Reviewer's report

Title: Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study.

Version: 1 Date: 28 January 2014

Reviewer: Simon Gates

Reviewer's report:

A useful piece of work that should help people to design pilot/feasibility studies.

Major compulsory revisions

1. My main criticism is something that isn’t in the paper, which I think ought to be at least mentioned. The whole article is based around the standard conventional method of calculating a sample size based on achieving a “statistically significant” result. Although this is the predominant method, it is by no means the only way of deciding on a sample size for a trial, and, if you don’t believe that a p-value is a good way of deciding whether a treatment works (and there are lots of reasons to think that), certainly not the best way (even though it tends to end up being the method used). There has for some time been a drive to consider magnitude of effects and uncertainty around them rather than significance, but determination of sample sizes has lagged behind and is still largely based on significance testing. I think it would be very helpful to state that the work reported here is primarily based on the traditional significance test based sample size methods (though some aspects of it are also relevant to other methods of course – knowing incidence of outcome or standard deviations will always be important). This would help to raise awareness that basing a sample size on a significance test isn’t the only way, or even necessarily a good way, to decide how big a trial needs to be. Most people still have a very naïve view of a sample size estimate as a kind of magic “number needed” and anything that helps to challenge this thinking would be a good thing. I think it would be outside the scope of this paper to get into the issue of whether it makes sense to base a sample size on a significance test, though the authors may wish to comment.

2. A more substantive point is that isn’t actually necessary to randomise to estimate most of the parameters. This is mentioned early on but tends to get a bit lost later e.g. in figure 5 it is assumed that the pilot/feasibility study will be randomised, which may well not be the case. It will often be possible to estimate important parameters (for example, the mean and sd, or incidence, of an outcome) by applying the eligibility criteria to a series of patients, either prospectively or retrospectively. This would usually be a more efficient way of doing it than running a randomised pilot. In my experience the main reason for doing a randomised pilot/feasibility phase is to test the feasibility of recruitment than estimation of sample size parameters. This may limit the applicability of the results presented here.
Minor essential revisions

1. Labelling of the figures needs to be sorted out – at present the labels on the figures don’t match up with the captions, which makes it a bit difficult to follow.
2. Background, end of second paragraph: “does not” should be “do not”.
3. Page 5, top: “Feasibility studies for randomised controlled trials may not themselves be randomised”; the “not” should come after “themselves”.
4. Page 5, bottom; presumably a typo: “so our aim should like…”
5. Page 6, near the bottom: “we can then calculate the required sample size for the definitive RCT.” See point 1 above. This is an instance of the tacit assumption that there is only one assumes that the “definitive” sample size would that that required to achieve a statistically significant result with specified power. As noted above, this the view that has become standard, to the extent that most people accept it unquestioningly, but if you take the view that a “significant” result is not a good way to decide on a treatment’s effectiveness (and there are many reasons to think that) then basing a sample size on it doesn’t make sense. I think it would be helpful to acknowledge that the sample size calculations mentioned are based on significance tests and that there may be other, better ways of determining a sample size.
6. Page 9: “For each simulated pilot study of size(nj:10(5)200)” needs to be explained.
7. It would be helpful to number the equations.
8. Page 10: there is some text that looks as though it should have been deleted: “Finally, the relative percentage gain in precision around the true binomial proportion per increase in 5 study participants for a fixed true success probability is defined as”.
9. Page 11, top line: “a strong bias”. This seems wrong – the bias looks very small to me. Am I looking at the correct figure?
10. Page 11, third paragraph: some of this paragraph should probably be in the Methods rather than the Results.
11. Figure 4 and 5 captions. I couldn’t see a bold vertical line on the figures.

Discretionary revisions

1. Background, first paragraph: “widely regarded as the most powerful research design”. It is probably fair to say that the RCT is the least biased research design.
2. Page 4, 3rd paragraph “Several UK publicly funded grant awarding bodies” is an awkward way of putting it. “UK public funding bodies“ would be better.
3. The discussion of definitions of feasibility and pilot studies (pages 4-5) isn’t really necessary here. I think it would be sufficient just to define what is meant by the terms used in this paper. The text is slightly confusing anyway; “So pilot studies are pieces of research done before a main study in order to answer the question “Can this study be done?”.” Surely that would be a feasibility study?
4. Slightly pedantic point; in the list of parameters that a pilot might estimate, “number of eligible patients” should really be number per unit time (i.e. rate), or number/centre/time.

**Level of interest:** An article of importance in its field

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests.