



Taking account of the play of chance

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

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

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

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

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

One approach that reduces the likelihood that we will be misled by chance effects involves estimating a range of treatment differences within which the real differences are likely to lie ([Gavarret 1840](#); Huth 2006). These range estimates are known as confidence intervals. As illustrated in the opening paragraph of this essay, repeating a treatment comparison is likely to yield varying estimates of the differential effects of treatments on outcomes, particularly if the estimates are based on small numbers of outcomes. Confidence intervals take account of this variation. Confidence intervals 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 a number of separate but similar studies - help us to take account of the play of chance and avoid concluding that treatment effects and differences exist when they don't, and don't exist when they do.

Cite As:

Editorial commentary (2007). Taking account of the play of chance. The James Lind Library

(www.jameslindlibrary.org).

[Home](#)

[Contents](#)