BMJ

BMJ 2013;347:f6282 doi: 10.1136/bmj.f6282 (Published 18 October 2013)

ENDGAMES

STATISTICAL QUESTION

The importance of statistical power

Philip Sedgwick *reader in medical statistics and medical education*

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

The effectiveness of a home based early childhood intervention on children's body mass index (BMI) at age 2 years was investigated. A randomised controlled superiority trial was used. The intervention consisted of eight home visits from specially trained community nurses in the first 24 months after birth. The intervention was in addition to the usual childhood nursing service from community health service nurses. The control group received the usual childhood nursing service alone. Participants were first time mothers and their infants.[1]

The primary outcome was children's BMI at age 2 years. The sample size calculation was based on having 80% power to detect a difference in mean BMI of 0.38 units between treatment groups at age 2 years, using a two sided hypothesis test and critical level of significance of 0.05. It was assumed that the standard deviation of observations in each group was the same and equal to 1.5 units. A total sample size of 504 participants (252 in each treatment arm) was needed. To allow for an estimated 25% drop-out rate the sample size was increased to 630 participants. In total, 667 first time mothers and their infants were recruited to the trial, with 337 allocated to intervention and 330 to control.

At age 2 years, mean BMI was significantly lower in the intervention group compared with the control group (16.53 *v* 16.82; difference −0.29, 95% confidence interval −0.55 to −0.02; P=0.04).

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

a) The difference in mean BMI of 0.38 between treatment groups is called the smallest effect of clinical interest

b) An increase in statistical power would require a smaller sample size

c) The trial was overpowered for the statistical test of the primary outcome

d) It can be inferred that a clinically important difference existed between treatment groups in the primary outcome because of the significant result (P=0.04).

Answers

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

The purpose of the above trial was to investigate the effectiveness of a home based early intervention on children's BMI at age 2 years. Control treatment was the usual childhood nursing service. The trial was designed as a randomised controlled superiority trial, as described in a previous question.[2]

For one of the treatments to have been considered clinically superior to the other in effectiveness, a significant difference of 0.38 units in mean BMI at age 2 years was required. This difference was called the smallest effect of clinical interest (*a* is true) and was proposed by the researchers on the basis of clinical experience or previous research. Obviously, larger differences between treatment groups would show clinical superiority, whereas smaller differences would not.

The smallest effect of clinical interest (0.38 units) may not exist for the population. That is, the difference in BMI at 2 years that would be seen between treatments groups if applied to the entire population of first time mothers and their infants may be less than 0.38 units. However, if the smallest effect of clinical interest does exist for the population, then the probability that it will be seen in the trial needs to be maximised. To do this, an optimal sample size is needed. This underlies the concept of statistical power. Statistical power is based on the hypothetical situation of repeating the above trial an infinite number of times and under the same conditions. Each trial would involve a statistical hypothesis test with a derived P value. The percentage of these repeated samples that would demonstrate the smallest effect of clinical interest (if it existed in the population) as a significant difference (P<0.05) is the statistical power of the calculated sample size in the above trial. To calculate the required sample size, in addition to the smallest effect of clinical interest and power, it was necessary to specify the critical level of significance and to provide some indication of the expected standard deviation of BMI at age 2 years. The standard deviation of BMI was assumed to be equal in each group and was based on previous research.

It was obviously essential that statistical power was as high as possible in the above trial. However, increased statistical power is associated with a larger sample size (*b* is false). This is intuitive, because as sample size increases and approaches that of the population, the observed difference in BMI in the trial

p.sedgwick@sgul.ac.uk

would become similar to that seen in the population. Therefore, as sample size increases so does power, because the smallest effect of clinical interest is more likely to be seen in the trial, if it exists in the population. To have 100% statistical power would require sampling the entire population, but this is not feasible. Therefore, a compromise was made between power and sample size in the above trial. The power was set to 80%, this being the minimum generally recommended when calculating sample size in clinical trials.

Determining the optimal sample size before starting the trial was an important ethical consideration. The trial needed to be adequately powered. If the sample size was too small it would not have inadequate power and might fail to detect the smallest effect of clinical interest, if it existed in the population. This would be considered unethical because time, effort, and resources might have been wasted in running a trial that had little potential to show clinical significance. Equally, too large a sample size would have recruited more participants than needed to show the smallest effect of clinical interest. The trial would be over powered. This would also be unethical because time, effort, and resources would have been wasted in recruiting too many participants.

It is important to appreciate the association between sample size and statistical significance when making conclusions based on study results. As described above, increased statistical power is associated with a larger sample size. However, as sample size and power increase, progressively smaller differences between treatment groups in the primary outcome will be identified as statistically significant. It is possible that differences smaller than the specified smallest effect of clinical interest would be identified as statistically significant. Therefore, differences between treatment groups identified as statistically significant may not be clinically significant.

In the above trial, the smallest effect of clinical interest was a difference of 0.38 units in BMI between treatment groups at age 2 years. However, the actual difference seen was 0.29 units. Although this difference was smaller than the smallest effect of clinical interest it was still significant at the 5% level of significance. This was because the trial was overpowered (*c* is true)—that is, the power was greater than 80% as specified in the sample size calculation. Several reasons may have accounted for this. Sample size calculations provide rough estimates of the number of participants, not least because some of the information required is difficult to predict. This includes the standard deviation of the primary outcome common to both treatment groups, which the researchers predicted to be 1.5 units. However, the actual standard deviation of the observations for the primary outcome was smaller. This meant the power of the trial was greater than 80%. Furthermore, the researchers recruited more subjects than was necessary. The required sample size was 504 but the researchers aimed to recruit 630 participants owing to an anticipated drop-out rate of 25%. This was close to the observed drop-out rate of about 26.5%. However, 667 participants were recruited in total, so the power of the study was increased further.

Despite ethical considerations, increased power in the above trial may have been beneficial because if the smallest effect of clinical interest did exist for the population then it would be more likely to be demonstrated. However, the downside of increased power was that the observed difference between treatment groups was significant although it was smaller than the smallest difference of clinical interest. Care is needed when interpreting results from overpowered studies, because statistical significance can be confused with clinical significance. This may have occurred in the trial above. On the basis of the statistical significance of the hypothesis test (P=0.04), the researchers concluded that home based early intervention delivered by trained community nurses was effective in reducing mean BMI for children at age 2 years. However, this is not the case because the smallest effect of clinical interest was specified as 0.38 before the trial started (*d* is false).

Competing interests: None declared.

1 Wen LM, Baur LA, Simpson JM, Rissel C, Wardle K, Flood VM. Effectiveness of home based early intervention on children's BMI at age 2: randomised controlled trial. *BMJ* 2012;344:e3732 [correction *BMJ* 2013;346:f1650].
2 Sedgwick P. What is a superiority trial? *BMJ* 2013;347:f5420.

Cite this as: *BMJ* 2013;347:f6282