Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2016, Vol. 25(3) 1039–1056

DOI: 10.1177/0962280215588242
smm.sagepub.com
SAGE

How big should the pilot study for my cluster randomised trial be?

Sandra M Eldridge,[1] Ceire E Costelloe,[1] Brennan C Kahan,[1] Gillian A Lancaster[2] and Sally M Kerry[1]

Abstract

There is currently a lot of interest in pilot studies conducted in preparation for randomised controlled trials. This paper focuses on sample size requirements for external pilot studies for cluster randomised trials. We consider how large an external pilot study needs to be to assess key parameters for input to the main trial sample size calculation when the primary outcome is continuous, and to estimate rates, for example recruitment rates, with reasonable precision. We used simulation to provide the distribution of the expected number of clusters for the main trial under different assumptions about the natural cluster size, intra-cluster correlation, eventual cluster size in the main trial, and various decisions made at the piloting stage. We chose intra-cluster correlation values and pilot study size to reflect those commonly reported in the literature. Our results show that estimates of sample size required for the main trial are likely to be biased downwards and very imprecise unless the pilot study includes large numbers of clusters and individual participants. We conclude that pilot studies will usually be too small to estimate parameters required for estimating a sample size for a main cluster randomised trial (e.g. the intra-cluster correlation coefficient) with sufficient precision and too small to provide reliable estimates of rates for process measures such as recruitment or follow-up rates.

Keywords

pilot, cluster randomised trial, sample size, intra-cluster correlation

1 Introduction

Cluster randomised trials are often larger and more expensive than individually randomised trials. The interventions they evaluate can be complex, trial logistics complicated, and execution difficult.[1,2] Potentially, some of the reported problems in such trials could have been avoided if procedures and activities had been tested beforehand. Therefore, pilot studies, carried out in advance of a main study to answer the question 'Will it be possible to conduct the main study and what, if anything,

[1]Centre for Primary Care and Public Health, Queen Mary University of London, London, UK
[2]Postgraduate Statistics Centre, Department of Mathematics and Statistics, University of Lancaster, Lancaster, UK

Corresponding author:
Sandra M Eldridge, Centre for Primary Care and Public Health, Queen Mary University of London, 58 Turner Street, E1 2AB London.
Email: s.eldridge@qmul.ac.uk

should be done differently?' may be even more important for these trials than for individually randomised trials.[3] In recent years, there has been considerable interest in pilot studies conducted in preparation for randomised controlled trials. This has included investigations to assess how large such studies should be.[4] However, none of these investigations have focussed on cluster randomised trials. This is the focus of the present paper.

Any research study, whatever the design, should be large enough to fulfil its objectives; not to do so would be unethical. Thus, careful articulation of objectives is important for determining the required sample size. This argument applies equally to pilot studies for cluster randomised trials. From the definition earlier, the overarching objective of a pilot study is to assess the feasibility of a main trial, an objective often stated in reports of pilot studies for cluster randomised trials.[5–8] However, investigators are usually interested in addressing the feasibility of some specific aspects of a trial. Broadly these aspects are: (i) sample size of the main trial; (ii) recruitment and randomisation; (iii) implementation of interventions; (iv) data collection and outcomes; (v) harms, benefits, and *potential* effectiveness. Objectives related to these five aspects are usually less well articulated in trial reports than an overarching feasibility objective, even though the importance of stating specific objectives has been stressed in several papers focusing on the methodology of pilot studies.[9,10] In addition, many studies labelled as pilot or feasibility randomised controlled trials continue to focus on assessing efficacy or effectiveness, rather than feasibility objectives.[11]

In fact, objectives related to the different aspects of feasibility lead to different methods of determining how large such a study should be. For example, to support the calculation of a sample size for the main trial, a pilot study should ideally be large enough to enable sufficiently precise estimation of key parameters to input into the calculation. Assessment of the feasibility of recruitment and randomisation may require a sample large enough to estimate recruitment rates to a sufficient degree of accuracy. To assess the implementation of the intervention, the intervention needs to be delivered to a sufficient number and spread of participants: for cluster randomised trials this may necessitate a certain number of clusters as well as a certain number of individuals. The appropriate spread of clusters will depend on the trial being planned. For example, if the intervention involves introducing changes to appointment systems within health care organisations it may be appropriate to include both large and small organisations and those representing all types of appointment system currently in place. To assess the feasibility of data collection and the use of certain outcomes, data need to be collected on a sufficient number of individuals to enable assessment of response rates. Finally, to assess harms and benefits, investigators may also need to consider what sort of precision they require in estimates of rates of harm or effect sizes. Note that a formal sample size calculation relating to the effectiveness of an intervention is not included amongst these methods for determining sample size because, using the definition of a pilot study given above, definitive estimation of effectiveness is not a valid objective for a pilot study. However, estimation of *potential* effectiveness via interim or surrogate endpoints may be a valid objective in some pilot randomised controlled trials.

While these specific objectives and their consequences for sample size are broadly similar for pilot studies for cluster and individually randomised trials, pilot studies for cluster randomised trials invariably involve recruiting clusters and investigators need to consider how many clusters to recruit as well as how many individual participants. In a pilot study, both aspects of sample size may differ from the main trial.

In this paper, we focus on the calculation of sample size for pilot studies for cluster randomised trials where the primary objective of the pilot is to determine parameters for a sample size calculation for the main trial. The most common way of estimating sample size for cluster

randomised trials is to calculate the sample size needed for an individually randomised trial testing the same intervention and then multiply by this by an estimate of the design effect

$$1 + (m - 1)ICC$$

where m is the cluster size assuming cluster sizes are fixed and ICC is the intra-cluster correlation coefficient.

The ICC measures the extent of variation between clusters and for continuous outcomes, which are the focus of this paper, can be calculated as the ratio of the between-cluster variance to the total variance (within plus between). Thus, for the sample size calculation for a cluster randomised trial, estimates of both within- and between-cluster variances are required. This is more complicated than in the case of individually randomised trials. Furthermore, an additional complication arises because estimates of these variances depend on the number of clusters as well as the number of individual participants, and the rather small number of clusters typically involved in a pilot study can potentially cause downward bias in estimates of between-cluster variation.[12] Following a brief overview of key parameters of published pilot studies for cluster randomised trials, we use simulation to explore the effect of various decisions about pilot sample size, in terms of both number of clusters and cluster size, on the likely accuracy of an estimate of sample size required for a future main trial and the usefulness of pilot studies for estimating such sample sizes. We also briefly explore the effect of pilot sample size on the likely accuracy of different rates (e.g. recruitment, response, harm) and the implications for achieving other objectives such as the assessment of feasibility of intervention implementation.

2 Overview of pilot studies for cluster randomised trials

To ensure the applicability of our results, we first reviewed a selection of reports of pilot studies for cluster randomised trials. These were sourced from personal collections and a systematic review of pilot trials for cluster randomised trials being conducted for a separate project. The electronic search strategy for that review was based on a previously developed and validated search for cluster randomised trials,[13] with the addition of terms specifying that the title of the trial report should include the words 'pilot' or 'feasibility'. The search will, therefore, have excluded non-randomised pilot and feasibility studies, although our experience is that there are far fewer published non-randomised than randomised pilot and feasibility studies especially for cluster randomised trials. Following screening of studies for inclusion in that review, we selected the eight most recent pilot trials for our brief review and additionally included three recent randomised pilot studies from our personal collections. We used these trials to provide appropriate examples and a broad idea of the size of existing pilot studies for cluster randomised trials.

The pilot study examples[5–8,14–20] are representative of a range of settings and vary in the included numbers of clusters. All involve randomisation of clusters into two or three groups, seven with equal allocation and three in which there was one extra cluster either in the intervention or control group. There was a diverse range of settings and clusters and although most took place in healthcare settings, three were in residential or community settings.

Four of the examples were conducted in hospital in-patients facilities and include as clusters: designated elderly care hospital bays in wards in England (one designated bay from each of eight hospitals randomised into two groups of four)[5]; adult non-maternity in-patient wards from two NHS Trusts in London (18 hospital wards randomised into three groups of six)[15]; UK general hospitals admitting emergency Acute Upper Gastrointestinal Bleeding cases (six academic centres

randomised into two groups of three)[16] and in-patient rehabilitation centres in New Zealand (four centres randomised into two groups of two).[20]

Three studies were conducted in community-based health care facilities or pharmacies; government primary health care and dispensary centres in Tanzania (36 dispensaries randomised into two groups of 18)[8]; primary care Family Medicine Groups (FMGs) from two selected health regions in Canada (four FMGs randomised to two groups of two, with one additional late entry added to control group)[18]; and Scottish community pharmacies (12 pharmacies randomised into two groups of six).[19] The final example in a medical setting involved medical conference workshops in Canada (three workshops with one randomised to traditional case-based teaching and two to simulation-based computerised teaching).[7]

The final three studies were shared apartment communities in Germany (eight sheltered housing units randomised into two groups of four)[14]; nursing homes in the USA (four community nursing homes matched to four Veteran Affairs facilities randomised within pair to receive intervention or control)[17]; and HIV community support groups for African-American women in the USA (29 support groups randomised into 15 intervention and 14 control groups).[6]

A summary of the aims, interventions, number of participants, and clusters for these trials in our brief review is shown in Table 1. It is notable that although most of these studies have an overall aim of assessing feasibility, many focused on the assessment of patient-centred outcomes in their reports, often including calculating effect sizes.

The numbers of clusters randomised in the pilot studies, as specified earlier, ranged from 3 to 36 and the mean cluster size from less than 2 to over 200. Most of the studies recruited more than 50 participants and some involved considerably more. The parameters used for our simulation study broadly reflect the range of cluster sizes and numbers of clusters in these example pilot studies. Only two studies reported an ICC which could be used for future sample size calculation. These ICCs were 0.02 for a trial where the clusters were hospital wards (15), and 0.03–0.4 for a range of outcomes in a study with in-patient rehabilitation facilities as clusters.[20]

3 Methods for assessing sample size required for pilot studies

3.1 Pilot sample size requirements to estimate the sample size required for the main trial

We performed a simulation study to explore the effect of various decisions about pilot sample size on likely accuracy of any future estimate of sample size required for a main trial and thereby assess the usefulness of pilot studies for this purpose. In the simulations we present here, we assumed that both within- and between-cluster variance were unknown and to be estimated from pilot data.

We simulated two distinct types of cluster: clusters that are naturally small, for example a family unit, and those that are naturally large, for example hospitals, schools, communities, or general practices. These two types of cluster often exhibit different characteristics. In trials with naturally small clusters, the extent of variation in the outcome between clusters, measured using the *ICC*, tends to be higher than for naturally large clusters.[21–23] In addition, there are usually a large number of potential clusters available, and generally, all eligible cluster members would be invited to take part in both the pilot and main trial. By contrast, for naturally larger clusters, sometimes due to costs or administrative burden only a subset of eligible members may be invited to participate and this may differ between the main trial and the pilot.

Investigators designing pilot cluster randomised studies are faced with a number of choices about the size of their study. Any two of the following parameters will completely determine the sample

Table 1. Key characteristics of recently published pilot cluster randomised trials.

Author, year, setting	Summary of aims and intervention for pilot	Definition of cluster	Number of clusters in analysis	Number of individual participants at follow up in pilot (minimum–maximum cluster size)
Colon-Emeric et al.,[17] USA	To pilot test the CONNECT intervention to improve staff connections and communications before implementation of a FALLS reduction training programme and estimate the effect size for fall rates	Nursing home	8	497 (not specified)
Drahota et al.,[5] UK	To assess the feasibility, potential benefits and harms, and to guide further research on the effect of shock absorbing flooring for the prevention of falls in elderly care wards	Hospital ward bay	8	571 (13–138)
Hawk,[6] USA	To assess feasibility of recruitment and likelihood of programme efficacy for a single-session intervention (the Girlfriends Project) to reduce risky behaviour, access HIV testing, and increase talking with partner about HIV and STI	Community group	29	149 (not specified)
Jairath et al.,[16] UK	To evaluate the feasibility and safety of implementing a restrictive versus a liberal red blood cell transfusion policy in adult patients admitted to hospital with acute upper gastrointestinal bleeding	Hospital	6	849 (not specified)
Kerr et al.,[7] Canada	To evaluate the feasibility of incorporating simulation-based training into an existing continuing medical education conference	Workshop	3	27 (8–10)

(continued)

Table 1. Continued

Author, year, setting	Summary of aims and intervention for pilot	Definition of cluster	Number of clusters in analysis	Number of individual participants at follow up in pilot (minimum–maximum cluster size)
Leblanc et al.,[18] Canada	To assess the feasibility and acceptability of the design, procedures, and interventions of the DECISION + programme, a continuing medical education programme in shared decision making on optimal use of antibiotics for ARIs	Family Medicine Group	5	544 (not specified)
Okwen et al.,[8] Tanzania	To learn whether enhanced skills-based supervision for health workers improves primary eye care knowledge and skills (n = 50 were interviewed at baseline in 36 clusters, 38 had FU at 6m in 26 clusters)	Dispensary	26	38 (not specified)
Porteous et al.,[19] Scotland	To assess the transferability of a goal-focused Australian intervention to improve self-management of intermittent allergic rhinitis to a UK context	Community pharmacy	12	144 (fixed cluster size of 12)
Reeves et al.,[15] England	To test the feasibility of conducting ward level surveys on nursing care, to examine the effectiveness of two interventions (basic feedback survey or Feedback Plus), and to estimate number of wards needed	Hospital ward	18	684 (13–203)
Taylor et al.,[20] New Zealand	To determine the feasibility, cluster design effect, and the variance and minimal clinical important difference in primary outcome in a pilot study of a structured approach to goal setting	In-patient rehabilitation facility	4	41 (6–17)
Teut et al.,[14] Germany	To evaluate feasibility of, and determine suitable outcomes for, an integrative medicine programme for geriatric patients	Sheltered housing	8	58 (3–10)

size required: (1) number of individual participants per cluster, (2) number of clusters, (3) total number participants. Frequently, investigators have made a decision about one of these parameters in advance and are trying to decide on the value of a second which will, in turn, determine the value of the third. We examined the following three situations, using a number of simulation scenarios for each:

(1) Investigators have made a decision to use the cluster size that they intend to use in the main trial, but there is some flexibility in the number of clusters to be used in the pilot and they must decide on this number. This will, in turn, determine the total number of participants in the study. Drahota et al.[5] and Teut et al.[14] are examples of this. In our simulations we applied this situation to studies involving both small and large natural cluster sizes.
(2) Investigators have identified a limited number of clusters available for the pilot study. They can choose to include the same number of participants from each cluster that they would include in the main trial or they can include fewer. The decision about number of participants in each cluster will, in turn, determine the total number of participants. Okwen et al.,[8] Jairath et al.,[16] and Leblanc et al.[18] are examples of this. In our simulations, we only applied this situation to studies involving naturally large clusters, because when clusters are naturally small entire clusters would usually be invited to participate leaving no flexibility in number of participants per cluster.
(3) Investigators have decided on the total number of participants they want to include in their pilot study but there is some flexibility in the number of clusters and, therefore, the number of participants per cluster. We only applied this situation to studies involving naturally large clusters as for situation 2. This may be a common situation when the costs of a pilot depend largely on the number of individuals recruited and finance for the pilot is limited, but this sort of decision making is unlikely to be reported in a pilot study.

For each of the three situations, we simulated continuous outcomes from the following model

$$Y_{ij} = \alpha + \beta x_{ij} + u_j + \varepsilon_{ij}$$

where

Y_{ij} is the outcome for the *ith* patient in the *jth* cluster
β is the treatment effect
x_{ij} is an indicator variable ($x_{ij} = 0$ for control participants and 1 for intervention participants)
u_j is a random cluster effect for *jth* cluster, following a normal distribution with mean 0 and standard deviation (SD) σ_μ
ε_{ij} is a random patient-level error term following normal distribution with mean 0 and SD σ_ε.

Using this model, σ_μ and σ_ε represent the true between- and within-cluster SDs, respectively. Variance estimates conditional on treatment allocation are independent of the size of treatment effect. We set β to 0 for simplicity and σ_ε to 1 for all simulations and specified a range of *ICC* values suitable for the particular simulation. Based on the specified *ICC* for the particular simulation, we calculated σ_μ as

$$\sigma_\mu^2 = \frac{\left(\sigma_\varepsilon^2 * ICC\right)}{1 - ICC}$$

Although the vast majority of cluster randomised trials involve variable sized clusters, for simplicity, we assumed an equal number of participants per cluster, m_p, in all simulated pilot trials and an equal number of participants per cluster, m_m, in all main trials. These assumptions are likely to underestimate the uncertainty in our results. We randomised clusters to intervention or control group, using block randomisation with a block size of two. For each of the three situations described earlier, we specified a number of simulation scenarios based on a range of true ICCs, a range of values for the fixed parameter (number of clusters, cluster size, total number of participants) for the scenario, and a range of values for one of the variable parameters. We undertook 110 simulations: 48 simulations for situation 1, 30 for situation 2, and 32 for situation 3. Parameters used in the simulations are shown in Table 2.

All simulations were performed in Stata v13.1, using 1000 replications for each simulation. For each replication, we used ANOVA to estimate the between- and within-cluster SDs ($\widehat{\sigma_\mu}$ and $\widehat{\sigma_\varepsilon}$) and the ICC ($ICC_{estimate}$). We used $\widehat{\sigma_\mu}$ and $\widehat{\sigma_\varepsilon}$ to estimate N, the sample size required for an individually randomised trial, based on 90% power, a 5% significance level, and a between-group difference of 0.3; 0.3 is commonly used as a standardised mean difference to calculate sample size for these trials. The true number required, based on a within-cluster SD of 1, is 468 participants in total. We then used $ICC_{estimate}$ to calculate the required number of clusters in the main trial, assuming a fixed cluster size of m_m, using the following formula

$$estimated\,number\,of\,clusters = \frac{N(1 + ICC_{estimate}(m_m - 1))}{m_m}$$

3.2 Sample size to fulfil other objectives of a pilot study for cluster randomised trials

As discussed in Section 1, pilot studies may have objectives that do not relate to sample size estimation for the main trial. For example, in the protocol for the 'Help for Hay Fever' pilot study investigators pose the following questions related to informing a future definitive trial: 'What are the likely pharmacy recruitment rates and completion rates?' and 'What are the likely customer recruitment rates and completion rates?',[19] while in the DECISION + pilot trial a number of thresholds for feasibility were reported, including: 'the proportion of contacted FMGs participating in the pilot study would be 50% or greater' and 'the mean level of satisfaction from family physicians regarding the workshops would be 65% or greater'.[18] Furthermore, Reeves et al. ask the question in relation to producing feedback summaries (their intervention): 'Could ward-level postal surveys be conducted successfully?'.[15] Presumably, if the response rate was very low the investigators would have considered their trial not feasible. These examples indicate that objectives around the feasibility of recruitment, retention, response, and intervention implementation can often be addressed using rates and proportions. In this section, we focus on sample sizes required to estimate rates pertaining to individual participants to a certain degree of accuracy, assuming that the sample in the pilot is representative of the sample to be included in the trial. Although investigators are often interested in rates pertaining to clusters, as described earlier, these will usually be much less precise due to the smaller number of clusters. In addition, methods of cluster recruitment are often different in pilot and main trials, making the extrapolation of recruitment rates in the pilot to rates in the main trial less valid.

Table 2. Specification of fixed parameter, cluster sizes (natural, in main trial and in pilot), ICC value, and number of clusters for 110 simulations.

Specification of scenario			Specification of simulated values for pilot study		
Situation	Natural cluster size	Size of cluster in main trial, m_m	ICC values, ICC	Size of cluster, m_p	Number of clusters, k
Cluster size fixed to be the same as the cluster size in main trial, number of clusters varied	Small	As in pilot	0.01, 0.1, 0.5	2, 5	5, 10, 20, 50
	Large	As in pilot	0.01, 0.05, 0.2	10, 50, 200	5, 10, 15, 30
Number of clusters limited to 10, cluster size varied up to the size planned to be used in the main trial	Large	50	0.01, 0.05, 0.2	5, 10, 15, 25, 50	10
		200	0.01, 0.05, 0.2	5, 10, 25, 50, 200	10
Total number of participants fixed, number of clusters varied (60, 240, 600, 1200, 2400)	Large	200	0.01, 0.05	Determined by combination of total number of participants and number of clusters	5, 10, 20, 30

To explore the effect of pilot sample size on the likely accuracy of different rates, we used half the width of the confidence interval as an expression of the margin of error in a rate, i.e. a measure of precision. This maximum likely error can be expressed as

$$\text{Maximum likely error} = 1.96\sqrt{((1 + (m_p - 1)ICC_{rate})^*_p(1-p)/km_p)}$$

where

ICC_{rate} is the ICC for the rate being estimated
p is the rate of interest

We assumed that most investigators would hope to have a maximum likely error of 10% or less for any calculated rates. We also hypothesised that because rates such as recruitment and response rates may depend to a greater extent on cluster staff and their behaviour, ICCs may be larger for these measures than for outcomes. We therefore used a range of ICCs from 0.05 to 0.5 in our analyses. We considered each of the three situations discussed in Section 3.1. By rearranging the equation above, and setting $p = 0.5$ to result in the largest possible maximum likely error for a given sample size we calculated the following:

(1) The number of clusters needed for a range of fixed cluster sizes and ICC values to be sure of having a maximum likely error of 10%.
(2) The number of participants per cluster needed for a range of fixed numbers of clusters and ICC values to be sure of having a maximum likely error of 10%.
(3) The total number of participants needed for a fixed range of numbers of clusters and ICC values to be sure of having a maximum likely error of 10%.

We used a range of clusters sizes and number of clusters similar to that used in Section 3.1.

4 Results

4.1 Pilot sample size requirements to estimate the sample size required for the main trial

4.1.1 When the pilot study has the same cluster size as planned for the main trial.

The distributions of expected number of clusters in the main trial estimated from pilots with naturally small clusters (cluster sizes of 2 and 5) are shown in Figure 1 and in the online appendix (available at: http://smm.sagepub.com/). In many scenarios the estimate of the number of clusters required for the main trial is biased downwards. This is more likely where pilot studies involve a small number of clusters or the true ICC is high, and generally occurs because the size of the between-cluster variance is underestimated. However, in certain circumstances, the within-cluster variance is also underestimated, exacerbating the problem. For example, for a pilot study with five clusters, two participants per cluster, and a true ICC of 0.50 the true number of clusters required is 351; however, the median estimate of sample size required from the simulations is only 282. This is because both the between-cluster and within-cluster SDs are underestimated (between-cluster SD; true = 1 versus estimated = 0.85: within-cluster SD; true = 1 versus estimated = 0.94).

Increasing the number of clusters included in the pilot reduces the bias, and increases the precision, of the estimates. For example, with two participants per cluster, and an ICC of 0.10,

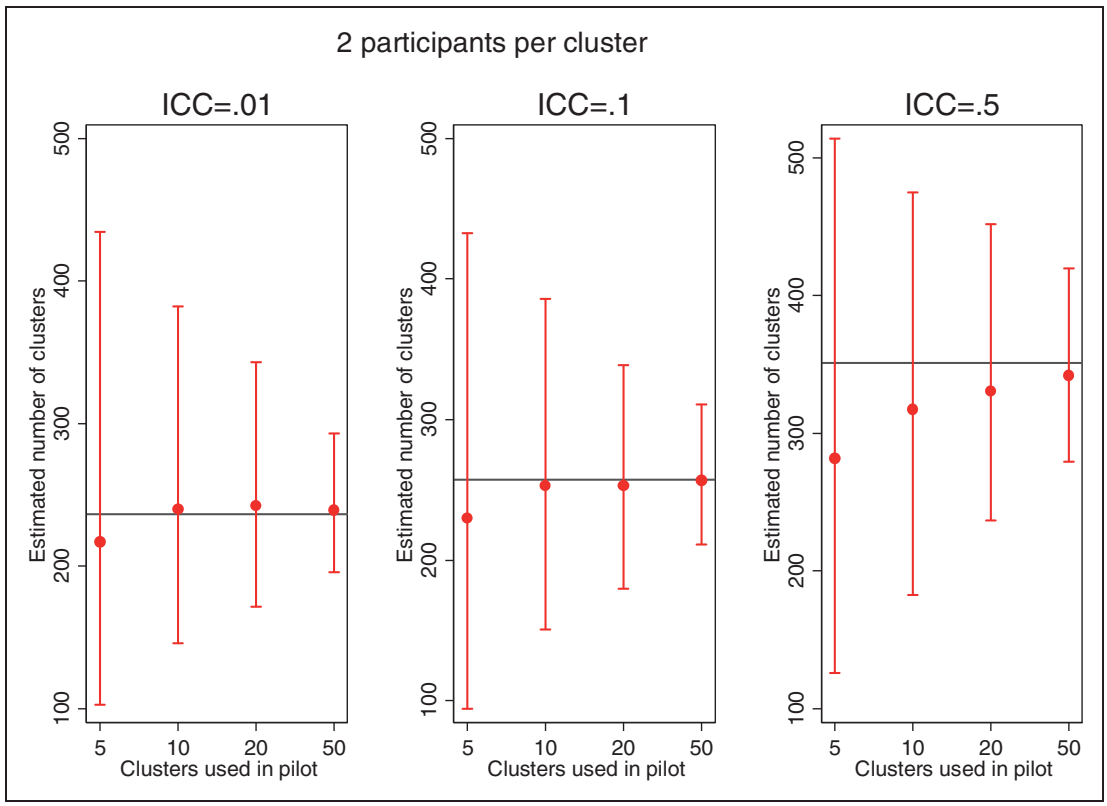

Figure 1. Distribution of expected number of clusters in main trial by number of clusters used in pilot for naturally small units. The y-axis represents the estimated number of clusters required to give 90% power in the main trial based on recruiting two participants to each cluster. The red points represent the median estimate, and the red vertical lines represent the 10th and 90th percentiles of these estimates. The black horizontal line represents the true number of clusters required to give 90% power in the main study.

using 10 or more clusters in the pilot leads to unbiased estimates for the total number of clusters required in the main trial, although precision is still poor with 10 or 20 clusters. However, with a high ICC, even using 50 clusters in the pilot still leads to bias in some situations. Although the bias was most often downward, in rare cases the number of clusters required for the main trial was biased upwards, for example, with five clusters, five participants per cluster, and an ICC of 0.01.

Results for naturally large clusters (50 and 200) are shown in Figure 2 and in the online appendix. As above, the estimated number of clusters required in the main study is biased downwards in most scenarios, and the bias is greater with a small number of clusters in the pilot, or with a high ICC. This is primarily caused by underestimation of the between-cluster variance in the pilot study.

In many scenarios, a pilot of 30 or more clusters would be required to obtain unbiased and accurate estimates; however, in most of the scenarios we considered, this represents a substantial proportion of the total number required for the main study. For example, with 10 participants per cluster and an ICC of 0.05, 68 clusters are required for the main trial. However, to reduce bias and increase precision to acceptable levels, we would need 30 or more clusters in the pilot, which is 44%

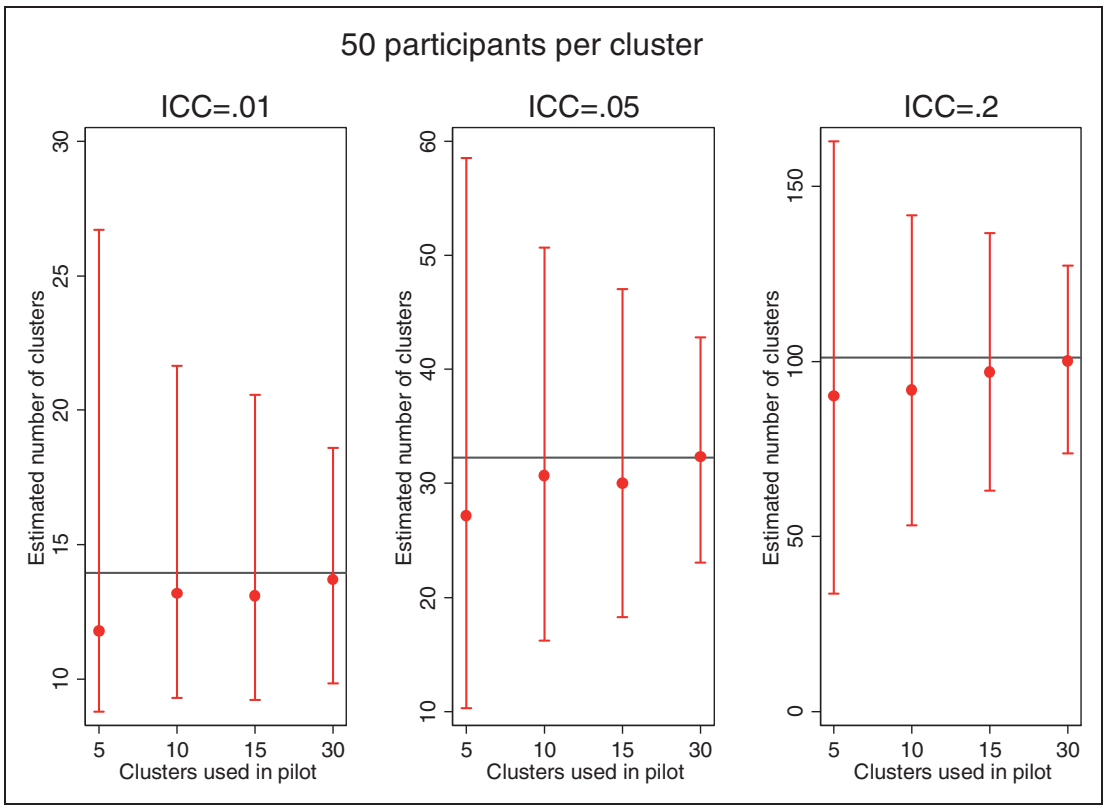

Figure 2. Distribution of number of clusters expected in main trial by number of clusters used in pilot for naturally large units. The y-axis represents the estimated number of clusters required to give 90% power in the main trial based on recruiting 50 participants to each cluster. The red points represent the median estimate, and the red vertical lines represent the 10th and 90th percentiles of these estimates. The black horizontal line represents the true number of clusters required to give 90% power in the main study.

of the required total. Likewise, with 200 participants per cluster and an ICC of 0.05, the number of clusters required in the main trial is 26. A pilot study with only five clusters would give a median estimate of 22 clusters required for the main study, with estimates of less than 7 or greater than 46 each occurring 10% of the time. To get a relatively unbiased estimate would require 15 or 30 clusters in the pilot, which is almost as large as or larger than the number required in the main study (and even in this case, the estimates would still be slightly biased).

4.1.2 A limited number of clusters are available for the pilot study

We simulated this situation for naturally large clusters when the number of clusters is limited to 10. The distribution of expected number of clusters in the main trial is shown in Figure 3 and in the online appendix. Recruiting fewer participants per cluster in the pilot than expected in the main trial leads to increased bias in the sample size estimate and reduced precision. For example, if the main trial recruited 200 participants per cluster with an ICC of 0.05, then 26 clusters would be required. Using a pilot study with 10 clusters and five participants per cluster would give a median estimate of the number of clusters required of 21 (10th–90th percentiles 2–81). Conversely, recruiting 200

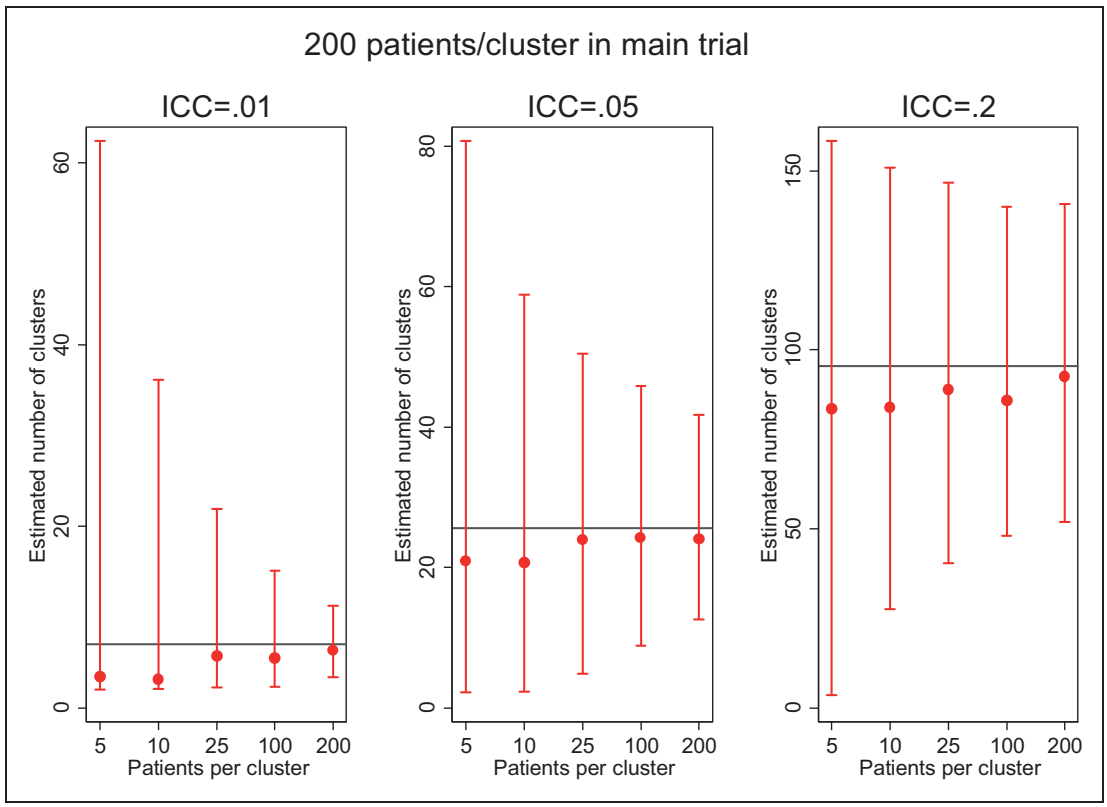

Figure 3. Distribution of number of clusters expected in main trial by number of participants per cluster in pilot, for naturally large units assuming exactly 10 clusters are to be recruited. The y-axis represents the estimated number of clusters required to give 90% power in the main trial based on recruiting 200 participants to each cluster. The red points represent the median estimate, and the red vertical lines represent the 10th and 90th percentiles of these estimates. The black horizontal line represents the true number of clusters required to give 90% power in the main study.

participants to each cluster in the pilot (the same number that would be recruited in the main study) would give a median estimate of 24 clusters (10th–90th percentiles 13–42); increasing the number of participants per cluster has reduced, but not eliminated the bias. Furthermore, this pilot requires 2000 participants, almost half of those that would be required for the main trial.

4.1.3 The total number of participants to be included in the pilot study has been decided

We simulated this situation for naturally large clusters. The distribution of expected number of clusters in the main trial is shown in Figure 4 and in the online appendix. For all scenarios, increasing the number of clusters reduces the bias in the estimates. However, for a small overall sample size, an increased number of clusters also leads to substantially reduced precision as compared to a small number of clusters. This arises because when ICCs are small, as is usually the case for cluster randomised trials, the ICC and its variance are heavily dependent on the within-cluster variance, thus any reduction in the number of individual units within a cluster – by increasing the number of clusters – results in an increase in this variance, and therefore reduced precision in an

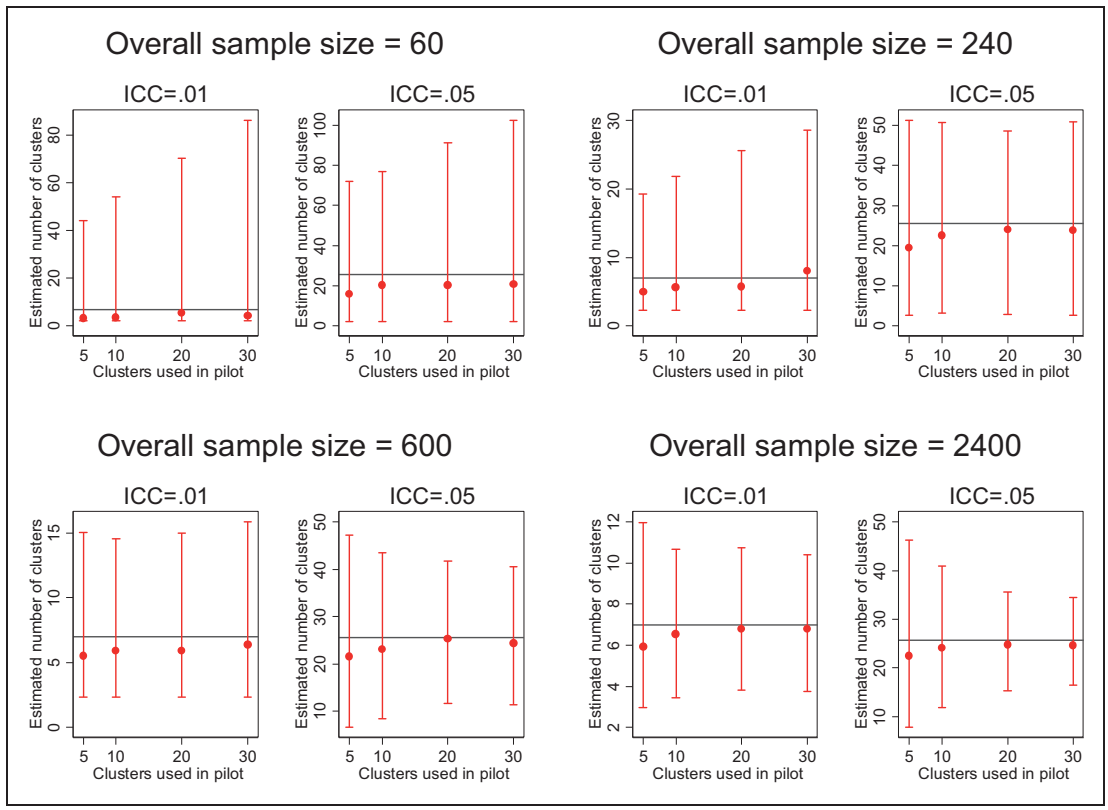

Figure 4. Distribution of number of clusters expected in main trial by number of clusters used in pilot for naturally large units when the total number of participants in the pilot is fixed at (a) 60, (b) 240, (c) 600, (d) 2400. The number of participants in each cluster is the overall sample size divided by the number of clusters, e.g. for an overall sample size of 60 and 10 clusters, the number of participants per cluster is 6. The y-axis represents the estimated number of clusters required to give 90% power in the main trial based on recruiting 200 participants to each cluster. The red points represent the median estimate, and the red vertical lines represent the 10th and 90th percentiles of these estimates. The black horizontal line represents the true number of clusters required to give 90% power in the main study.

estimate of the sample size. It is not until a sample size of 600 or more (with an ICC of 0.05) that increasing the number of clusters both reduces the bias and increases the precision.

4.2 Sample size to fulfil other objectives of a pilot study for cluster randomised trials

Table 3 shows the required size of a pilot study to achieve a maximum likely error in a rate of less than 10%. With naturally small cluster sizes, a relatively large number of clusters are needed even with an ICC of only 0.05. When cluster sizes are naturally large, the number of clusters needed depends on the assumed ICC with, for example, six clusters needed when the ICC is 0.05 and the number of participants per cluster is 200, but 20 clusters needed when the ICC is assumed to be 0.2. When the number of clusters is fixed it is sometimes impossible to achieve a maximum likely error of

Table 3. Size of pilot studies needed under different scenarios to achieve maximum likely error of less than 10% in a rate.

a. Fixed cluster size:

	Number of clusters needed by cluster size, m_p, and ICC			
ICC	$m_p = 2$	$m_p = 5$	$m_p = 50$	$m_p = 200$
0.5	72	58	49	49
0.2	58	35	21	20
0.1	53	27	12	10
0.05	51	23	7	6

b. Fixed number of clusters:

	Cluster size needed by number of clusters, k, and ICC			
ICC	k = 5	k = 10	k = 15	k = 20
0.5	X	X	X	X
0.2	X	X	X	97
0.1	X	219	16	9
0.05	461	18	9	6

X indicates situations in which it is not possible to get a maximum likely error of 10% or less.

10% or less, and where it is possible this can never be achieved with fewer than 120 individual participants.

5 Discussion

We have shown that unless pilot studies for cluster randomised trials include a large number of clusters and individual participants, they are unlikely to be useful in determining sample size for the main trial because estimates from feasibly sized pilot studies lack precision and between-cluster variances are typically biased downwards for the number of clusters investigators usually use. Using parameter estimates from pilot studies with, for example, only 10 clusters risks main trials being considerably under- or over-powered; even when such pilots recruit 50 individuals per cluster (i.e. 500 participants in total) there is greater than 20% chance that investigators will over- or under-recruit by more than 35%. Thus, to provide adequate precision, and to avoid the downward bias in estimating sample size required for a main trial, investigators will often have to invest almost as much effort as for the main trial. Most pilot studies also will not produce very precise estimates of rates such as recruitment and response rates, again because they are usually too small. Furthermore, any estimates will only strictly be valid if the clusters and individuals in the pilot are representative of the larger population from which the main study sample will be drawn. While a sample of individuals in a pilot study for an individually randomised trial may be representative of those in the main trial, this is less likely to be the case for the relatively small number of clusters usually used in a pilot study for a cluster randomised trial.

Our results are in line with a recent paper looking at sample size required for external pilot studies for individually randomised trials in which the aim of the pilot study is to estimate sample size for

the main trial. That paper reported that quite large sample sizes are needed to obtain reasonable precision and that for small sample sizes variance estimates are biased downwards.[4] Our findings also concur with previous research that indicates that the uncertainty around ICC estimates is considerable even for relatively large samples and that estimating ICCs only from pilot studies for cluster randomised trials is not optimal.[3,24] This is in spite of the fact that our brief review indicates that pilots for cluster randomised trials often include considerably more participants than pilots for individually randomised trials.[9] This review also shows that although it is 10 years since the first of several papers outlining the importance of appropriate objectives, design, conduct, and analysis for pilot studies,[9,25] many recent pilots for cluster randomised trials still assess effectiveness.

In this paper, we have only looked at the use of pilot studies to estimate sample size for main trials with continuous primary outcomes and equal cluster sizes, although we do not expect results for binary outcomes to be very different, and if trials have variable cluster sizes estimates are likely to have greater uncertainty than we have reported here. We also assumed both within- and between-cluster variation were to be estimated from the pilot study for a cluster randomised trial. In some cases, total variation or within-cluster variation may be estimated from other sources. Again we do not expect this to make a substantial difference to our results.

The implications of our research are that, in general, pilot studies should not be used to estimate parameters for sample size calculations for cluster randomised trials. In fact, to do so could be misleading. For example, Taylor et al. decided not to proceed with a main trial partly on the basis of large ICC estimates from a small pilot trial with four clusters and 41 patients, even though they report confidence intervals for their ICCs that all include zero.[20] It has been suggested that the most reliable method of estimating ICCs is from patterns in these measures drawn from a number of sources.[3] Sources could include previous trials and observational studies, pilot studies, a growing number of papers that now report lists of ICC values (recent examples are Littenberg and MacLean[25] and Bell and McKenzie[26]), or be based on patterns in ICCs. Patterns in ICCs have been explored by a number of authors and these are summarised in the book by Eldridge and Kerry, where the issue of how to combine estimates from different sources is also discussed.[3] Briefly, it has been shown that ICCs for outcomes that measure the process of care are generally higher than ICCs for clinical outcomes, that ICCs are smaller for clusters with naturally large cluster sizes, that ICCs are smaller for more extreme prevalences. Thus, ICC estimates from a pilot study could potentially contribute to the body of knowledge about patterns in ICCs but should not be relied upon as the only source of information for input into a sample size calculation for a main trial. Furthermore, any downwards bias in between-cluster variation should be acknowledged. Statistical methods to alleviate this bias could usefully be investigated. Where pilot studies report estimates of ICCs without confidence intervals, relatively simple analytic formulae can be used to construct approximate confidence intervals and thus aid interpretation.[3] Some research suggests that internal pilot studies may be good sources of information for ultimate sample size required for a cluster randomised trial, although this research used different methods and scenarios from the ones we used.[27,28] Previous research in relation to individually randomised trials has suggested inflating sample sizes estimated from pilot studies using upper confidence interval limits to counteract imprecision in these estimates.[29] However, a more recent study concludes that if individually randomised trials are designed using parameters from a pilot study then in most cases a trial designed to have 90% power will have 80% power well over 75% of the time,[4] thus in the majority of cases the loss of power due to uncertainty is not catastrophic, and the authors counsel against inflating sample sizes calculated from pilot study estimates. Given that ICC

estimates from pilot studies, in particular, are likely to have very wide confidence intervals we do not recommend inflation of parameter estimates using confidence interval limits as this could result in considerably over-powered trials. As suggested earlier obtaining ICC estimates from a number of sources and using information on patterns in ICCs is likely to be more efficient.

One trend in the interpretation of rates from pilot studies is to set thresholds to be reached in order for a main trial to be considered feasible. The DECISION+ trial described in Table 1 and Section 3.2 is a good example of this.[18] However, our results show that most estimates of rates for pilots of cluster randomised trials are likely to be subject to considerable uncertainty so we suggest researchers need to be cautious about setting definitive thresholds that may then be missed by chance. It may be wise not to rely on single pilot studies to try and estimate such rates.

In spite of the difficulties in relying on pilot studies for quantitative information in preparation for a cluster randomised trial, there are, as outlined in Section 1, compelling reasons for conducting pilot studies for these trials related to understanding the feasibility of trial processes and intervention implementation. We are not suggesting that estimates of ICC and rates should never be estimated in these studies but we strongly suggest that obtaining these estimates should be secondary objectives of a pilot study for a cluster randomised trial and, as discussed earlier, they should never be used to estimate likely values in a main trial without other information. Given the results from our brief review, there remains further work to be done in disseminating the importance of a clear articulation of specific objectives in pilot studies. These objectives, and particularly the primary objective, should then drive discussion and choice of the sample size for pilot studies for cluster randomised trials.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Eldridge SM, Ashby D, Feder GS, et al. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clin Trials* 2004; **1**: 80–90.
2. Farrin A, Russell I, Torgerson D, et al. Differential recruitment in a cluster randomized trial in primary care: the experience of the UK back pain, exercise, active management and manipulation (UK BEAM) feasibility study. *Clin Trials* 2005; **2**: 119–124.
3. Eldridge S and Kerry S. *A practical guide to cluster randomised trials in health services research*. Sussex, UK: Wiley, 2012.
4. Teare MD, Dimairo M, Shephard N, et al. Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials* 2014; **15**: 264.
5. Drahota AK, Ward D, Udell JE, et al. Pilot cluster randomised controlled trial of flooring to reduce injuries from falls in wards for older people. *Age Ageing* 2013; **42**: 633–640.
6. Hawk M. The Girlfriends Project: results of a pilot study assessing feasibility of an hiv testing and risk reduction intervention developed, implemented, and evaluated in community settings. *Aids Educ Prev* 2013; **25**: 519–534.
7. Kerr B, Hawkins TLA, Herman R, et al. Feasibility of scenario-based simulation training versus traditional workshops in continuing medical education: a randomized controlled trial. *Med Educ Online* 2013; **18**: 21312.
8. Okwen M, Lewallen S and Courtright P. Primary eye care skills scores for health workers in routine and enhanced supervision settings. *Public Health* 2014; **128**: 96–100.
9. Lancaster GA, Dodd S and Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *J Eval Clin Pract* 2004; **10**: 307–312.
10. Thabane L, Ma J, Chu R, et al. A tutorial on pilot studies: the what, why and how. *BMC Med Res Methodol* 2010; **10**: 1.
11. Shanyinde M, Pickering RM and Weatherall M. Questions asked and answered in pilot and feasibility randomized controlled trials. *BMC Med Res Methodol* 2011; **11**: 117.

12. Mass CJM and Hox JJM. Sufficient sample sizes for multilevel modeling. *Methodology* 2005; **1**: 7.
13. Diaz-Ordaz K, Froud R, Sheehan B, et al. A systematic review of cluster randomised trials in residential facilities for older people suggests how to improve quality. *BMC Med Res Methodol* 2013; **13**: 127.
14. Teut M, Schnabel K, Baur R, et al. Effects and feasibility of an Integrative Medicine program for geriatric patients – a cluster-randomized pilot study. *Clin Interv Aging* 2013; **8**: 953–961.
15. Reeves R, West E and Barron D. Facilitated patient experience feedback can improve nursing care: a pilot study for a phase III cluster randomised controlled trial. *BMC Health Serv Res* 2013; **13**: 259.
16. Jairath V, Kahan BC, Gray A, et al. Restrictive vs liberal blood transfusion for acute upper gastrointestinal bleeding: rationale and protocol for a cluster randomized feasibility trial. *Transfus Med Rev* 2013; **27**: 146–153.
17. Colon-Emeric CS, McConnell E, Pinheiro SO, et al. CONNECT for better fall prevention in nursing homes: results from a pilot intervention study. *J Am Geriatr Soc* 2013; **61**: 2150–2159.
18. LeBlanc A, Legare F, Labrecque M, et al. Feasibility of a randomised trial of a continuing medical education program in shared decision-making on the use of antibiotics for acute respiratory infections in primary care: the DECISION plus pilot trial. *Implement Sci* 2011; **6**: 5.
19. Porteous T, Smith S, Sheikh A, et al. ''Help for hay fever'': a pilot cluster RCT of a community pharmacy delivered goal-focussed intervention for people with intermittent allergic rhinitis. *Clin Exp Allergy* 2013; **43**: 1444–1445.
20. Taylor WJ, Brown M, William L, et al. A pilot cluster randomized controlled trial of structured goal-setting following stroke. *Clin Rehabil* 2012; **26**: 327–338.
21. Donner A. An empirical-study of cluster randomization. *Int J Epidemiol* 1982; **11**: 283–286.
22. Eldridge S, Cryer C, Feder G, et al. Sample size calculations for intervention trials in primary care randomizing by primary care group: an empirical illustration from one proposed intervention trial. *Stat Med* 2001; **20**: 367–376.
23. Gulliford MC, Ukoumunne OC and Chinn S. Components of variance and intraclass correlations for the design of community-based surveys and intervention studies – data from the Health Survey for England 1994. *Am J Epidemiol* 1999; **149**: 876–883.
24. Ukoumunne OC. A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Stat Med* 2002; **21**: 3757–3774.
25. Littenberg B and MacLean CD. Intra-cluster correlation coefficients in adults with diabetes in primary care practices: the Vermont Diabetes Information System field survey. *BMC Med Res Methodol* 2006; **6**: 20.
26. Bell ML and McKenzie JE. Designing psycho-oncology randomised trials and cluster randomised trials: variance components and intra-cluster correlation of commonly used psychosocial measures. *Psycho-oncology* 2013; **22**: 1738–1747.
27. Lake S, Kammann E, Klar N, et al. Sample size re-estimation in cluster randomization trials. *Stat Med* 2002; **21**: 1337–1350.
28. van Schie S and Moerbeek M. Re-estimating sample size in cluster randomised trials with active recruitment within clusters. *Stat Med* 2014; **33**: 3253–3268.
29. Sim J and 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**: 301–308.