A simple formula for the calculation of sample size in pilot studies

Journal of Clinical Epidemiology 68 (2015) 1375–1379

Journal of Clinical Epidemiology

A simple formula for the calculation of sample size in pilot studies

Wolfgang Viechtbauer[a], Luc Smits[b], Daniel Kotz[c,d], Luc Budé[e], Mark Spigt[c,f], Jan Serroyen[g], Rik Crutzen[h,*]

[a]*Department of Psychiatry and Psychology, MHeNS School for Mental Health and Neuroscience, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands*

[b]*Department of Epidemiology, CAPHRI School for Public Health and Primary Care, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands*

[c]*Department of Family Medicine, CAPHRI School for Public Health and Primary Care, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands*

[d]*Institute of General Practice, Medical Faculty, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany*

[e]*Department of Midwifery Education and Studies, Research Centre for Midwifery Science, Zuyd University, PO Box 1256, 6201 BG Maastricht, The Netherlands*

[f]*Department of Community Medicine, General Practice Research Unit, University of Tromsø, Tromsø, Norway*

[g]*Department of Methodology and Statistics, CAPHRI School for Public Health and Primary Care, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands*

[h]*Department of Health Promotion, CAPHRI School for Public Health and Primary Care, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands*

Accepted 29 April 2015; Published online 6 June 2015

Abstract

One of the goals of a pilot study is to identify unforeseen problems, such as ambiguous inclusion or exclusion criteria or misinterpretations of questionnaire items. Although sample size calculation methods for pilot studies have been proposed, none of them are directed at the goal of problem detection. In this article, we present a simple formula to calculate the sample size needed to be able to identify, with a chosen level of confidence, problems that may arise with a given probability. If a problem exists with 5% probability in a potential study participant, the problem will almost certainly be identified (with 95% confidence) in a pilot study including 59 participants.

Keywords: Pilot study; Sample size; Problem detection; Rule of three; Unforeseen problems

1. Introduction

A pilot study can be defined as a small-scale study that helps to examine the practicality and feasibility of the methods to be used in a subsequent larger and more comprehensive investigation [1]. Because conducting an adequately powered study often requires the inclusion of a large number of participants and therefore may be very costly in terms of time and money, piloting a study on a smaller scale can help to identify unforeseen problems that could compromise the quality or flow of the study [2].

For example, one may encounter nonanticipated reasons why potential participants have to be excluded, questionnaire items that are interpreted in unintended ways by the participants or whose answer options are not sufficiently comprehensive, or unclear information about the delivery of the intervention (eg, dosing or visiting schedules).

If such problems are discovered during the course of a pilot study, the necessary steps can be taken before the actual large-scale study is started to minimize or entirely avoid their negative impact. For example, the study protocol or materials (eg, questionnaire items) could be adapted accordingly, or contingency plans could be set up ahead of time to handle any problems adequately and in a timely manner. However, this is only possible if such problems are actually discovered during the conduct of the pilot study. Therefore, a pilot study

W.V. derived the equation. W.V. wrote the first draft and produced the figure. All authors contributed to revising the article and approved the final version. W.V. is the guarantor.

Conflict of interest/Financial disclosure: Neither conflict of interest nor financial support regarding the present study.

* Corresponding author. Tel.: +31 43 388 28 28

E-mail address: rik.crutzen@maastrichtuniversity.nl (R. Crutzen).

http://dx.doi.org/10.1016/j.jclinepi.2015.04.014

What is new?

Key findings

- A simple formula to calculate the sample size needed to be able to identify, with a chosen level of confidence, problems that may arise with a given probability in a pilot study.

What this study adds to what was known?

- Although sample size calculation methods for pilot studies have been proposed, none of them are directed at the goal of problem detection.

What is the implication and what should change now?

- The equation can be easily adopted as a method for sample size calculations in pilot studies that is simple to use, but provides a basis for more reasoned decisions about sample sizes in pilot studies.

aimed at discovering such problems should have sufficient power to do so, or in other words, the sample size of a pilot study must be sufficiently large, such that the probability of detecting such problems is high.

Existing methods for sample size calculations typically focus on how to select an appropriate sample size for a pilot study such that various parameters of interest can be estimated with sufficient precision (eg, the effect size, the standard deviation of the outcome measure, its reliability, or adherence or attrition rates) [1,3–5]. Such calculations may also play an important role in deciding whether to proceed with the primary trial in the first place [6,7]. These considerations have led to various guidelines for choosing an appropriate sample size for a pilot study, such as 12 participants per group [3], values in the range of 10 to 40 participants per group depending on the parameter of interest [4,5], at least 9% of the main trial's sample size [6], or at least 50 participants [8].

However, none of these approaches is directly applicable when the goal of a pilot study was the detection of unforeseen problems. Therefore, in this article, we take a different approach and describe a simple method for determining the sample size necessary to identify problems with a chosen level of confidence in pilot studies.

Not surprisingly, the sample size determined in this manner depends not only on the confidence level with which we would like to detect a particular problem but also the actual probability that the problem manifests itself in a potential study participant. Because the true problem probability is unknown in practice, what we really need to consider is a lower bound for the problem probability: if the true probability is in fact this low (or higher), then we achieve (or exceed) the desired confidence level, but if it is really lower (so that we are more likely than desired to miss the problem), then it would be infrequent enough to not be considered a problem worthy of detection. Choice of this value therefore depends on the context and on how detrimental a problem would be to a trial. We will return to this issue further in the following.

2. Required sample size to detect a problem in a pilot study

For now, assume that a particular problem has a given probability of occurring in a potential study participant. For example, if there is a 0.15 probability of encountering unanticipated reasons for exclusion in a given participant, then there is 0.85 probability that this problem does not manifest itself. In a group of n participants, there is then a 0.85^n probability that the problem will not occur at all. Therefore, the probability of observing at least one occurrence of this problem in n participants is given by

$$P(x>0) = 1 - (1 - \pi)^n,$$

where x denotes the number of participants (of the n participants) in whom the problem manifests itself and π denotes the problem probability. We now want to choose n so that $P(x>0)$ exceeds a certain threshold of confidence, which we denote by $100\% \times \gamma$. In other words, how many participants must be included so that we can be $100\% \times \gamma$ certain that the problem will manifest itself at least once during our pilot study? Solving the equation for n yields:

$$n = \frac{\ln(1-\gamma)}{\ln(1-\pi)}. \quad (1)$$

Because n will typically not be a whole number, we round the value obtained with Equation (1) up to the nearest integer. An online calculator is available at http://www.pilotsamplesize.com.

For example, for $\pi=0.15$, including $n=\ln(1-0.95)/\ln(1-0.15) = 18.43$, or rather 19 participants will ensure that we will encounter at least one incident of the problem at a 95% confidence level. A problem that manifests itself less frequently, for example, with only $\pi=0.05$ probability, will require the inclusion of $\ln(1-0.95)/\ln(1-0.05) = 58.40$ or rather 59 participants in the pilot study so that the problem can be detected with a high confidence level.

Fig. 1 shows the required sample size (ie, n) as a function of the problem probability (ie, π) for three different levels of confidence (ie, γ equal to 0.95, 0.90, and 0.80). Not surprisingly, higher confidence levels and the detection of less-frequent problems require larger sample sizes.

3. Sample size calculations in practice

To use Equation (1) for deciding on a sample size for a pilot study, we must first choose values for γ and π. By

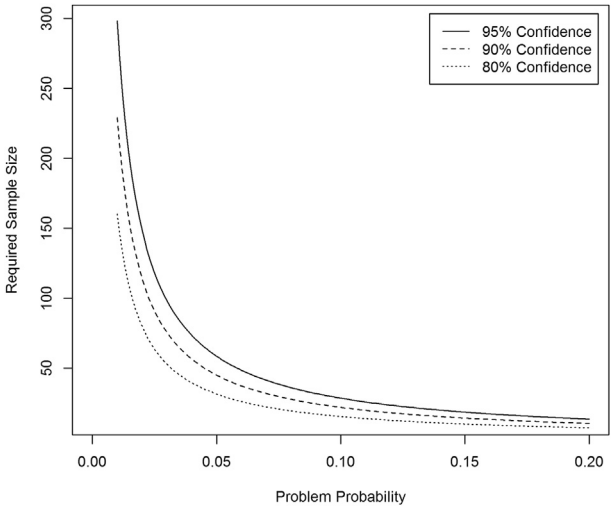

Fig. 1. Required sample size to detect a problem with a given level of confidence.

convention, we could adopt a confidence level of 95% (analogous to the level commonly chosen for confidence intervals) or choose the confidence level in accordance with the severity of the issue that we want to detect. (ie, a potentially disastrous problem that could ruin the entire study should be detected with higher confidence than a problem that would merely be a nuisance to deal with.)

With respect to π, we need to consider that very infrequent problems will naturally require a large sample size to detect. However, infrequent problems (unless they are so disastrous as to require the immediate discontinuation of the study) can be handled on an as-needed basis, without disrupting or jeopardizing the entire study. When choosing π, one should therefore consider the amount of resources (eg, time and personnel) available to handle unforeseen difficulties and their impact on the trial. In addition, because the ''true'' probability of a particular problem is technically unknown in practice, we suggest thinking of π as the ''minimum'' probability of the problem that we would like to detect with the desired confidence level.

For example, suppose the inclusion or exclusion criteria for a study have been written to the best of the researchers' knowledge, there may be participants for whom the criteria do not yield an unambiguous outcome. Whenever such a case presents itself, for example, during a screening visit, a decision about the eligibility of the participant needs to be made. Suppose that a single investigator is responsible for making this decision. A large number of such cases would then quickly overburden this investigator. Accordingly, it is decided that, if such difficulties present themselves with ''at least'' $\pi=0.10$ probability (ie, in at least 1 of 10 participants), it would be good to detect this problem already during the pilot study. Accordingly, Equation (1) then indicates that 29 participants need to be screened to be 95% confident that one or more such cases are in fact encountered.

If the true probability is actually higher than $\pi=0.10$, then the achieved confidence level will exceed 95% when 29 participants are screened (eg, approximately 99% if $\pi=0.15$, as one can easily verify with Equation 1). Therefore, screening 29 participants will ensure a high level of confidence (ie, at least 95%) for the chosen minimum problem probability. On the other hand, if the actual problem probability is lower than $\pi=0.10$, then 29 participants will not be sufficient to reach the desired 95% confidence level. However, in that case, because the true problem probability is actually lower than the minimum level deemed to be important for detection, we can also consider it acceptable that we run a higher risk of ''not'' encountering at least one case of ambiguity (because the investigator can easily handle a smaller number of ''on the spot'' decisions).

Note that the ''minimum problem probability'' is context dependent. For example, suppose the presence or absence of a particular condition constitutes the primary outcome variable in a trial. Diagnoses are made independently by two clinicians, and disagreements would demand additional follow-up assessments. If such assessments are costly and/or invasive, we may already find a disagreement probability of (at least) $\pi=0.01$ to be problematic and therefore would want to make sure that this issue is likely to be discovered during a pilot study (eg, so that ways of reducing the occurrence of disagreements can be considered). On the other hand, if costs and invasiveness are minimal, we may consider a higher probability of $\pi=0.05$ or $\pi=0.10$ to be acceptable (but anything higher might call into question the reliability with which diagnoses are made).

Similarly, if responses to a questionnaire are used to measure the primary outcome in a trial, then we may again choose a lower minimum probability for detecting particular problems (eg, nonresponses, item misinterpretations) than if the same questionnaire was part of a process evaluation, where problems that occur with a low probability are not detrimental to the trial itself. However, if problems manifest themselves with at least a $\pi=0.05$ or $\pi=0.10$ probability, we would still hope to be alerted of their existence—even if it is not the primary outcome in the trial.

Finally, in practice, one will often be interested in not just a single but an entire collection of different problems (eg, difficulties in applying the inclusion or exclusion criteria, ambiguities in questionnaire items, problems in the application of the treatment). Equation (1) can then be used to determine the required sample size to detect each individual problem. The largest value of n obtained this way then guarantees that each individual problem can be detected with ''at least'' the desired confidence level specified.

For example, suppose that, in addition to potential ambiguities with the inclusion or exclusion criteria, the researchers also want to detect any problems participants may have with the interpretation of the items on a questionnaire. Because incorrectly interpreted items may

Box The rule of three

A method for constructing confidence intervals for π when observing zero events (ie, when $x=0$) in a study [13,14], sometimes called the "rule of three," is closely related to Equation (1). Applied to the present context, the method works as follows: If the problem does not manifest itself in the n participants included in the pilot study, then an approximate 95% confidence interval for π is given by the interval with endpoints 0 and $3/n$. For example, suppose with 60 participants, we never observe the problem in the pilot study. Then, the endpoints of an approximate 95% confidence interval for π are equal to 0 and 0.05. Note that Equation (1) for $\pi=0.05$ leads in fact to $n=59$, the difference resulting only from the approximation used in the derivation of the confidence interval [13]. Therefore, Equation (1) essentially implies that one should include $3/\pi$ participants if one wants to be approximately 95% certain that the problem will manifest itself at least once during the pilot study.

completely invalidate the responses from the participants that encounter such difficulties, the investigators want to be 95% certain that they will detect this issue during the pilot study if the problem manifests itself with "at least" $\pi=0.05$ probability (ie, in at least 1 of 20 participants). As described earlier, 59 participants should then be included in the pilot study. This will also ensure that any ambiguities with the inclusion or exclusion criteria are detected with more than a sufficient confidence level.

4. Discussion

Despite the extensive levels of planning involved in the design of a study, experience shows that unforeseen problems can and typically will arise during the conduct of a study that must be handled with care. A pilot study provides an excellent opportunity to uncover such problems ahead of time, minimizing the need to adapt procedures or to develop contingency plans on short notice when the larger study is being conducted. In this article, we described a simple method for choosing a sample size for a pilot study that ensures the discovery of potential problems with high confidence.

How have existing pilot studies fared in the context of these results? Arain et al [9] found a median sample size of 76 participants in pilot studies published in 2007 or 2008 in four general medical and three specialist journals. For this median sample size, problems with a probability of approximately 0.04 could have been detected with a 95% confidence level, whereas problems that occurred with a probability of 0.02 could still be detected with almost 80% confidence.

It needs to be emphasized that the method does not indicate the appropriate sample size for estimating the actual probability of a particular problem (ie, π) with a given level of precision [1,5]. Instead, the method is used to determine the necessary sample size so that the problem is likely to be observed at least once during the course of the pilot study. The emphasis therefore is on "problem detection" and not on the estimation of the "problem frequency."

Similarly, it is often suggested that pilot studies are useful for determining an appropriate sample size for large-scale investigations, for example, by providing information about the variability in the outcome variable of interest or regarding the size of the effect. However, because pilot studies are by definition conducted with small sample sizes, they usually do not provide accurate information about such parameters and therefore may lead to erroneous decisions when such estimates are used without further qualification [2,10]. The same issue applies when using pilot studies to estimate other parameters of interest when planning studies (eg, adherence or attrition rates).

Finally, it is worth noting that Equation (1) has been discussed in other contexts, such as the number of animals required to detect at least a single case of disease with a given confidence level during long-term holding of laboratory rodents [11] and the required number of objects (typically precincts) that must be audited to detect at least one case of election fraud [12]. In addition, Equation (1) is related to what is sometimes called "the rule of three," a method for computing confidence intervals for proportions (see Box). As described in the present article, the equation can be easily adopted as a method for sample size calculations in pilot studies, which is simple to use but provides a basis for more reasoned decisions about sample sizes in pilot studies.

References

[1] Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, et al. A tutorial on pilot studies: the what, why and how. BMC Med Res Methodol 2010;10:1.
[2] Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA. Caution regarding the use of pilot studies to guide power calculations for study proposals. Arch Gen Psychiatry 2006;63:484–9.
[3] Julious SA. Sample size of 12 per group rule of thumb for a pilot study. Pharm Stat 2005;4:287–91.
[4] Hertzog MA. Considerations in determining sample size for pilot studies. Res Nurs Health 2008;31:180–91.
[5] Johanson GA, Brooks GP. Initial scale development: sample size for pilot studies. Educ Psychol Meas 2010;70:394–400.
[6] Cocks K, Torgerson DJ. Sample size calculations for pilot randomized trials: a confidence interval approach. J Clin Epidemiol 2013;66:197–201.
[7] Stallard N. Optimal sample sizes for phase II clinical trials and pilot studies. Stat Med 2012;31:1031–42.
[8] Sim J, Lewis M. The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency. J Clin Epidemiol 2012;65:301–8.

[9] Arain M, Campbell MJ, Cooper CL, Lancaster GA. What is a pilot or feasibility study? A review of current practice and editorial policy. BMC Med Res Methodol 2010;10:67.

[10] Browne RH. On the use of a pilot sample for sample size determination. Stat Med 1995;14:1933−40.

[11] Institute of Laboratory Animal Resources. Long-term holding of laboratory rodents. ILAR News 1976;19:1−25.

[12] Aslam JA, Popa RA, Rivest RL. On estimating the size and confidence of a statistical audit. Proceedings of the 2007 USENIX Workshop on Accurate Electronic Voting Technology; 2007, Boston.

[13] Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. J Am Med Assoc 1983;249:1743−5.

[14] Eypasch E, Lefering R, Kum CK, Troidl H. Probability of adverse events that have not yet occurred: a statistical reminder. BMJ 1995;311:619−20.