BMJ 2015;350:h1690 doi: 10.1136/bmj.h1690 (Published 27 March 2015)

STATISTICAL QUESTION

Randomised controlled trials: understanding effect sizes

Philip Sedgwick *reader in medical statistics and medical education*

Institute for Medical and Biomedical Education, St George's, University of London, London, UK

Researchers assessed whether a multimodal group exercise intervention had physiological, functional, and emotional benefits for patients with cancer. A randomised controlled trial was performed. The intervention was supervised exercise comprising high intensity cardiovascular and resistance training, relaxation and body awareness training, and massage. The intervention was delivered for nine hours weekly for six weeks in addition to conventional care. The control treatment was conventional care alone. Participants were 269 patients with cancer undergoing adjuvant chemotherapy or treatment for advanced disease. The primary outcome was fatigue at six weeks as assessed using the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30); scores range from 0 to 100 and higher scores represent higher levels of fatigue. Secondary outcomes included measures of physical capacity, general wellbeing, physical activity, and quality of life.[1]

At six weeks of follow-up, the fatigue score was significantly lower in the intervention group than in the control group (adjusted difference 6.6 points, 95% confidence interval 0.9 to 12.3; P=0.02). The difference between treatment groups in outcome was adjusted for baseline fatigue score, disease, and demographic covariates. The observed reduction in fatigue after the intervention compared with the control had an effect size of 0.33. The researchers also reported that the intervention significantly improved vitality, aerobic capacity, muscular strength, emotional wellbeing, and physical and functional activity. No significant effect was seen for quality of life.

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

a) The effect size was calculated using the standard error of the mean difference between treatment groups in fatigue

b) The effect size was measured on the same scale as the original measurement of fatigue

c) The effect size permitted a direct comparison of treatment effects across trials that used different scales to assess fatigue

d) Effect sizes are always positive in value

e) The effect size can be described as large

Answers

Statement *c* is true, whereas *a*, *b*, *d*, and *e* are false.

It was reported that the intervention significantly reduced fatigue when compared with the control (adjusted difference 6.6 points, 0.9 to 12.3; P=0.02). The mean difference between treatment groups in the outcome of fatigue is called the treatment effect, and it compares the effectiveness of the intervention with that of the control treatment.[2] The treatment effect was quantified by the effect size, estimated to be 0.33. It was calculated as the difference between the treatment group means—that is, the mean value in the control group minus the mean value in the intervention group, divided by a pooled standard deviation for the outcome measure scores across treatment groups (*a* is false). The mean difference between treatment groups is sometimes divided by the sample standard deviation for one of the groups, typically the control group. Effect sizes therefore take into account the variability in outcome scores. This is important when determining the magnitude of the difference between treatment groups in outcome. The smaller the variability in the outcome measure the greater the difference between treatment groups in treatment effect. This results in a larger effect size. The effect size is denoted by *d* and occasionally referred to as Cohen's *d* after the statistician who first suggested it.

An effect size is a measure of the magnitude of the difference between treatment groups in the primary outcome, expressed as a multiple of the standard deviation of the outcome scores. In the example above, the intervention reduced fatigue compared with the control after six weeks of follow-up and the effect size was 0.33. Therefore, the intervention group had a mean fatigue score that was 0.33 standard deviations smaller than that for the control group.

The effect size is a ratio, and the numerator and denominator are measured in the same units as the original outcome. The measure therefore has no units and does not depend on the original measurement scale (*b* is false). Hence, the effect size is a standardised measure that is measured on a common scale. This permits a direct comparison of the treatment effect across

p.sedgwick@sgul.ac.uk

trials that used different scales to assess the outcome and would otherwise not be directly comparable (*c* is true). Effect sizes are commonly used in meta-analyses and have similar properties to those of standardised mean differences, which have been described in a previous question.[3]

The effect size may be positive or negative (*d* is false), depending on whether the sample mean for the control group is subtracted from the intervention group mean and whether an increase or decrease in the outcome measure is beneficial. Nevertheless, effect sizes for outcome measures are typically presented as positive. In the above example, the primary outcome was fatigue at six weeks as assessed using the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30); scores range from 0 to 100 and a higher score represents higher levels of fatigue. The intervention resulted in a reduction in the mean fatigue symptom score; the sample mean for the intervention group was subtracted from the control group mean, resulting in a positive effect size.

Generally 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 0.2 (one fifth of a standard deviation) is “small,” a value of 0.5 (half a standard deviation) is “medium,” a value of 0.8 (8/10ths of a standard deviation) is large, and an effect size of 1.3 (one and a third of a standard deviation) is “very large.”

Although the reduction in fatigue for the intervention compared with the control was significant, the effect size of the intervention was small to medium (*e* is false). This is probably because the trial was overpowered,[4] resulting in a reduction in fatigue that was statistically significant yet not clinically significant. The observed mean difference between treatment groups in fatigue was 6.6 points. However, the researchers reported that the sample size calculation for the trial was based on the smallest effect of clinical interest of 10 points on the fatigue scale. To estimate the smallest effect of clinical interest with 80% power, 98 patients were needed in each treatment group, and this number was adjusted to 135 to take into account expected dropout during follow-up. In total, 269 patients were recruited and randomised to the intervention (n=135) or control (n=134). Overall 235 patients (intervention 118; control 117) completed follow-up and so there were more patients than were needed to demonstrate the smallest effect of clinical interest with 80% power. As a result, the power of detecting the smallest effect of clinical interest was greater than 80% and the trial is referred to as being overpowered. However, a consequence of the larger sample size was that a treatment effect smaller than the smallest effect of clinical interest was identified as statistically significant while statistical power was maintained at about 80%. Hence a reduction in fatigue that was not clinically significant was identified as statistically significant.

The effect size facilitates the comparison of treatment effects between clinical trials on a common scale. However, the use of the effect size and the expressions small, medium, large, and very large to describe the effectiveness of intervention may be problematic if used out of context. Effect sizes should be used to compare treatment effects only if the intention is to investigate the effectiveness of similar treatments with regard to the same outcome (although possibly measured on different scales). Furthermore, the practical importance of the effectiveness of a treatment depends on the relative clinical benefits and costs. The effect size *d* is not routinely reported as part of the results in clinical studies. The presentation of the effect size in addition to the 95% confidence interval for the population parameter of the treatment effect and the P value could help inform decision making. Effect sizes can be derived for statistics other than mean differences between treatment groups in outcome, including relative risks and correlation coefficients. These will be discussed in a later question.

Competing interests: None declared.

1 Adamsen L, Quist M, Andersen C, et al. Effect of a multimodal high intensity exercise intervention in cancer patients undergoing chemotherapy: randomised controlled trial. *BMJ* 2009;339:b3410.

2 Sedgwick P. Treatment effects and placebo effects. *BMJ* 2015;350:h267.

3 Sedgwick P, Marston L. Meta-analyses: standardised mean differences. *BMJ* 2013;347:f7257.

4 Sedgwick P. Randomised controlled trials: the importance of sample size. *BMJ* 2015;350:h1586.

Cite this as: *BMJ* 2015;350:h1690