ELSEVIER

Journal of Clinical Epidemiology 65 (2012) 301−308

Journal of Clinical Epidemiology

The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency

Julius Sim*, Martyn Lewis

Arthritis Research UK Primary Care Centre, Keele University, Keele, Staffordshire ST5 5BG, UK

Accepted 26 July 2011; Published online 9 December 2011

Abstract

Objective: To investigate methods to determine the size of a pilot study to inform a power calculation for a randomized controlled trial (RCT) using an interval/ratio outcome measure.

Study Design: Calculations based on confidence intervals (CIs) for the sample standard deviation (SD).

Results: Based on CIs for the sample SD, methods are demonstrated whereby (1) the observed SD can be adjusted to secure the desired level of statistical power in the main study with a specified level of confidence; (2) the sample for the main study, if calculated using the observed SD, can be adjusted, again to obtain the desired level of statistical power in the main study; (3) the power of the main study can be calculated for the situation in which the SD in the pilot study proves to be an underestimate of the true SD; and (4) an ''efficient'' pilot size can be determined to minimize the combined size of the pilot and main RCT.

Conclusion: Trialists should calculate the appropriate size of a pilot study, just as they should the size of the main RCT, taking into account the twin needs to demonstrate efficiency in terms of recruitment and to produce precise estimates of treatment effect.

Keywords: Pilot study; Randomized controlled trial; Sample size; Statistical power; Standard deviation; Confidence interval

1. Introduction

A sample size calculation for a randomized controlled trial (RCT) is undertaken to estimate the minimum number of participants required to detect as significant a prespecified effect, with a stated level of statistical power and at a chosen significance level [1]. Here, the significance level equates to the risk of incorrectly rejecting the null hypothesis (a type 1 error), and power is the probability of detecting as statistically significant an effect of a specified magnitude, if it exists; this is equivalent to the probability of avoiding a type 2 error.

Where the outcome variable of interest is an interval/ratio scale and the effect in question is a mean difference, the sample size calculation depends in part on the value of the standard deviation (SD) of the outcome variable in the main RCT. This is unknown and often estimated by the SD from a pilot study; this process is equivalent to the estimation of a population parameter. However, given that the pilot SD is a random variable, it may be an under- or overestimate of the SD in the main RCT. Accordingly, an RCT may turn out to be under- or overpowered to detect the specified effect, owing to the SD of data in the trial—on which the hypothesis test is based—being respectively larger or smaller than the value used in the prior sample size calculation. Under- or overestimation of the SD in the main RCT may be for two reasons. First, the estimate used in the calculation may not be appropriate for the clinical population in which the trial is conducted (e.g., it was derived from a previous study of patients whose age, chronicity, or symptom severity differed from that of the patients in the RCT). That is, the SD used in the sample size calculation may be biased (systematic error). Alternatively, as a random variable, the SD used in the calculation may have under- or overestimated the SD in the main RCT simply through sampling fluctuation (random error).

A pilot study can help to remedy the problem of bias in the estimate of the SD as it can be conducted on the same clinical population as will be included in the subsequent RCT. However, the estimate of the SD from a pilot study may still under- or overestimate the SD in the main RCT through random error. The more pressing concern is the possibility of underestimation of the SD, with the consequence of underestimation of the sample size for the main RCT. This appears to be a common phenomenon [2] and prevents

* Corresponding author. Arthritis Research UK Primary Care Centre, Keele University, Keele, Staffordshire ST5 5BG, UK. Tel.: +44-1782-734253; fax: +44-1782-734255.

E-mail address: j.sim@keele.ac.uk (J. Sim).

0895-4356/$ - see front matter
doi: 10.1016/j.jclinepi.2011.07.011

What is new?

- Small pilot studies may provide imprecise estimates of the standard deviation (SD), and resulting power calculations for the main study may be correspondingly imprecise, but guidance on the appropriate size of a pilot study is sparse.
- Basing the power calculation on a value at the upper confidence limit for the SD can provide reassurance that the main study will have the desired level of statistical power.
- An inflation factor can be used to calculate this adjusted value for the SD, in relation to a pilot study of a given size, and can also be used to determine an adjusted sample size.
- The size of a pilot study for a randomized controlled trial using an interval or ratio outcome measure should be determined through a calculation based on the precision of the estimate of the SD.

clear conclusions from being drawn from the individual RCTs concerned. It is, therefore, important that a pilot study provides an acceptably precise estimate of the SD so as to reduce the likelihood that the trial is underpowered to detect the prespecified clinical difference. In essence, an acceptably precise estimate of the SD requires the pilot study itself to be of sufficient size. It has been suggested that $n = 30$ is an acceptable size for a pilot study [3]. With specific reference to estimates of the SD, Julious [4] proposes at least $n = 12$ per group, equivalent to $n = 24$ for a traditional two-group study, a figure similar to that proposed by other authors [5,6]. However, there is otherwise little in the way of specific guidelines on the appropriate size of a pilot study.

The aim of this article is twofold: (1) to guide trialists, at the developmental stage of their research, as to the appropriate size of a pilot study to gain sufficiently precise estimates of the true SD and (2) to help inform trialists, post-pilot, on how to adjust their estimate of the SD, for purposes of the sample size calculation of the main trial, to be confident of not underpowering their main study. Our focus is on the situation in which a prior pilot study is conducted independent of the main RCT, rather than on that in which an internal pilot study is performed; that is, where the required sample size is recalculated on the basis of an estimate of the SD derived from the first patients recruited to the main RCT [5−8].

2. Illustrative example

Suppose an RCT is being designed to detect a mean difference in systolic blood pressure of at least 8 mm Hg between two treatment groups. A pilot study is conducted and provides an estimated pooled SD on this scale of 20 mm Hg. Based on this estimate—and assuming 80% power and a 5% two-tailed significance level—100 participants per group would be needed for the analysis of the main study [9].

However, the SD for the main study could be either larger or smaller than that estimated in the pilot study, causing the power of the statistical test in the main study to differ from that on which the sample size calculation is based. We can construct a confidence interval (CI) around the estimate of the SD from the pilot study to quantify its precision [10]. When focusing on underpowering as the principal concern, we should adjust the size of the main study to ensure that it is large enough to detect the clinical difference stipulated with at least the power originally intended (the nominal power). This requires us to examine the upper limit of a CI for the pilot SD [11], using the following formula:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}} \le \sigma \le \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}}$$

where s^2 is the sample variance, σ is the population SD, $1-\alpha$ is the chosen confidence level, and χ^2 denotes the chi-square distribution. Fig. 1 shows 95% CIs for the SD at various sizes of pilot studies. It can be seen that there is greater potential for underestimation than overestimation of the SD from a small pilot study. This is because the positive skew of the chi-square distribution for small sample sizes gives rise to asymmetry in the CI [12]. The necessary sample size of the trial is accordingly likely to be more markedly underestimated than overestimated in such a case.

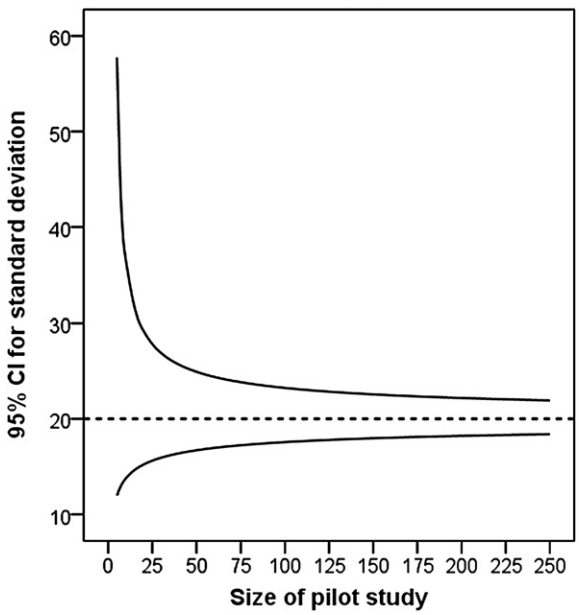

Fig. 1. Ninety-five percent two-sided confidence intervals (CIs) around a standard deviation of 20 for pilot studies of between $n = 5$ and $n = 250$.