TESTING TREATMENTS
Chapter 6, 6.3

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.[4] 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

Want to see this Testing Treatments extract in context? click here

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?

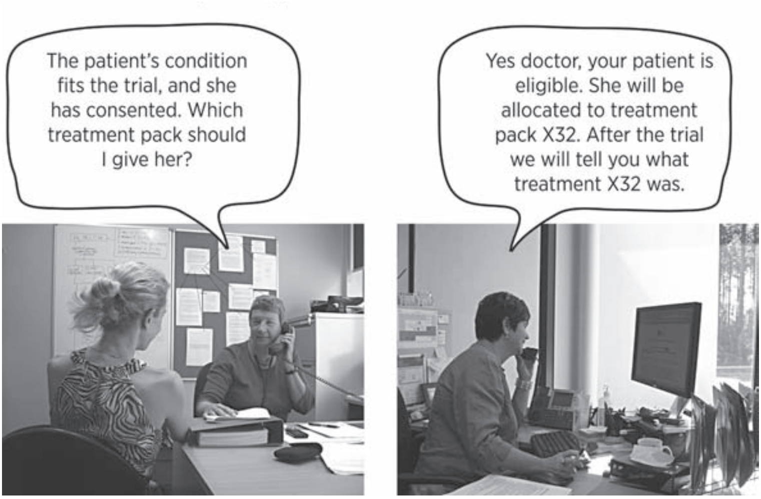
Unbiased, prospective allocation to different treatments

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.[5] 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).


The patient's condition fits the trial, and she has consented. Which treatment pack should I give her?
Yes doctor, your patient is eligible. She will be allocated to treatment pack X32. After the trial we will tell you what treatment X32 was.

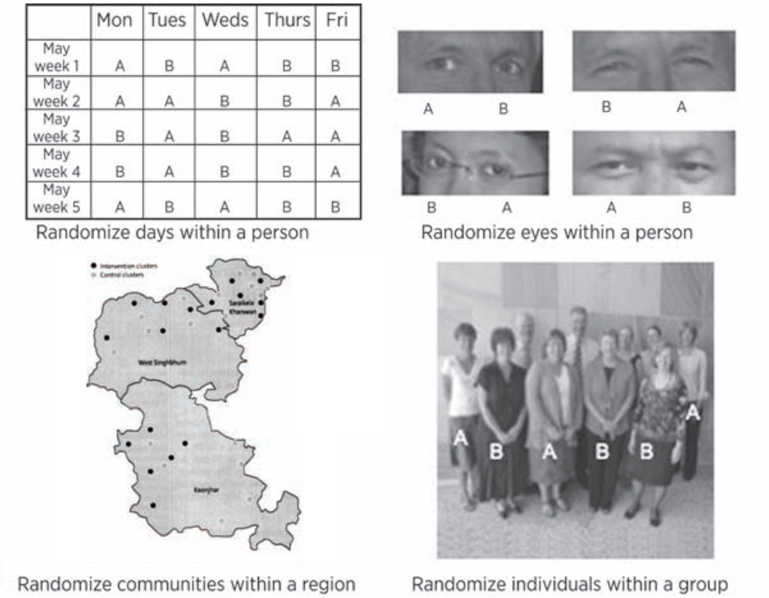
Concealing treatment allocation in a trial using telephone randomization.

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 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.[6]

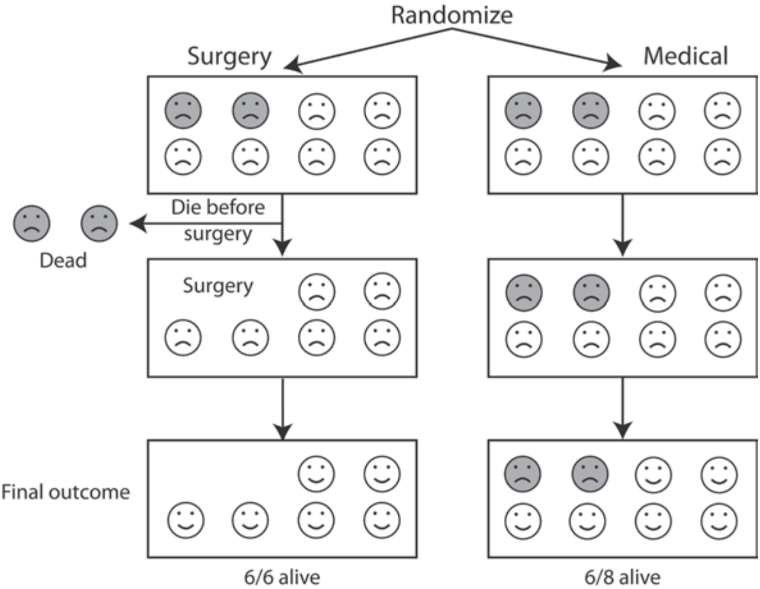
However by far the most common use of random allocation is its use to decide which patient will receive which treatment.

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



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.[7] 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 –