BMJ

BMJ 2013;346:f1041 doi: 10.1136/bmj.f1041 (Published 15 February 2013)

ENDGAMES

Sample size: how many participants are needed in a trial?

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 investigated the effectiveness of a home based early intervention on children's body mass index (BMI) at age 2 years. 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; this 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. The primary outcome was children's BMI at age 2.[1]

The sample size calculation was based on having 80% power to detect a difference in mean BMI of 0.25 units between treatment groups at age 2, 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 BMI units. A total sample size of 504 participants (252 in each treatment arm) was required. To allow for an estimated 20% dropout rate it would be necessary to recruit 630 first time mothers. In total, 667 first time mothers and their infants were recruited to the trial.

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). The researchers concluded that a home based early intervention delivered by trained community nurses was effective in reducing mean BMI in children at age 2 years.

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

a) A difference in mean BMI of 0.25 between treatments was the smallest effect of clinical interest

b) If power was increased to 90%, the required sample size would increase

c) The type I error rate was fixed at 5% for the statistical test of the primary outcome

d) Increasing sample size will lead to a reduction in the type I error rate

e) It can be concluded that a difference in mean BMI of at least 0.25 between treatments definitely exists in the population

Answers

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

The aim of this superiority trial was to establish whether home based early intervention was superior in effectiveness to the control treatment, or whether the control treatment was superior. Superiority trials have been described in a previous question.[2] Although it was predicted that home based early intervention would reduce mean BMI at age 2 compared with the control treatment, sometimes results are unexpected and it was important that statistical hypothesis testing allowed for the possibility of the control treatment being superior. Therefore, traditional statistical hypothesis testing with a two sided alternative hypothesis was used to compare treatment groups in the outcome measure of BMI.[3]

One of the treatments would have been considered more effective than the other if the difference in mean BMI between treatment groups was at least 0.25 units. This difference is called the smallest effect of clinical interest (*a* is true), and represents the smallest difference in mean BMI needed for one treatment to be considered clinically more effective than the other. Larger differences would obviously also demonstrate superiority—that is, a significant difference between treatment groups. However, a significant difference between treatment groups would not be demonstrated with the calculated sample size if the difference between treatment groups was smaller. The smallest effect of clinical interest was proposed by the researchers on the basis of clinical experience or previous research.

The observed difference in BMI between treatment groups in the trial estimated the population effect—that is, the difference that would be seen between treatments groups if applied to the entire population of first time mothers and their infants. The smallest effect of clinical interest may not exist in the population, but if it does, the probability that it will be seen in the trial needs to be maximised. To maximise this probability, an optimal sample size was needed. To calculate the sample size, in addition to specifying the smallest effect of clinical interest, the researchers needed to specify the required power and critical level of significance; they also needed to provide some indication of the expected standard deviation of BMI in each treatment group. The standard deviation of BMI was

p.sedgwick@sgul.ac.uk

assumed to be equal in each group and was based on previous research.

To establish whether the observed difference in mean BMI was significant, a statistical hypothesis test was undertaken and P value derived. Hypothesis testing is based on the hypothetical situation of sampling an infinite number of times. For the example above, each of the infinite number of samples would be exactly the same size and obtained under the same conditions. Power is the percentage of these repeated samples (set at 80% in the example above) that would demonstrate the smallest effect of clinical interest if it existed in the population.

It is generally recommended that power is set to a minimum of 80% when calculating sample size. Typically power is fixed at 80% or 90%. Increasing power in a sample size calculation has the effect of increasing the required sample size (*b* is true). This may be intuitive, because as sample size increases and approaches that of the population, the observed difference in BMI in the trial will become similar to that 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 compare the intervention and control groups, a two sided hypothesis test with a critical level of significance of 0.05 was proposed. The critical level of significance is typically set at 0.05 in statistical hypothesis testing. Obviously, before the trial started it was not known if there was a difference in mean BMI between treatments in the population. If no difference existed, it was important that the probability of making a type I error was minimised. A type I error would occur if the null hypothesis was rejected in favour of the alternative when there was no difference in mean BMI between treatments in the population. Setting the critical level of significance in advance ensured that the maximum probability of a type I error occurring was 0.05 (5%) (*c* is true).

As described above, hypothesis testing is based on the hypothetical situation of sampling an infinite number of times. Because the critical level of significance was set at 0.05, the null hypothesis would be rejected in favour of the alternative for 5% of these infinite number of samples. Therefore, for any hypothesis test the maximum probability of rejecting the null hypothesis is 0.05. Because any hypothesis test could result in a type I error, the maximum probability of a type I error was 0.05 (*c* is true). The probability of making a type I error is influenced by sample size. As sample size increases and approaches that of the population, the difference in mean BMI in the trial will become similar to that in the population, making it less likely that a type I error will occur (*d* is true).

Although the study found a significant difference between treatments in mean BMI at age 2 years, it cannot be concluded that a difference in mean BMI of at least 0.25 units (smallest effect of clinical interest) definitely exists between treatments in the population (*e* is false). The trial provided sufficient evidence to reject the null hypothesis in favour of the alternative, with the conclusion that a difference exists between treatments. However, it is always possible that this result was a type I error, although, as described above, the probability of this was at most 0.05 (5%). It is not possible to say whether this significant result is a type I error.

It was essential that the researchers calculated the optimal sample size. If the sample size had been too small it may not have been representative of the population, and this could have led to the trial lacking power. Too large a sample may have been time consuming, expensive, and possibly unethical. The required sample size was adjusted for an estimated 20% dropout rate. It is not uncommon for participants to leave a trial for a variety of reasons, so the sample size had to be adjusted to account of this. The extent of dropout would have been estimated from previous trials.

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.

2 Sedgwick P. Superiority trials. *BMJ* 2011;342:d2981.

3 Sedgwick P. Statistical hypothesis testing. *BMJ* 2010;340:c2059.

Cite this as: *BMJ* 2013;346:f1041