.2 Preparing and maintaining systematic reviews of all the relevant evidence (http://www.jameslindlibrary.org/essays/4-2-preparing-and-maintaining-systematic-reviews-of-all-the-relevant-evidence/)

Systematic reviews of methodology:

- Baudard M, Yavchitz A, Ravaud P, Perrodeau E, Boutron I (2017). Impact of searching clinical trial registries in systematic reviews of pharmaceutical treatments: methodological systematic review and reanalysis of meta-analyses. BMJ 356:j448

- Hopewell S, McDonald S, Clarke MJ, Egger M (2017). Grey literature in meta-analyses of randomized trials of health care interventions. Cochrane Database of Systematic Reviews (2):MR000010
- Martínez García L, Pardo-Hernandez H, Superchi C, Niño de Guzman E, Ballesteros M, Ibargoyen Roteta N, McFarlane E, Posso M, Roqué I Figuls M, Rotaeche Del Campo R, Sanabria AJ, Selva A, Solà I, Vernooij RWM, Alonso-Coello P (2017). Methodological systematic review identifies major limitations in prioritization processes for updating. Journal of Clinical Epidemiology 86:11-24
- Moher D, Tsertsvadze A, Tricco A, Eccles M, Grimshaw J, Sampson M, Barrowman N. When and how to update systematic reviews (2008). Cochrane Database of Systematic Reviews (1):MR000023
- Page MJ, McKenzie JE, Kirkham J, Dwan K, Kramer S, Green S, Forbes A (2014). Bias due to selective inclusion and reporting of outcomes and analyses in systematic reviews of randomised trials of healthcare interventions. Cochrane Database of Systematic Reviews (10):MR000035
- Schmucker CM, Blümle A, Schell LK, Schwarzer G, Oeller P, Cabrera L, von Elm E, Briel M, Meerpohl JJ; OPEN consortium (2017). Systematic review finds that study data not published in full text articles have unclear impact on meta-analyses results in medical research. PLoS One 12(4):e0176210
- Tudur Smith C, Marcucci M, Nolan SJ, Iorio A, Sudell M, Riley R, Rovers MM, Williamson PR (2016). Individual participant data meta-analyses compared with meta-analyses based on aggregate data. Cochrane Database of Systematic Reviews (9):MR000007

4.3 Using the results of research

All research has been done in the past, but the results of research need to be used today and tomorrow to inform decisions in health care. Trustworthy evidence from research is necessary, but not sufficient, to improve the quality of health care.

Over recent years, it has been realised increasingly that systematic reviews of research are needed to express fully the significance of fair tests of treatments. This trend has been reflected in a rapid increase in the numbers of reports of systematic reviews being published on paper and electronically. Sometimes reviews will show that no reliable evidence exists, and this is one of their most important functions. Similarly, reviews may sometimes confirm that reliable evidence is limited to a single study; and here, too, it is important to make this explicit.

Systematic reviews of research are being used widely (a) to inform clinical practice, often through clinical practice guidelines; (b) to assess which medical treatments are cost-effective; (c) to shape the agenda for additional research; and (d) to meet the needs of patients for reliable information about the effects of treatments.

These developments show that people trying to improve access to the evidence that is needed to inform choices in health care have accepted the importance of systematic reviews, but there is still a long way to go. Many thousands of systematic reviews will be needed to cover existing research evidence, and then kept up to date as new evidence emerges. Indeed, one journal editor suggested in 1993 that there should be a moratorium on all new research until we've caught up with what existing evidence can tell us. That didn't happen and new research continues to appear at an overwhelming pace.

Those responsible for disbursing funds for research must ensure that resources are provided to cope with this ever-increasing backlog. Support for new studies

Andy Oxman and Elizabeth Paulsen recommend two sources of information that are derived from systematic reviews, presented in an open access format, and using language that is lay-friendly: Cochrane Evidence and Informed Health.

Oxman AD, Paulsen EJ (2019)

Who can you trust? A review of free online sources of "trustworthy" information about treatment effects for patients and the public. BMC Med Inform Decis Mak. Feb 20;19(1):35.

should only be provided 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 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.

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.

Systematic reviews are necessary but insufficient for informing decisions about treatments for individual patients and policies. Other important factors – values, preferences, needs, resources and priorities – must be considered. And this is the point at which the art as well as the science of health care needs to be deployed for the benefit of patients and the public. We hope that this book helps everyone to achieve this.

In more depth

The James Lind Library 4.3 Using the results of research (http://www.jameslindlibrary.org/essays/4-3-using-the-results-of-up-to-date-systematic-reviews-of-research/)

Systematic reviews of methodology:

- Akl EA, Oxman AD, Herrin J, Vist GE, Terrenato I, Sperati F, Costiniuk C, Blank D, Schünemann H. Framing of health information messages (2011). Cochrane Database of Systematic Reviews (12):CD006777
- Chiu K, Grundy Q, Bero L (2017). 'Spin' in published biomedical literature: A methodological systematic review. PLoS Biology 15(9):e2002173
- Covey J (2007). A meta-analysis of the effects of presenting treatment benefits in different formats. Medical Decision Making 27(5):638-54
- Moxey A, O'Connell D, McGettigan P, Henry D (2003). Describing treatment effects to patients. Journal of General Internal Medicine 18(11):948-59

Acknowledgements

The co-authors are grateful to Iain Milne and Estela Dukan for making available material in the Sibbald Library of the Royal College of Physicians of Edinburgh, and to Phoebe Marson Smith for helpful comments on earlier drafts.

Copyright