Interaction 2: compare effect sizes not \( P \) values

John N Matthews, Douglas G Altman

As we have previously described, \(^1\) the statistical term interaction relates to the non-independence of the effects of two or more factors. For example, in a controlled trial comparing a new treatment with a standard treatment we may want to examine whether the observed benefit was the same for different subgroups of patients. A common approach to answering this question is to analyse the data separately in each subgroup. Here we illustrate this approach and explain why it is incorrect.

One of several subgroups analyses in a trial of antenatal steroids for preventing neonatal respiratory distress syndrome\(^2\) was performed to see whether the effect of treatment was different in mothers who did or did not develop pre-ecampsia. Among mothers with pre-ecampsia 21.2\% (7/33) of babies whose mothers were given dexamethasone developed neonatal respiratory distress syndrome compared with 27.3\% (7/25) of babies whose mothers received placebo, giving \( P = 0.57\). Among mothers who did not have pre-ecampsia 7.9\% (21/267) of babies in the steroid group and 14.1\% (17/262) of babies in the placebo group developed neonatal respiratory distress syndrome, giving \( P = 0.021\).

There is a temptation to claim that the difference in \( P \) values establishes a difference between subgroups because “there is a treatment effect in mothers without pre-ecampsia but not in those with pre-ecampsia.” This argument is false: the key to realising this is to recall that a statement such as \( P = 0.57\) does not mean there is no difference, merely that we have found no evidence that there is a difference. A \( P \) value is a composite which depends not only on the size of an effect but also on how precise the effect has been estimated (its standard error). So differences in \( P \) values can arise because of differences in effect sizes or differences in standard errors or a combination of the two.

This is well illustrated by the present example. If we measure treatment effect by the difference in percentages developing neonatal respiratory distress syndrome in the placebo and steroid groups, then the treatment effect among mothers with pre-ecampsia, namely 27.3 - 21.2 = 6.1\%, is very close to the effect among mothers without pre-ecampsia, which is 14.1 - 7.9 = 6.2\%. The difference in \( P \) values has arisen only because the \( n\) of mothers pre-ecampsia (86 out of 595), so the former treatment effect is estimated much more precisely than the latter.

Another example can be found in a study of the effect of vitamin \( D \) supplementation for preventing neonatal hypocalcemic exposure mothers were given either supplements or placebo and the serum calcium concentration of the baby was measured at one week. The benefit of supplementation was investigated separately for breast and bottle fed infants, and a test to compare the treatment groups gave \( P = 0.40\) in the breast fed group and \( P = 0.0006\) in the bottle fed group.

As we have seen, it would be wrong to infer that vitamin \( D \) supplementation had a different effect on breast and bottle fed babies on the basis of these two \( P \) values: the correct way to proceed is to compare directly the sizes of the treatment effects. The effect of vitamin \( D \) supplementation can be measured by the difference in mean serum calcium concentrations between supplement and placebo groups and this gives effects of 0.04 mmol/L in the breast fed babies and 0.10 mmol/L in bottle fed babies. In order to interpret the difference in effect sizes, namely 0.06 mmol/L, we need to construct a confidence interval or perform a test of the null hypothesis that the true effect sizes are the same in each subgroup, a 95\% confidence interval for the difference in effect sizes is -0.05 to 0.17\% and a test of the null hypothesis gives \( P = 0.28\). There is then no evidence that the effect of vitamin \( D \) supplementation differs between breast and bottle fed infants. Comparing \( P \) values alone can be misleading.

Details of how to construct relevant confidence intervals and carry out associated tests are contained in a subsequent Statistics Note.

John N Matthews, Douglas G Altman

5. Bland JM, Altman DG.1994 SEE 48H Health Sciences, PO Box 777, Oxford OX1 3LF
6. Correspondence to: Dr Matthews.