BMJ

BMJ 2012;345:e7370 doi: 10.1136/bmj.e7370 (Published 5 November 2012)

ENDGAMES

STATISTICAL QUESTION

Effect sizes

Philip Sedgwick *reader in medical statistics and medical education*

Centre for Medical and Healthcare Education, St George's, University of London, Tooting, London, UK

Researchers assessed the effectiveness of an intervention designed to improve the mother-infant relationship and the security of infants in their attachment with their mother. A randomised controlled trial study design was used. Participants were pregnant women in a South African peri-urban settlement with adverse socioeconomic circumstances. The intervention, support and guidance in parenting, was delivered from late pregnancy and for six months post partum by community workers. Women in the control group received no therapeutic input.[1]

The main outcome measures included the quality of mother-infant interactions at six months post partum, as measured by the parent/caregiver involvement scale. The scale measured the responses of the mother to her infant's needs and initiations—in particular, maternal sensitivity and intrusiveness. The intervention was associated with significant benefit to the mother-infant relationship. At six months post partum, compared with the controls, the intervention group were significantly more sensitive (mean difference 0.77 (SD 0.37), P<0.05, effect size *d*=0.24) and less intrusive (mean difference 0.68 (SD 0.36), P<0.05, effect size *d*=0.26) in their interactions with their infants.

Which of the following statements, if any, are true?

a) Effect sizes are always positive in value.

b) Effect sizes are measured on the same scale as the outcome measure.

c) Effect sizes allow a direct comparison of different interventions on the same outcome.

d) As the difference between the intervention and control groups in maternal sensitivity was significant, it could be concluded that the effect size for the intervention was large.

Answers

Statements *c* is true, while *a*, *b*, and *d* are false.

The trial investigated the effectiveness of an intervention designed to improve the mother-infant relationship and the security of infants in their attachment with their mother. The researchers reported that at six months post partum, when compared with controls, mothers in the intervention group were significantly more sensitive and less intrusive in their interactions. Although the difference between intervention and control was statistically significant, it did not necessarily mean that the difference was important or would be useful in decision making. Before the trial, the researchers would have performed a sample size calculation to identify the smallest effect of clinical interest in the main outcome variables (maternal sensitivity and intrusiveness at six months), such that if a difference existed between treatment groups it would be significant. The smallest effect of clinical interest represented the smallest difference required for the intervention to be considered clinically effective in comparison with the control intervention. However, it could be difficult to compare the results of the above trial with other studies that investigated the effectiveness of the same or a different intervention on the same outcome. Comparing studies on the basis of statistical significance alone does not necessarily allow a direct comparison of the effectiveness of the interventions. In particular, other trials may have used a smallest effect of clinical interest of a different size when calculating sample size, or a trial may or may not have produced a statistically significant result because of sampling error. Sampling error has been described in a previous question.[2]

To establish the importance of the difference between treatment groups in each outcome measure, the effect size *d* for the difference was calculated. The effect size *d* is sometimes referred to as Cohen's *d* after the statistician who first suggested it. In the above example, the effect sizes for each of the outcome measures of maternal sensitivity and intrusiveness were calculated as the difference between the treatment group sample means—that is, intervention minus control, divided by a pooled standard deviation for the scores of the outcome measure across treatment groups. Sometimes the mean difference between treatment groups is divided by the sample standard deviation of one of the groups, typically the control group.

Cohen's *d* is a measure of the relevant magnitude of the difference between two groups in an outcome measure, expressed as a multiple of the standard deviation of the outcome scores. As described in a previous question,[3] the standard deviation describes the spread of measurements in a distribution; approximately 99% of the sample measurements for a variable

p.sedgwick@sgul.ac.uk

will be contained within the range of three standard deviations either side of the mean. Therefore, the mean for the intervention group will typically be no further than three standard deviations away from the mean for the control group; hence for most practical purposes *d* will have a value between −3 and 3 (*a* is false). The sample mean for the control group was subtracted from the intervention group mean, and therefore whether *d* was positive or negative depended on whether an increase or decrease in the mean outcome measure was beneficial. In the above example, the intervention was associated with significant benefit to the mother-infant relationship. The score on sensitivity was raised for the intervention, when compared with control, giving an effect size that was positive. The score on intrusiveness was reduced for the intervention, when compared with control, giving an effect size that was negative. Nevertheless, as is typical, the effect sizes for both outcome measures were both presented as positive.

Effect sizes are a standardised measure and are on a common scale. They have no units; the difference between treatment groups is standardised to the same scale regardless of the scale on which the outcome measure was originally measured (*b* is false). This allows the effectiveness of different trials or interventions on the same outcome to be compared (*c* is true).

The larger the effect size, the greater the difference between treatment groups in the outcome measure. There are no universally accepted standards for describing values of *d*. However, it has been suggested that an effect size of 0.8 (8/10ths of a standard deviation) is "large," a value of 0.5 (half a standard deviation) is "medium," and a value of 0.2 (one fifth of a standard deviation) is small. Although the difference between the intervention and control groups in maternal sensitivity was significant, the effect size of intervention was small to medium (*d* is false). Similarly, for the outcome measure of intrusiveness the difference between the intervention and control groups was significant, yet the effect size of intervention was small to medium. Therefore, although the trial showed significant results on the basis of estimated smallest effects of clinical interest between treatment groups, the actual effect sizes are regarded as small.

Effect sizes are often used to calculate sample sizes for trials. However, effect sizes can be calculated only after data have been collected from the study participants. Therefore, it is necessary to use an estimate for *d*. An effect size of at least 0.5 is typically used in sample size calculations, as it would indicate a medium (or bigger) treatment effect.

Competing interests: None declared.

1 Cooper PJ, Tomlinson M, Swartz L, Landman M, Molteno C, Stein A, et al. Improving quality of mother-infant relationship and infant attachment in socioeconomically deprived community in South Africa: randomised controlled trial. *BMJ* 2009;338:b974.
2 Sedgwick P. What is sampling error? *BMJ* 2012;344:e4285.
3 Sedgwick P. Standard deviation versus standard error. *BMJ* 2011;343:d8010.

Cite this as: *BMJ* 2012;345:e7370