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

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