BMJ 2015;350:h3475 doi: 10.1136/bmj.h3475 (Published 26 June 2015)

ENDGAMES

STATISTICAL QUESTION

What is significance?

Philip Sedgwick *reader in medical statistics and medical education*

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

The effects of pelvic floor muscle training on pelvic floor symptoms were investigated using a randomised controlled trial. The intervention consisted of pelvic floor muscle training combined with home exercises. The control intervention consisted of watchful waiting. The length of follow-up was three months. The participants were women recruited from a primary care population, aged 55 years or more, who had symptomatic mild pelvic organ prolapse.[1]

The primary outcome was the change in bladder, bowel, and pelvic floor symptoms at follow-up from baseline as measured by the Pelvic Floor Distress Inventory-20 (PFDI-20). Higher scores on the inventory indicated a greater severity of symptoms. To calculate the required sample size it was assumed that the watchful waiting group would have a PDFI-20 score of 60 points at baseline with no subsequent change in symptoms at three months. The sample size was based on having 80% power to detect a difference between treatment groups of 15 points (25% reduction) in the PFDI-20 score, with a standard deviation of 36 points at three month follow-up. To achieve this difference using a critical level of significance of 0.05 and two sided alternative hypothesis, 92 women were needed in each treatment arm. To account for an estimated dropout rate of 15%, the required sample size was adjusted to 216.

In total, 287 women were recruited and randomised to pelvic floor muscle training (n=145) or watchful waiting (n=142). Overall, 250 (87%) women completed follow-up. At the end of follow-up the intervention group had a significant improvement in symptoms compared with the watchful waiting group, with an average reduction of 9.1 (95% confidence interval 2.8 to 15.4; P=0.005) points on the PFDI-20.

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

a) The proposed difference between treatment groups of 15 points on the PFDI-20 used to calculate the sample size was the smallest effect of clinical interest

b) The intervention would be considered clinically effective if the intervention group had an improvement in mean PFDI-20 score of 15 points or more compared with the control group

c) Because the difference between treatment groups in the primary outcome was statistically significant, it can be inferred that pelvic floor muscle training was clinically effective

d) The trial was overpowered for the statistical test of the difference between treatment groups in the primary outcome

Answers

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

When compared with the watchful waiting group the intervention group showed an improvement in bladder, bowel, and pelvic floor symptoms, with a mean reduction of 9.1 points on the PFDI-20 score. The trial was designed as a superiority one, described in a previous question.[2] The null hypothesis stated that there was no difference between the treatment groups in the change in symptoms from baseline in the population from which the sample was obtained. The alternative hypothesis was two sided and stated that a difference existed between treatment groups. The critical level of significance for statistical testing was set at 0.05 (5%). The reported P value (0.005) measured the strength of the evidence in support of the null hypothesis. Because the P value was smaller than the critical level of significance, there was little evidence to support the null hypothesis, and it was rejected in favour of the alternative hypothesis. The inference was that the intervention group showed a statistically significant reduction in symptoms compared with the control group. The inference of statistical significance could have been made on the basis of the 95% confidence interval for the difference between treatment groups in mean PFDI-20 score (2.8 to 15.4), which permits a test of the statistical hypotheses at the 5% level.[3] Because the 95% confidence interval did not include zero, it can be inferred that the P value for the test of the statistical hypotheses was smaller than the critical level of significance (0.05).

It was essential that sample size was considered when planning the trial. The number of women required was based principally on clinical significance. It was assumed that the control group would have a mean PDFI-20 score of 60 points at baseline and would not show any change in symptoms after three months of follow-up. For the intervention to be considered clinically

Correspondence to: P Sedgwick p.sedgwick@sgul.ac.uk

effective, the pelvic floor muscle training group had to show at least a 25% reduction (15 points) in mean PFDI-20 score compared with the control group at follow-up. The difference of a 25% reduction (15 points) in the mean score is called the smallest effect of clinical interest (*a* is true). Differences larger than the smallest effect of clinical interest—that is, 15 points or more on the PFDI-20 score—would be regarded as clinically significant and the intervention clinically effective (*b* is true), whereas smaller ones would not. The smallest effect of clinical interest may not exist in the population. That is, the difference in mean symptom scores that would be seen between treatments groups if applied to the entire population may be less than 15 points. However, if the smallest effect of clinical interest does exist for the population, the probability that it will be seen in the trial needs to be maximised. This underlies the concept of statistical power, as described in a previous question.[4] Although it is desirable for statistical power to be as large as possible, an increase in power results in an increase in sample size. Therefore, a compromise between power and sample size is made. The power was set to 80% in the above trial, which is the minimum generally recommended when calculating sample size in clinical trials. The derived sample size was the number of women who needed to be recruited to demonstrate the smallest effect of clinical interest (or a larger difference) as statistically significant in the trial if that effect existed in the population.

At the end of follow-up the intervention group showed an improvement in symptoms compared with the watchful waiting group, with an average reduction of 9.1 points on the PFDI-20 score. This difference was statistically significant (P=0.005). However, although the reduction in symptoms was statistically significant, the researchers concluded that it was probably not clinically significant or relevant (*c* is false). This was because the reduction in symptoms was less than the smallest effect of clinical interest (15 points). However, the authors indicated that because limited data are available on the minimal clinically important difference for the PFDI-20 questionnaire in women with mild prolapse, the proposed smallest effect of clinical effect may have been too large.

The most likely explanation for the difference in PFDI-20 scores being statistically significant although it was smaller than the smallest effect of clinical interest was that the trial was overpowered (*d* is true). In particular, the actual power of the trial was in excess of 80%, as originally specified, because more women than needed were recruited. To observe the smallest effect of clinical interest with 80% power, it was estimated that 184 women needed to be recruited. The required sample size was adjusted for an estimated dropout rate of 15%, resulting in a total sample size of 216. In total, 287 women were recruited, of whom 250 (87%) completed follow-up. As described above, an increase in sample size results in an increase in power. Hence, the power for the above trial based on the specified smallest effect of clinical interest was greater than 80%, making the trial overpowered. A disadvantage of having a larger than needed sample size was that statistical power could be maintained close to 80%, and if differences between treatment groups smaller than the smallest effect of clinical interest were observed they might have been statistically significant.

The concept of significance is important in clinical research and clinical medicine. If the difference between treatment groups in an outcome for a trial is statistically significant, it implies that the observed difference also exists in the population. Clinical significance implies that the difference between treatments in effectiveness is clinically important, and it is possible that practice will change if such a difference is seen in a trial. If statistical significance exists then it may be used to inform clinical significance. In particular, this would be the case for a clinical trial, where clinical significance is used to obtain the required sample size so that statistical significance will most likely be observed in the trial if the smallest effect of clinical interest exists in the population. Unfortunately, clinical significance and statistical significance are often confused. The terms are often used interchangeably, although one does not necessarily imply the other. Researchers sometimes infer that the effectiveness of a treatment is clinically significant because the difference between treatments is statistically significant. However, clinical significance cannot necessarily be inferred from statistical significance, and statistical significance cannot be inferred from clinical significance. As in the trial above, a statistically significant difference existed between treatment groups in the primary outcome, yet the researchers concluded that it was unlikely to be of clinical significance.

Presumably the confusion between clinical significance and statistical significance exists because it is difficult to set aside the everyday meaning of clinical significance when discussing statistical significance. In statistics the use of the word significant does not imply that something is important. It may be straightforward to consider the association between statistical significance and clinical significance in trials, particularly for the primary outcome measure on which the sample size calculation is based. However, the importance of statistically significant results in observational studies may be less easy to discern. This is particularly true for datasets that consist of tens of thousands of cases, as is typical of cohort studies nowadays, and many comparisons or estimates of effects are statistically significant. The interpretation of statistical significance in observational studies will be discussed in a later question.

In addition to clinical significance and statistical significance, the concept of "patient significance" has been suggested. In the above trial, the smallest effect of clinical interest in the measurement of symptoms using the PFDI-20 was chosen on the basis of clinical expertise and previous research. It was not indicated whether the smallest effect of clinical interest was important to participants. Although any improvement in symptoms would presumably be of benefit to women, the significance of such changes may be based on the time and effort needed on their behalf. Patient groups are becoming increasingly involved in the design of clinical trials, enabling their views of significant treatment effectiveness to be acknowledged.

Competing interests: None declared.

1 Wiegersma M, Panman CMCR, Kollen BJ, et al. Effect of pelvic floor muscle training compared with watchful waiting in older women with symptomatic mild pelvic organ prolapse: randomised controlled trial in primary care. *BMJ* 2014;349:g7378.
2 Sedgwick P. What is a superiority trial? *BMJ* 2013;347:f5420.
3 Sedgwick P. Randomised controlled trials: inferring significance of treatment effects based on confidence intervals. *BMJ* 2014;349:g5196.
4 Sedgwick P. Randomised controlled trials: understanding power. *BMJ* 2015;350:h3229.

Cite this as: *BMJ* 2015;350:h3475