

# TESTING TREATMENTS

## Chapter 6

### 6 Fair tests of treatments

---

The principles underlying fair tests of treatments may not be familiar to many readers, but they are not complicated. In fact, much of our everyday, intuitive grasp of the world depends on them. Yet they are not taught well in schools and are often needlessly wrapped up in complex language. As a result, many people shy away from the subject, believing that it is beyond their ability to comprehend. We hope this and the following two chapters will persuade you that you are actually already aware of the key principles, and so will readily understand why they are so important. Readers who would like to explore these issues in more detail will find additional material at [www.testingtreatments.org](http://www.testingtreatments.org) and in *The James Lind Library* ([www.jameslindlibrary.org](http://www.jameslindlibrary.org)).

#### WHY ARE FAIR TESTS OF TREATMENTS NEEDED?

##### **Nature, the healer**

Many health problems will tend to get worse without treatment, and some will get worse in spite of treatment. However, some get better by themselves – that is, they are ‘self-limiting’. As one researcher involved in testing a proposed treatment for the common cold put it: ‘if a cold is treated energetically it will get well in seven days, while if left to itself it will get well in a week.’<sup>1</sup> Put more cynically, ‘Nature cures, but the doctor takes the fee.’

And of course, treatment may actually make matters worse.

It is because people often recover from illness without any specific treatment that the ‘natural’ progress and outcome of illnesses without treatment must be taken into account when treatments are being tested. Think about a time when you have had a sore throat, a stomach cramp, or an unusual skin rash. These will often resolve on their own, without formal treatment. Yet, if you *had* received treatment (even an ineffective treatment), you might have assumed that the treatment caused the symptoms to disappear. In short, knowledge of the natural history of an illness, including the likelihood that it will get better on its own (spontaneous remission), can prevent use of un-needed treatments and false beliefs in unproven remedies.

When symptoms of an illness come and go, it is especially difficult to try to pin down the effects of treatments. Patients with arthritis, for example, are most likely to seek help when they are having a particularly bad flare-up – which, by its very nature, is unlikely to be sustained. Whether the treatment they then receive is mainstream or complementary, effective or ineffective, it is

### MISTAKING THE CURE

. . . ‘it is alleged to be found true by proof, that by the taking of *Tobacco*, divers and very many do find themselves cured of divers diseases; as on the other part, no man ever received harm thereby. In this argument there is first a great mistaking, and next a monstrous absurdity: . . . when a sick man has his disease at the height, he hath at that instant taken *Tobacco*, and afterward his disease taking the natural course of declining and consequently the patient of recovering his health, O then the *Tobacco* forsooth, was the worker of that miracle.’

James Stuart, King of Great Britaine, France and Ireland. 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, 1616: pp 214-222.

likely that their pain will improve after receiving it, simply because the flare-up dies down. Understandably, however, practitioners and patients will tend to attribute such improvements to the treatment taken, even though it may have had nothing to do with the improvements.

### **The beneficial effects of optimism and wishful thinking**

The psychological reasons for people attributing any improvement in their condition to the treatment they received are now better understood. We all have a tendency to assume that if one event follows another, the first may have been responsible for the second. And we are inclined to see patterns where none exist – a phenomenon that has been demonstrated many times in areas as diverse as coin tossing, stock market prices and basketball shots. We are all also prone to a problem known as confirmation bias: we see what we expect to see – ‘believing is seeing’. Any support we find for our beliefs will boost our confidence that we are right. Conversely, we may not recognize or readily accept information that contradicts our views, and so tend to turn a blind eye to it – often unconsciously.

#### **BELIEVING IS SEEING**

The British doctor Richard Asher noted in one of his essays for doctors:

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

Asher R. Talking sense (Lettsomian lecture, 16 Feb, 1959). *Transactions of the Medical Society of London*, vol LXXV, 1958-59. Reproduced in: Jones, FA, ed. *Richard Asher talking sense*. London: Pitman Medical, 1972.

Most patients and clinicians hope, of course, that treatments will help. They may conclude that something works simply because it agrees with their belief that it *should* work. They do not look for, or they discard, information that is contrary to their beliefs. These psychological effects also explain why patients who believe that a treatment will help to relieve their symptoms may well experience improvements in their condition – even though the treatment, in fact, has no active ingredient (a ‘sham’, often known as a ‘placebo’). Patients have reported improvements after being given pills made of sugar, injections of water, treatments with inactivated electric gadgetry, and surgery where nothing happened other than a small cut being made and sewn up again.

Take the example of a test comparing different weight-reducing diets. Researchers recruited viewers of a popular television programme who wanted to lose weight and assigned them to one of six diets. One of the diets – *bai lin* tea – had been promoted as a successful way of losing weight. The average weight of the slimmers went down in all six groups, but in some much more than in others. However, when the results were presented on television, it was revealed that one of the diets – ‘the carrot diet’ – was not a slimming diet at all. It had been included in the test to provide a ‘bench mark’ of weight loss which was due not to any of the six diets, but to changes in eating habits resulting from other factors that had motivated participants to eat differently.<sup>2</sup>

### **The need to go beyond impressions**

If patients believe that something helps them, isn’t that enough? Why is it important to go to the trouble and expense of doing research to try to assess the effects of the treatment more formally, and perhaps to try to find out whether and if so *how* it has helped them? ‘There are at least two reasons. One is that treatments that do not work may distract us from treatments that do work. Another reason is that many (if not most) treatments have adverse side-effects, some short term, some longer term, and some still unrecognized. If patients do not use these treatments, they can be spared the unwanted effects. So it is worth identifying treatments that are very unlikely to help or might cause more harm than benefit. Research may also uncover important information about

how treatments work, and so indicate possibilities for developing better and safer treatments.

Research about the effects of treatments is relevant everywhere, but especially in communities that endeavour to share healthcare resources fairly among all patients – for example, in the British National Health Service, or the US Veterans Health Administration. In these circumstances, decisions always have to be taken about which treatments represent good value for the inevitably limited resources available for healthcare. If some patients are given treatments that have not been shown to be useful, this may mean depriving other patients of treatments that have been shown to be beneficial.

None of this should suggest that patients' and clinicians' impressions and ideas about the effects of treatments are unimportant. Indeed they are often the starting point for formal investigation of apparently promising new treatments. Following up such impressions with formal research can sometimes lead to the identification of both harmful and useful effects of treatments. For example, it was a woman who had been treated with the drug diethylstilboestrol (DES) during pregnancy two decades earlier who first suggested that this might have caused her daughter's rare vaginal cancer (see Chapter 2, p15-16). And when a patient mentioned unexpected side-effects of a new treatment prescribed for his raised blood pressure, neither he nor his doctor could have imagined that his comment would lead to the identification of an all-time best-selling drug – sildenafil (Viagra).

So, individuals' impressions about the effects of treatments should not be ignored, but they are seldom a reliable basis for drawing sound conclusions about the effects of treatments, let alone for recommending treatments to others.

### **So what are fair tests?**

Most of us know that it can be a mistake to take a media report of some new medical advance at face value. But the sad truth is that one must also be cautious about reports of treatments even in apparently reputable journals. Misleading and overblown claims about treatments are common, and it is important to be able to assess their reliability.

We run two risks in taking reports of the effects of treatments at face value. We could wrongly conclude that a helpful treatment is actually useless or even dangerous. Or we could wrongly conclude that a useless or even dangerous treatment is actually helpful. Fair tests of treatments are designed to obtain reliable information about the effects of treatments by (i) comparing like with like, to reduce distorting influences (biases); (ii) taking account of the play of chance; and (iii) assessing all the relevant, reliable evidence. This chapter and the next two chapters deal with these three principal features of fair tests.

## COMPARING LIKE WITH LIKE

### Comparisons are key

Comparisons are key to all fair tests of treatments. Clinicians and patients sometimes compare in their minds the relative merits of two treatments. For example, they may form an impression that they or others are responding differently to a treatment compared with responses to previous treatments. Sometimes the comparisons are made more formally. As early as the ninth-century, the Persian physician al-Razi compared the outcome of patients with meningitis treated with blood-letting with the outcome of those treated without it to see if blood-letting could help.

Treatments are usually tested by comparing groups of patients who have received different treatments. If treatment comparisons are to be fair, the comparisons must ensure that like will be compared with like: that the only systematic difference between the groups of patients is the treatments they have received. This insight is not new. For example, before beginning his comparison of six treatments for scurvy on board HMS *Salisbury* in 1747, James Lind (i) took care to select patients who were at a similar stage of this often lethal disease; (ii) ensured that the patients had the same basic diet; and (iii) arranged for them to be accommodated in similar conditions (see Chapter 1, p1-3). Lind recognized that factors other than the treatments themselves might influence his patients' chances of recovery.

One way to make a test *unfair* would have been to give one of the treatments recommended for scurvy – say, sulphuric acid,

which was being recommended by the Royal College of Physicians of London – to patients who were less ill to begin with and in the early stages of the disease, and another treatment – say, citrus fruits, which were being recommended by some sailors – to patients who were already approaching death. This would have made sulphuric acid appear to be better, even though it was actually worse. Biases such as these can arise unless care is taken to ensure that like is being compared with like in all relevant respects.

### **Treatments with dramatic effects**

Sometimes patients experience responses to treatments which differ so dramatically from their own past experiences, and from the natural history of their illness, that confident conclusions about treatment effects can be drawn without carefully done tests (see Chapter 5, p50-53).<sup>3</sup> For a patient with a collapsed lung (pneumothorax), inserting a needle into the chest and letting out the trapped air causes such immediate relief that the benefits of this treatment are clear. Other examples of dramatic effects include morphine on pain, insulin in diabetic coma, and artificial hip joints on pain from arthritis. Adverse effects of treatment can be dramatic as well. Sometimes drugs provoke severe, even lethal, allergic reactions; other dramatic effects include the rare limb deformities caused by thalidomide (see Chapter 1, p4-5).

However, such dramatic effects of treatments, whether beneficial or harmful, are rare. Most treatment effects are more modest, but still worth knowing about. For example, carefully done tests are needed to identify which dosage schedules for morphine are effective and safe; or whether genetically engineered insulin has any advantages over animal insulins; or whether a newly marketed artificial hip that is 20 times more expensive than the least expensive variety is worth the extra cost in terms that patients can appreciate. In these common circumstances we all need to avoid unfair (biased) comparisons, and the mistaken conclusions that can result from them.

## Treatments with moderate but important effects

### *Comparing patients given treatments today with apparently similar patients given other treatments in the past for the same disease*

Researchers sometimes compare patients given treatments today with apparently similar patients given other treatments in the past for the same disease. Such comparisons can provide reliable evidence if the treatment effects are dramatic – for example, when a new treatment now leads some patients to survive from a disease that had been universally fatal. However, when the differences between the treatments are not dramatic, but nevertheless worth knowing about, such comparisons using ‘historical controls’ are potentially problematic. Although researchers use statistical adjustments and analyses to try to ensure that like will be compared with like, these analyses cannot take account of relevant features of patients in the comparison groups which have not been recorded. As a result, we can never be completely confident that like is being compared with like.

The problems can be illustrated by comparing the results of the same treatment given to similar patients, but at different points in time. Take an analysis of 19 such instances in patients with advanced lung cancer comparing the annual death rates experienced by similar patients treated at different points in time with exactly the same treatments. Although few differences in death rates would have been expected, in fact the differences were considerable: death rates ranged from 24% better to 46% worse.<sup>4</sup> Clearly, these differences were not because the treatments had changed – they were the same – or because the patients were detectably different – they weren’t. The differing death rates presumably reflected either undetected differences between the patients, or other, unrecorded changes over time (better nursing or control of infection, for example), which could not be taken into account in the comparisons.

### *Comparing apparently similar groups of patients who happen to have received different treatments in the same time period*

Comparing the experiences and outcomes of apparently similar groups of patients who happen to have received different treatments

in the same time period is still used as a way to try to assess the effects of treatments. However, this approach too can be seriously misleading. The challenge, as with comparisons using 'historical controls', is to know whether the groups of people receiving the different treatments were sufficiently alike before they started treatment for a valid comparison to be possible – in other words, whether like was being compared with like. As with 'historical controls', researchers may use statistical adjustments and analyses to try to ensure that like will be compared with like, but only if relevant features of patients in the comparison groups have been recorded and taken into account. So seldom will these conditions have been met that such analyses should always be viewed with great caution. Belief in them can lead to major tragedies.

A telling example concerns hormone replacement therapy (HRT). Women who had used HRT during and after the menopause were compared with apparently similar women who had not used it. These comparisons suggested that HRT reduced the risk of heart attacks and stroke – which would have been very welcome news if it were true. Unfortunately it wasn't. Subsequent comparisons, which were designed before treatment started to ensure that the comparison groups would be alike, showed that HRT had exactly the opposite effect – it actually increased heart attacks and strokes (see Chapter 2, p16-18). In this case, the apparent difference in the rates of heart attacks and strokes was due to the fact that the women who used HRT were generally healthier than those who did not take HRT – it was not due to the HRT. Research that has not ensured that like really is being compared with like can result in harm being done to tens of thousands of people.

As the HRT experience indicates, the best way to ensure that like will be compared with like is to assemble the comparison groups before starting treatment. The groups need to be composed of patients who are similar not just in terms of known and measured factors, such as age and the severity of their illness, but also in terms of unmeasured factors that may influence recovery from illness, such as diet, occupation and other social factors, or anxiety about illness or proposed treatments. It is always difficult – indeed often impossible – to be confident that treatment groups are alike if they have been assembled after treatment has started.

The critical question then is this: do differences in outcomes reflect differences in the effects of the *treatments* being compared, or differences in the *patients* in the comparison groups?

*Unbiased, prospective allocation to different treatments*

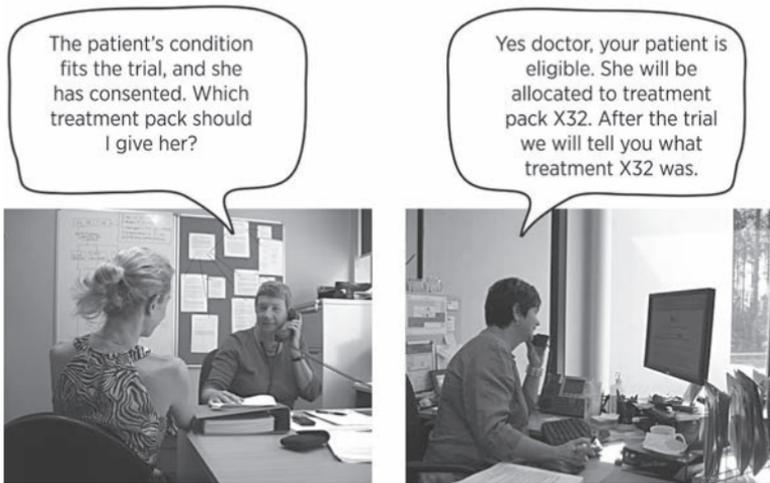
In 1854, Thomas Graham Balfour, an army doctor in charge of a military orphanage, showed how treatment groups could be created to ensure that like would be compared with like. Balfour wanted to find out whether belladonna protected children from scarlet fever, as some people were claiming. So, ‘to avoid the imputation of selection’ as he put it, he allocated children *alternately* either to receive the drug, or not to receive it.<sup>5</sup> The use of alternate allocation, or some other unbiased way of creating comparison groups, is a key feature of fair tests of treatments. It increases the likelihood that comparison groups will be similar, not just in terms of known and measured important factors, but also of unmeasured factors that may influence recovery from illness, and for which it is impossible to make statistical adjustments.

To achieve fair (unbiased) allocation to different treatments it is important that those who design fair tests ensure that clinicians and patients cannot know or predict what the next allocation will be. If they do know, they may be tempted, consciously or unconsciously, to choose particular treatments. For example, if a doctor knows that the next patient scheduled to join a clinical trial is due to get a placebo (a sham treatment), she or he might discourage a more seriously ill patient from joining the trial and wait for a patient who was less ill. So even if an unbiased allocation *schedule* has been produced, unbiased *allocation* to treatment groups will only occur if upcoming allocations in the schedule are successfully concealed from those taking decisions about whether or not a patient will join a trial. In this way, no one will be able to tell which treatment is going to be allocated next, and tempted to depart from the unbiased allocation schedule.

Allocation concealment is usually done by generating allocation schedules that are less predictable than simple alternation – for example, by basing allocation on random numbers – and by concealing the schedule. Several methods are used to conceal allocation schedules. For example, random allocation can be

assigned remotely – by telephone or computer – for a patient confirmed as eligible to participate in the study. Another way is to use a series of numbered envelopes, each containing an allocation – when a patient is eligible for a study, the next envelope in the series is opened to reveal what the allocation is. For this system to work, the envelopes have to be opaque so that doctors can't 'cheat' by holding the envelope up to the light to see the allocation inside.

This approach is recognized today as a key feature of fair tests of treatments. Studies in which random numbers are used to allocate treatments are known as 'randomized trials' (see box in Chapter 3, p26).



**Concealing treatment allocation in a trial using telephone randomization.**

*Ways of using unbiased (random) allocation in treatment comparisons*

Random allocation for treatment comparisons can be used in various ways. For example, it can be used to compare different treatments given at different times in random order to the same patient – a so-called 'randomized cross-over trial'. So, to assess whether an inhaled drug could help an individual patient with a persistent, dry cough, a study could be designed to last a few months. During some weeks, chosen randomly, the patient

## 6 FAIR TESTS OF TREATMENTS

would use an inhaler containing a drug; during the other weeks the patient would use an identical-looking inhaler which did not contain the drug. Tailoring the results of research to individual patients in this way is clearly desirable if it can be done. But there are many circumstances in which such crossover studies are simply not possible. For example, different surgical operations cannot be compared in this way, and nor can treatments for ‘one-off’, acute health problems, such as severe bleeding after a road crash.

Random allocation can also be used to compare different treatments given to different parts of the same patient. So, in a skin disorder such as eczema or psoriasis, affected patches of skin can be selected at random to decide which should be treated with ointment containing a drug, and which with ointment without the active ingredients. Or in treating illness in both eyes, one of the eyes could be selected at random for treatment and comparison made with the untreated eye.

Another use of random allocation is to compare different treatments given to different populations or groups – say, all the people attending each of a number of primary care clinics

	Mon	Tues	Weds	Thurs	Fri
May week 1	A	B	A	B	B
May week 2	A	A	B	B	A
May week 3	B	A	B	A	A
May week 4	B	A	B	B	A
May week 5	A	B	A	B	B

Randomize days within a person



Randomize eyes within a person



Randomize communities within a region



Randomize individuals within a group

**Different possible units for random allocation.**

or hospitals. These comparisons are known as ‘cluster (or group) randomized trials’. For example, to assess the effects of the Mexican universal health insurance programme, researchers matched 74 pairs of healthcare catchment areas – clusters that collectively represented 118,000 households in seven states. Within each matched pair one was allocated at random to the insurance programme.<sup>6</sup>

However by far the most common use of random allocation is its use to decide which patient will receive which treatment.

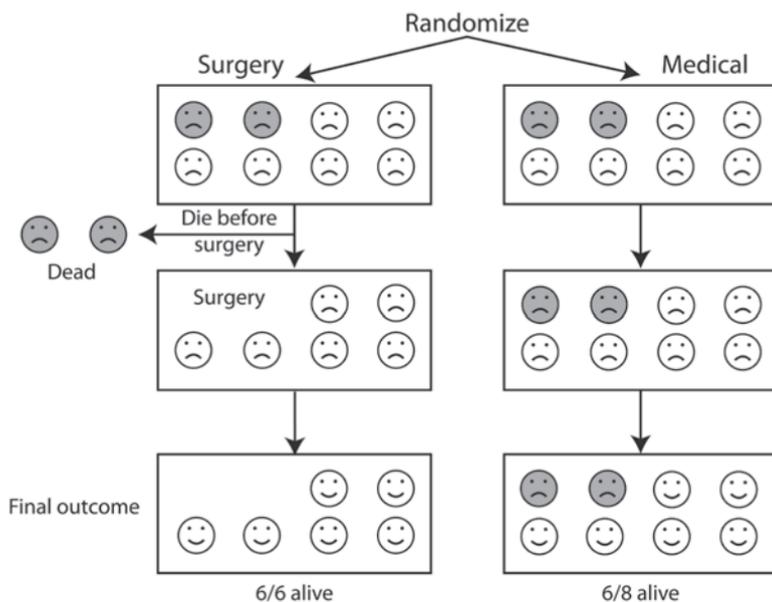
### *Following up everyone in treatment comparisons*

After taking the trouble to assemble comparison groups to ensure that like will be compared with like, it is important to avoid introducing the bias that would result if the progress of some patients were to be ignored. As far as possible, all the patients allocated to the comparison groups should be followed up and included in the main analysis of the results of the group to which they were allocated, irrespective of which treatment (if any) they actually received. This is called an ‘intention-to-treat’ analysis. If this is not done, like will no longer be compared with like.

At first sight it may seem illogical to compare groups in which some patients have not received the treatments to which they were assigned, but ignoring this principle can make the tests unfair and the results misleading. For example, patients who have partial blockages of blood vessels supplying the brain and who experience dizzy spells are at above average risk of having a stroke. Researchers conducted a test to find out whether an operation to unclog blood vessels in these patients would reduce subsequent strokes. They rightly compared all the patients allocated to have the operation, irrespective of whether they survived the surgery, with all those allocated not to have it. If they had recorded the frequency of strokes only among patients who survived the immediate effects of the operation, they would have missed the important fact that the surgery itself can cause stroke and death and, other things being equal, the surviving patients in this group will have fewer strokes. That would have been an unfair test of the effects of the operation, the risks of which need to be factored into the assessment.

The outcomes of surgery and medical treatment shown in the

## 6 FAIR TESTS OF TREATMENTS



**Why all patients randomized should be included in the final outcome ('intention to treat').**

figure are actually equal. However, if the two people allocated to surgery die before operation and are then excluded from consideration, the comparison of the two groups will be biased. It will suggest that surgery appears to be better when it is not.

### *Dealing with departures from allocated treatments*

For all the reasons given so far in this chapter, you will have realized that fair tests of treatments have to be planned carefully. The documents setting out these plans are known as research protocols. However, the best-laid plans may not work out quite as intended – the treatments actually received by patients sometimes differ from those they were allocated. For example, patients may not take treatments as intended; or one of the treatments may not be given because supplies or personnel become unavailable. If such discrepancies are discovered, the implications need to be considered and addressed carefully.

During the 1970s and 1980s, there were remarkable advances in the treatment of children with acute lymphoblastic leukaemia,

the most common type of leukaemia in this age group. However, it was puzzling that American children were doing substantially better than British children who, on the face of it, were receiving exactly the same drug regimens.<sup>7</sup> During a visit to a children's cancer centre in California, an astute British statistician noticed that American children with leukaemia were being treated far more 'aggressively' with chemotherapy than children in the UK. The treatment had nasty side-effects (nausea, infection, anaemia, hair loss, and so on) and when these side-effects were particularly troublesome, British doctors and nurses, unlike their American counterparts, tended to reduce or pause the prescribed treatment. This 'gentler approach' appears to have reduced the effectiveness of the treatment, and was probably a reason for the differences in British and American treatment success.

#### *Helping people to stick to allocated treatments*

Differences between intended and actual treatments during treatment comparisons can happen in other ways that may complicate the interpretation of tests of treatments. Participants in research should not be denied medically necessary treatments. When a new treatment with hoped-for, but unproven, beneficial effects is being studied in a fair test, therefore, participating patients should be assured that they will all receive established effective treatments.

If people know who is getting what in a study, several possible biases arise. One is that patients and doctors may feel that people allocated to 'new' treatments have been lucky, and this may cause them unconsciously to exaggerate the benefits of these treatments. On the other hand, patients and doctors may feel that people allocated 'older' treatments are hard done by, and this disappointment may cause them to under-estimate any positive effects. Knowing which treatments have been allocated may also cause doctors to give the patients who have been allocated the older treatments some extra treatment or care, to compensate, as it were, for the fact that they had not been allocated to receive the newer, but unproven treatments. Using such additional treatments in patients in one of the comparison groups but not in the other group complicates the evaluation of a new treatment, and risks

making the comparison unfair and the results misleading. A way to reduce differences between intended and actual treatment comparisons is to try to make the newer and older treatments being compared look, taste and smell the same.

This is what is done when a treatment with hoped-for beneficial effects is compared with a treatment with no active ingredients (a sham treatment, or placebo), which is designed to look, smell, taste and feel like the 'real' treatment. This is called 'blinding', or 'masking'. If this 'blinding' can be achieved (and there are many circumstances in which it cannot), patients in the two comparison groups will tend to differ in only one respect – whether they have been allocated to take the new treatment or the one with no active ingredients. Similarly, the health professionals caring for the patients will be less likely to be able to tell whether their patients have received the new treatment or not. If neither doctors nor patients know which treatment is being given, the trial is called 'double blind'. As a result, patients in the two comparison groups will be similarly motivated to stick to the treatments to which they have been allocated, and the clinicians looking after them will be more likely to treat all the patients in the same way.

### **Fair measurement of treatment outcome**

Although one of the reasons for using sham treatments in treatment comparisons is to help patients and doctors to stick to the treatments allocated to them, a more widely recognized reason for such 'blinding' is to reduce biases when the outcomes of treatments are being assessed.

Blinding for this reason has an interesting history. In the 18th century, Louis XVI of France called for an investigation into Anton Mesmer's claims that 'animal magnetism' (sometimes called 'mesmerism') had beneficial effects. The king wanted to know whether the effects were due to any 'real force', or rather to 'illusions of the mind'. In a treatment test, blindfolded people were told either that they were or were not receiving animal magnetism when in fact, at times, the reverse was happening. People only reported feeling the effects of the 'treatment' when they had been told that they were receiving it.

For some outcomes of treatment – survival, for example –

biased outcome assessment is very unlikely since there is little room for doubt about whether or not someone has died. However, assessing most outcomes will entail some subjectivity, because outcomes should and often do involve patients' experiences of symptoms such as pain and anxiety. People may have individual reasons for preferring one of the treatments being compared. For example, they may be more alert to signs of possible benefit when they believe a treatment is good for them, and more ready to ascribe harmful effects to a treatment about which they are worried.

In these common circumstances, blinding is a desirable feature of fair tests. This means that the treatments being compared must appear to be the same. In a test of treatments for multiple sclerosis, for example, all the patients were examined both by a doctor who did not know whether the patients had received the new drugs or a treatment with no active ingredient (that is, the doctor was 'blinded'), and also by a doctor who knew the comparison group to which the patients had been allocated (that is, the doctor was 'unblinded'). Assessments done by the 'blinded' doctors suggested that the new treatment was not useful whereas assessments done by the 'unblinded' doctors suggested that the new treatment was beneficial.<sup>8</sup> This difference implies the new treatment was not effective and that knowing the treatment assignment led the 'unblinded' doctors to have 'seen what they believed' or hoped for. Overall, the greater the element of subjectivity in assessing treatment outcomes, the greater the desirability of blinding to make tests of treatments fair.

Sometimes it is even possible to blind patients as to whether or not they have received a real surgical operation. One such study was done in patients with osteoarthritis of the knee. There was no apparent advantage of a surgical approach that involved washing out the arthritic joints when this was compared with simply making an incision through the skin over the knee under anaesthesia, and 'pretending' that this had been followed by flushing out the joint space.<sup>9</sup>

Often it is simply impossible to blind patients and doctors to the identity of treatments being compared – for example, when comparing surgery and a drug treatment or when a drug has a characteristic side-effect. However, even for some outcomes for

which bias might creep in – say, in assigning a cause of death, or judging an X-ray – this can be avoided by arranging for these outcomes to be assessed independently by people who do not know which treatments individual patients have received.

### **Generating and investigating hunches about unanticipated adverse effects of treatments**

#### *Generating hunches about unanticipated effects of treatments*

Unanticipated effects of treatments, whether bad or good, are often first suspected by health professionals or patients.<sup>10</sup> Because the treatment tests needed to get marketing licences include only a few hundred or a few thousand people treated over a few months, only relatively short-term and frequent side-effects are likely to be picked up at this stage. Rare effects and those that take some time to develop will not be discovered until the treatments have been in more widespread use, over a longer time period, and in a wider range of patients than those who participated in the pre-licensing tests.

In an increasing number of countries – including the UK, the Netherlands, Sweden, Denmark, and the USA – there are facilities for clinicians and patients to report suspected adverse drug reactions, which can then be investigated formally.<sup>11</sup> Although none of these reporting schemes has been especially successful in identifying important adverse reactions to drugs, there are instances where they have been. For example, when the cholesterol-lowering drug rosuvastatin was launched in the UK in 2003, reports soon began to identify a serious, rare, unanticipated adverse effect on muscles called rhabdomyolysis. In this condition, muscles break down rapidly and the breakdown products can cause serious kidney damage. Further investigation helped to show that the patients most at risk of this complication were those taking high doses of the drug.

#### *Investigating hunches about unanticipated effects of treatments*

Hunches about adverse effects often turn out to be false alarms.<sup>10</sup> So how should hunches about unanticipated effects of treatments be investigated to find out whether the suspected effects are real? Tests to confirm or dismiss suspected unanticipated effects

## THE YELLOW CARD SCHEME

The Yellow Card Scheme was launched in Britain in 1964 after the thalidomide tragedy highlighted the importance of following up problems that occur after a drug has been licensed. Reports are sent to the Medicines and Healthcare products Regulatory Agency (MHRA), which analyzes the results. Each year, the MHRA receives more than 20,000 reports of possible side-effects. Initially, only doctors could file the reports, but then nurses, pharmacists, coroners, dentists, radiographers and optometrists were encouraged to do so. Since 2005, patients and carers have been invited to report suspected adverse reactions. Reports can be filed online at [www.yellowcard.gov.uk](http://www.yellowcard.gov.uk), by post, or by phone.

One patient summarised her experience this way: 'Being able to report side effects through the Yellow Card Scheme puts you in control. It means that you can report directly without having to wait for a busy healthcare professional to do it . . . It's about putting patients at the centre of care. It's a quantum leap for patient involvement, and marks the beginning of the way forward and a sea change in attitude.'

Bowser A. A patient's view of the Yellow Card Scheme. In: *Medicines & Medical Devices Regulation: what you need to know*. London: MHRA, 2008. Available at [www.mhra.gov.uk](http://www.mhra.gov.uk)

must observe the same principles as studies to identify hoped-for, anticipated effects of treatments. And that means avoiding biased comparisons, ensuring that 'like is compared with like', and studying adequate numbers of instances.

As with hoped-for effects of treatments, unanticipated dramatic effects are easier to spot and confirm than less dramatic treatment effects. If the suspected, unanticipated treatment outcome is normally very unusual but occurs quite often after a treatment has been used, it will generally strike both clinicians and patients that something is wrong. In the late 19th century, a Swiss surgeon, Theodor Kocher, learned through a general practitioner

that one of the girls whose thyroid goitre Kocher had removed some years previously had become dull and lethargic. When he looked into this and other former goitre patients on whom he had operated, he discovered that complete removal of the enlarged thyroid gland had resulted in cretinism and myxoedema – rare, serious problems resulting from lack of the hormone produced by the gland, as we now know.<sup>12</sup> The unanticipated effects of thalidomide (see Chapter 1, p4-5) were suspected and confirmed because the association between use of the drug in pregnancy and the birth of babies born without limbs was dramatic. Such abnormalities were previously almost unheard of.

Less dramatic unanticipated effects of treatments sometimes come to light in randomized trials designed to assess the relative merits of alternative treatments. A randomized comparison of two antibiotics given to newborn infants to prevent infection revealed that one of the drugs interfered with the body's processing of bilirubin, a waste product from the liver. The build up of the waste product in the blood led to brain damage in babies who had received one of the antibiotics being compared.<sup>13</sup>

Sometimes further analyses of randomized trials done in the past can help to identify less dramatic adverse effects. After it had been shown that the drug diethylstilboestrol (DES) given to women during pregnancy had caused cancer in the daughters of some of them, there was speculation about other possible adverse effects. These were detected by contacting the sons and daughters of the women who had participated in controlled trials. These follow-up studies revealed genital abnormalities and infertility in men as well as in women. More recently, when rofecoxib (Vioxx), a new drug for arthritis, was suspected of causing heart attacks, more detailed examination of the results of the relevant randomized trials showed that the drug did indeed have this adverse effect (see Chapter 1, p5-7).<sup>14</sup>

Follow-up of patients who have participated in randomized trials is obviously a very desirable way of ensuring that like will be compared with like when hunches about unanticipated effects of treatment are being investigated. Unfortunately, unless advance provision has been made for it, this is seldom an option. Investigating hunches about possible adverse effects of treatments

would present less of a challenge if contact details of people who have been participants in randomized trials were collected routinely. They could then be re-contacted and asked for further information about their health.

Investigation of suspected adverse effects of treatments is made easier if the suspected adverse effects concern a totally different health problem from the one for which the treatment has been prescribed.<sup>15</sup> For example, when Dr Spock recommended that babies should be put to sleep on their tummies, his prescription was for *all* babies, not those believed to be at above average risk of cot death (see Chapter 2, p13-14). The lack of any link between the prescribed advice ('put babies to sleep on their tummies') and the suspected consequence of the advice (cot death) helped to strengthen the conclusion that the observed association between the prescribed advice and cot death reflected cause and effect.

By contrast, investigating hunches that drugs prescribed for depression lead to an increase in the suicidal thoughts that sometimes accompany depression presents far more of a challenge. Unless there are randomized comparisons of the suspect drugs with other treatments for depression, it is difficult to assume that people who have and have not taken the drugs are sufficiently alike to provide a reliable comparison.<sup>16</sup>

### KEY POINTS

- Fair tests of treatments are needed because we will otherwise sometimes conclude that treatments are useful when they are not, and vice versa
- Comparisons are fundamental to all fair tests of treatments
- When treatments are compared (or a treatment is compared with no treatment) the principle of comparing 'like with like' is essential
- Attempts must be made to limit bias in assessing treatment outcomes