How many do I need? Basic principles of sample size estimation

Declan Devane BSc MSc RM RGN RNT
Midwifery Doctoral Student and Research Assistant, School of Nursing and Midwifery Studies, University of Dublin Trinity College, Dublin, Ireland

Cecily M. Begley MSc PhD RGN RM RNT FFNRCrSI
Chair of Nursing and Midwifery; and Director, School of Nursing and Midwifery Studies, University of Dublin Trinity College, Dublin, Ireland

Mike Clarke DPhil
Director and Co-Chair of the Cochrane Collaboration Steering Group, UK Cochrane Centre, Oxford, UK

Submitted for publication 4 December 2003
Accepted for publication 9 February 2004

Correspondence:
Declan Devane,
School of Nursing and Midwifery Studies,
University of Dublin Trinity College,
St James’s Hospital,
James Street,
Dublin 8,
Ireland.
E-mail: ddevane@tcd.ie

How many do I need? Basic principles of sample size estimation

Background. In conducting randomized trials, formal estimations of sample size are required to ensure that the probability of missing an important difference is small, to reduce unnecessary cost and to reduce wastage. Nevertheless, this aspect of research design often causes confusion for the novice researcher.

Aim. This paper attempts to demystify the process of sample size estimation by explaining some of the basic concepts and issues to consider in determining appropriate sample sizes.

Method. Using a hypothetical two group, randomized trial as an example, we examine each of the basic issues that require consideration in estimating appropriate sample sizes. Issues discussed include: the ethics of randomized trials, the randomized trial, the null hypothesis, effect size, probability, significance level and type I error, and power and type II error. The paper concludes with examples of sample size estimations with varying effect size, power and alpha levels.

Conclusion. Health care researchers should carefully consider each of the aspects inherent in sample size estimations. Such consideration is essential if care is to be based on sound evidence, which has been collected with due consideration of resource use, clinically important differences and the need to avoid, as far as possible, types I and II errors. If the techniques they employ are not appropriate, researchers run the risk of misinterpreting findings due to inappropriate, unrepresentative and biased samples.

Keywords: sample size, power, midwifery/nursing research

Introduction

Few activities in conducting research appear to cause as much concern to the novice researcher as estimating the sample size required for a quantitative study. How big is big enough? The importance of formal estimation of the required sample size is shown by the increasing number of health-related journals requiring that randomized controlled trials (RCTs) are described in accordance with the Consolidated Standards of Reporting Trials (CONSORT) statement (Moher et al. 2001), which includes the need to report the methods for determining sample sizes. In addition, many grant awarding bodies demand sample size justification in an effort to ensure that the studies they fund produce as unambiguous results as possible.
Finally and perhaps most importantly, if a sample size is too small then the study might be a waste of resources and potentially unethical, because any ‘true’ difference in outcome between the interventions under study is unlikely to be differentiated from the effects of chance. The research might not detect a true difference between two different interventions simply because the trial was too small to show it. Similarly, too large a sample size may also be a waste of resources, as more participants than needed would be recruited to the study. Again, this raises ethical issues because more participants than necessary would be exposed to a potentially less beneficial intervention.

In the past, nursing and midwifery research in particular has been criticized for its poor level of statistical quality (Brogan 1989). Sampling theory, in particular, was recommended as something that nurse and midwife researchers should address in order to improve the rigour of their work (Nadzam et al. 1989). This paper attempts to demystify some of the basic concepts and considerations in determining sample size. Although sample size determination is important in all research, the focus of this paper is on the experimental study and, in particular, the randomized trial. Throughout the paper, the concepts underpinning sample size estimations are reinforced with an example of a hypothetical trial of midwifery practice.

The randomized trial

The randomized trial has been heralded as the ‘gold standard’ (Robson 1993) and the only design with the ‘potential directly to affect patient care’ (Altman 1996, p. 570). While the limitations of randomized trials of socially complex interventions are acknowledged (Lindsay 2004), they have been used in many areas of midwifery and nursing, e.g. to assess the relative effects of different modes of managing the third stage of labour (Begley 1990) and the effects of relaxation and music on postsurgical pain (Good et al. 1999). In such a study, participants are allocated randomly to the interventions under investigation – often, but not always, a control and an experimental group.

For the purpose of this discussion we will use as an example a comparison of the effectiveness of a new type of tissue adhesive for perineal skin closure vs. the traditional three-stage perineal repair (using a polyglycolic acid suture and sub-cuticular skin closure), with the outcome of most interest being short-term (24–48 hours) postpartum perineal pain. We will assume that pain will be measured at fixed time periods and that the instrument used is a reliable way to measure pain. Women meeting the inclusion criteria and consenting to take part in the trial would be randomly assigned to the ‘tissue adhesive’ group or the ‘suture’ group. This random assignment ensures that each woman has an equal chance of having either the tissue adhesive or suturing repair but, importantly, the group to which an individual woman will be allocated cannot be predicted in advance (Altman & Bland 1999). The randomization attempts to ensure that both known and unknown factors, or ‘confounding variables’, that might affect the outcome (in this case, pain), are distributed evenly between the two groups so that any differences that are subsequently found will be due to the true effects of the intervention and not the formation of the two groups. The possibility that chance itself, through the allocation process, will generate randomized groups with different propensities to the outcome of interest is why trials have to be large enough to ensure that true differences between the interventions are not overwhelmed by these chance effects.

For example, it might be the case that a previous third degree tear or previous dyspareunia after perineal repair could affect a woman’s experience of subsequent perineal pain. Where the sample size is large enough, it is likely, although it cannot be guaranteed, that randomization will ensure that women with such experiences will be allocated evenly to both groups. The larger the number of women randomized, the more likely it is that this balancing will occur. In addition, researchers might use techniques such as stratification or minimization to make it more likely that participants with known prognostic factors are evenly distributed between the two groups. In stratification, separate randomization lists are used to select participants from each prognostic subgroup (Roberts & Torgerson 1998) or ‘stratum’. Alternatively, minimization attempts to ensure balance between groups for specified prognostic factors. At the start of the study, the first participant is randomly allocated to either group. Subsequent participants are assigned to whichever group would minimize the imbalance between the groups on prespecified prognostic factors (Roberts & Torgerson 1998). However, both these techniques require that the confounding variables are known and can be measured and recorded before randomization. Variables that cannot be measured or whose importance is not known can only be dealt with through the randomization of a sufficiently large number of participants. When the study has been completed and before analysing the results, the distribution of the confounding variables between the groups should be examined to ascertain whether any serious imbalances in known variables were generated by chance alone.

The null hypothesis

In planning a randomized trial, the first steps are to set out its purpose and, traditionally, to establish one or more
hypotheses to be tested. A hypothesis is a prediction of the relationship between two or more variables, such as the type of suture and pain (Burns & Grove 1997, Polit & Hungler 1999). The logic of statistical inference requires that hypotheses are framed in terms of there being no relationship between the variables. Such framing is termed the null hypothesis ($H_0$) and this is tested even if we expect a relationship to exist. If, for example, our presumption is that women allocated to the tissue adhesive group experience less pain than those allocated to the suturing group, this brings into question how much less pain. If we attempted to assess exactly how much less pain would be experienced by women in the trial, the list of possible reductions would become quite exhaustive and it would be almost impossible to test for all differences. For example, would 1%, 2%, 5%, 10%, etc. fewer women experience pain above a specific level on a pain scale? The null hypothesis overcomes this difficulty by starting out with an assumption that there is no relationship between the variables, i.e. that there is no difference in pain experienced by women in the tissue adhesive and suture groups. The null hypothesis is therefore assumed to be true unless we find evidence to the contrary. If the null hypothesis is rejected, we do not specifically accept any alternative hypothesis; rather we conclude that there is a difference in pain.

The number of women required in our hypothetical study depends on several factors. These are: (1) the anticipated differences between the tissue adhesive and suture groups (i.e. the ‘effect size’), (2) the level of statistical significance we consider appropriate (‘P-value’, or ‘alpha level’) and (3) the chance of detecting the difference we anticipate (or the ‘power’ of the test) (Machin et al. 1997). It is important to note that because some of the parameters required for sample size and power ‘calculations’ are based on assumptions, such calculations provide estimates rather than definitive numbers needed for a study (Friedman et al. 1998).

**Effect size**

To determine the appropriate sample size for our study, we must prespecify the magnitude of the difference between the two groups that we would regard as clinically meaningful and important (Friedman et al. 1998). This is known as the ‘effect size’, which is a measure of how ‘wrong’ the null hypothesis is. If we already knew the true effect size, we would not, of course, be planning to do the trial. In determining the predicted effect size, researchers might use data from a pilot study. Effect sizes can also be elicited by determining what would be a clinically important effect (SPSS 2000) by consulting with experts in the clinical outcome under investigation, discussions with patients, and/or through the use of survey data. The appropriate effect size will vary widely between studies, since it should represent the smallest effect that would be regarded as clinically meaningful and important, and is therefore dependent on factors such as the convenience of the intervention, the severity of the condition and the outcome being measured. For example, an intervention that reduces maternal mortality by as little as 1% would certainly be regarded as clinically important, while an intervention that reduces the incidence of neonatal physiological jaundice by 10% might be regarded as less clinically important.

In practice, determining effect size is often arbitrary. If power and significance level are held constant (see below), the larger the effect size the smaller the sample required, and the smaller the effect size the larger the sample required (see Table 1). It is not unusual for clinicians and researchers to base their sample size estimation on a given effect size, only to revise the effect size with an aim of detecting a rather larger difference than had originally been intended when they realize that they are unlikely to recruit a sufficient number of participants.

One such example is the randomized controlled trial of cardiotocography vs. Doppler auscultation of the foetal heart at admission in labour in a low risk obstetric population by Mires et al. (2001). The researchers report that with a significance level of 5% and a power of 80%, they required a sample size of more than 2500 women to detect a 3% difference in metabolic acidosis between the two groups. However, due to recruitment difficulties, they subsequently revised their estimate for the effect size by increasing it to 4%, which with the same power and significance level resulted in a required sample size of about 1700 women. This illustrates how a fairly small alteration in the effect size estimate can have a large impact on the estimated sample size and, perhaps, on the feasibility of the trial. However, such changes to the sample size calculation do mean that the trial

<table>
<thead>
<tr>
<th>Effect size (%)</th>
<th>Power (%)</th>
<th>Alpha (%)</th>
<th>Sample size (per group)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>80</td>
<td>5</td>
<td>384</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>5</td>
<td>1536</td>
</tr>
<tr>
<td>10</td>
<td>90</td>
<td>5</td>
<td>513</td>
</tr>
<tr>
<td>10</td>
<td>95</td>
<td>5</td>
<td>635</td>
</tr>
<tr>
<td>10</td>
<td>80</td>
<td>1</td>
<td>572</td>
</tr>
<tr>
<td>10</td>
<td>90</td>
<td>1</td>
<td>727</td>
</tr>
<tr>
<td>10</td>
<td>95</td>
<td>1</td>
<td>870</td>
</tr>
<tr>
<td>5</td>
<td>95</td>
<td>1</td>
<td>3531</td>
</tr>
</tbody>
</table>
might no longer be sufficiently powerful to detect the effect size that was deemed clinically important when the trial was initially designed and are therefore, inevitably, controversial (Murray 2001, Nesheim 2001).

**Probability (P value)**

When experimental research is analysed using inferential statistics, which infer conclusions about a population from sample data, suitable tests are applied and the result is checked on the appropriate probability table to determine the level of statistical significance. Even when our null hypothesis is true, that is, when there is absolutely no difference between the groups, slight differences may well be observed by chance alone. Therefore, the statistical tests applied by researchers seek to determine the likelihood that any differences observed between the groups are due to chance rather than being true differences generated by the interventions. The probability of obtaining the observed difference between the two groups if our null hypothesis is true is termed the ‘P value’.

In our hypothetical study, we would assess whether any difference in pain experienced by women in the tissue adhesive versus the suture group might be due to chance, when the null hypothesis is true (Altman 1991). If the P value is very small, the likelihood that the observed difference is due to chance alone would be small; we would reject the null hypothesis and conclude that the two methods of perineal repair do differ in terms of postpartum pain experienced by the women.

**Significance level and type I error**

In estimating our required sample size, the significance level, denoted as α (alpha), is the critical value we chose for the probability that the null hypothesis is true. If the P value for our trial is less than α we would reject the null hypothesis and, conversely, if P > α we would not reject it (Machin et al. 1997). If, for example, we set alpha at the conventional level of 0.05 (i.e. 5%), we are stating that we are (holding effect size and power constant) willing to accept a 5% probability of falsely rejecting the null hypothesis that there is no difference between the ‘tissue adhesive’ and ‘suture’ group in terms of pain. However, this is little more extreme than the chances of rolling a total 11 with two dice, the odds of which are 1 in 18. Given the large amount of health care research, using such a threshold is likely to lead to many ‘false positive’ conclusions of a difference between treatments when the true difference is much smaller or, even, the opposite of that found by the research. The use of a more cautious P value of 0.01 will increase the required sample sizes but will also reduce the likelihood that we will falsely conclude that there really is a difference between two interventions, since the probability of a result that favours one treatment being due to chance would be less than 1 in 100.

To falsely reject a true null hypothesis is to make a type I error, where we have falsely concluded that there is a difference between our groups when no such difference really exists; that is, a false positive. Where α = 0.05, we are accepting a 5% possibility of the differences between the two groups having occurred as a result of chance rather than as a result of the study intervention. As noted above, it might be preferable to adopt a more cautious approach to rejecting the null hypothesis and we can do this by selecting an alpha of 0.01.

**Power and type II error**

The ability of a test to identify correctly that there is a difference between the groups in a trial is called the power of the test (Pallant 2001); that is, the ability of a test to reject the null hypothesis when it should be rejected (Anthony 1999). Statistical tests vary in their power to detect differences when they actually exist (e.g. where certain assumptions are met, such as normality of distribution, parametric tests have more power than non-parametric tests). The other determinants of power are the sample size, effect size and level of significance or alpha level.

Our study could yield a difference between the tissue adhesive and suture groups that would lead to a probability (P) greater than our chosen alpha level (α) (i.e. P > α), which would lead us to accept the null hypothesis and conclude that there is no statistically significant difference in terms of pain between the two groups. However, it is possible for P to be greater than α when the null hypothesis is in fact false, i.e. we conclude that there is no significant difference between the groups when there really is. In this situation, we fail to reject the null hypothesis when it is in fact false. This is termed a type II error or a false negative.

A minimum power of 0.80 or 80% is usually chosen in clinical trials (Cohen 1988, Machin et al. 1997), although 0.90 is not uncommon. The former means that the researcher is willing to accept a chance of 1 in 5 that they will conclude that there is no difference between the interventions being assessed even when there is one. This can be contrasted with the higher level that is used for alpha (usually 0.05 or 95%), implying that researchers are more inclined to accept that an experimental intervention will be no better than the existing intervention against which it is compared. Thus, health care researchers set higher standards for accepting that there is a difference between two interventions than for concluding
that their effects are similar. Such cautious behaviour may stem from what Machin et al. (1997) term innate conservatism. Researchers are, he suggests, more willing to accept a traditional intervention rather than risk adopting a newer one. In fact, this conservatism would be heightened if alpha levels of 0.01 were used without also raising the level used for power.

Estimation of sample size

Turning to the calculation of the sample size for our hypothetical trial, other work has shown that the prevalence of pain felt by women at 24–48 hours following a three-stage perineal repair using a polyglycolic acid suture and subcuticular skin closure is 62% (Gordon et al. 1998). With this in mind, we might consider that a 10% difference between the groups in our trial would be the smallest effect that would be clinically important to detect (i.e. a reduction from 62% to 52%).

If we follow convention and set the alpha level at 0.05 and the power at 80% for our trial, we can calculate the sample size to test the null hypothesis that the proportion of women experiencing pain at 24–48 hours is the same in the two treatment groups. Using SamplePower, Version 2.00. (SPSS 2000) with two-tailed tests and based on the above difference in proportions, our study would require a sample size of 384 in each of the two groups.

Table 1 shows how the sample size would vary if we altered our assumptions and expectations. For example, if we were interested in detecting a smaller difference, of say 5%, rather than 10%, the sample size would almost quadruple to 1556 per group. This is because the smaller the effect size we are interested in, the more likely it is to be swamped by the effects of chance when randomizing participants into the study; the only way to overcome this is by randomizing more participants. If we wish to retain ‘use of’ a 10% reduction in pain but want to reduce the chances of a false negative, then we can increase the power to, for example, 90% or 95%, leading to sample size estimates of 513 and 635 per group, respectively. If, instead, we wish to retain a power of 80% but to be more cautious about false positives, we can decrease the alpha to 1% – giving a sample size of 572 per group. Finally, if we want to conduct a trial with a high power to avoid false negatives, a more cautious approach to false positives and a desire to detect reliably a difference of just 5%, we would need to do a trial with 3531 women in each group. Although a single trial might not be able to recruit such a large number of participants, it may be that, through combination with similar trials in a meta-analysis, enough randomized evidence would be brought together for this purpose.

Conclusion

Novice researchers often express concerns about sample size estimation. We hope that this paper has helped to demystify this process by explaining some of the basic concepts and issues necessary to determine the minimum sample size required for a randomized trial. Essential to the planning of a randomized trial is an estimation of the required sample size. Health care researchers should ensure that a power analysis is conducted when the study is being planned, in order to confirm that there is sufficient power to detect, as statistically significant, the smallest effect that would be regarded as clinically important. To achieve this, researchers should give proper consideration to each aspect of the calculation and, importantly, be pragmatic about what they hope their study will achieve.

References


