OPEN ACCESS

Ethical implications of excessive cluster sizes in cluster randomised trials

Karla Hemming,[1] Monica Taljaard,[2] Gordon Forbes,[3] Sandra M Eldridge,[3] Charles Weijer[4]

[1]Institute of Applied Health Research, University of Birmingham, Birmingham, UK
[2]Department of Clinical Epidemiology, Ottawa Hospital Research Institute, Ottawa, Canada
[3]Primary Care and Public Health, Queen Mary University of London, London, UK
[4]Rotman Institute of Philosophy, Western University, London, UK

Correspondence to
Dr Karla Hemming, School of Health and Population Sciences, University of Birmingham, Birmingham b152tt, UK; k.hemming@bham.ac.uk

Check for updates

ABSTRACT

The cluster randomised trial (CRT) is commonly used in healthcare research. It is the gold-standard study design for evaluating healthcare policy interventions. A key characteristic of this design is that as more participants are included, in a fixed number of clusters, the increase in achievable power will level off. CRTs with cluster sizes that exceed the point of levelling-off will have excessive numbers of participants, even if they do not achieve nominal levels of power. Excessively large cluster sizes may have ethical implications due to exposing trial participants unnecessarily to the burdens of both participating in the trial and the potential risks of harm associated with the intervention. We explore these issues through the use of two case studies. Where data are routinely collected, available at minimum cost and the intervention poses low risk, the ethical implications of excessively large cluster sizes are likely to be low (case study 1). However, to maximise the social benefit of the study, identification of excessive cluster sizes can allow for prespecified and fully powered secondary analyses. In the second case study, while there is no burden through trial participation (because the outcome data are routinely collected and non-identifiable), the intervention might be considered to pose some indirect risk to patients and risks to the healthcare workers. In this case study it is therefore important that the inclusion of excessively large cluster sizes is justifiable on other grounds (perhaps to show sustainability). In any randomised controlled trial, including evaluations of health policy interventions, it is important to minimise the burdens and risks to participants. Funders, researchers and research ethics committees should be aware of the ethical issues of excessively large cluster sizes in cluster trials.

THE CLUSTER RANDOMISED CONTROLLED TRIAL

The cluster randomised trial (CRT) is increasingly being used to evaluate interventions that cannot be evaluated using the conventional individually randomised trial. CRTs proceed by randomising clusters of individuals to intervention or control conditions, with typical examples of clusters being hospitals, general practices and schools. These designs are used across a breadth of contexts, and evaluate a diverse range of interventions from health services or policy interventions[1] to drug therapies.[2–4]

As with all research involving human participants, CRTs should be conducted in accord with appropriate scientific and ethical principles. One scientific and ethical standard is that an appropriate power calculation should be carried out, to ensure the study has enough participants to detect the target difference. However, it has been argued that trials that accrue needlessly large numbers of participants may violate ethical principles by exposing excessive numbers of participants to the burdens and risks of study participation.[5 6] We use the term 'excessively large cluster sizes' to notionally represent cluster sizes in which some observations make a negligible contribution to the study's primary aims.

In this paper, we argue that this issue is of particular importance in a CRT. This is because, uniquely to a CRT, some participants in a cluster may contribute little information to the study.[7–10] The statistical implications of these excessive cluster sizes have been outlined elsewhere, including how simple changes to the trial design can deliver close to the desired statistical power without compromising the value of the trial.[11] However, the ethical implications of excessive cluster sizes may vary according to the type of CRT, but have not yet been fully articulated.

When interventions are low risk and study participation poses little or no burden, for example, when data are routinely collected and non-identifiable,[12 13] there may be no ethical implications or implications may be inconsequential. However, excessive cluster sizes will have ethical implications when the intervention poses some risk

to participants or there is some burden posed with trial participation. The objective of this paper is to raise awareness of these issues to prompt funders, researchers and ethics committees to ensure maximum social and scientific value of cluster trials, while at the same time minimise the number of participants who are exposed to the burden and risks of trial participation.

We introduce two case studies with large cluster sizes. These case studies raise a number of important issues, many of which are common to large pragmatic CRTs. We discuss each of these case studies in turn, considering for each the ethical implications of including excessive cluster sizes. We consider in particular the risks and burdens for participants, and whether these can be justified when it is known that some participants provide little information to the primary outcome.
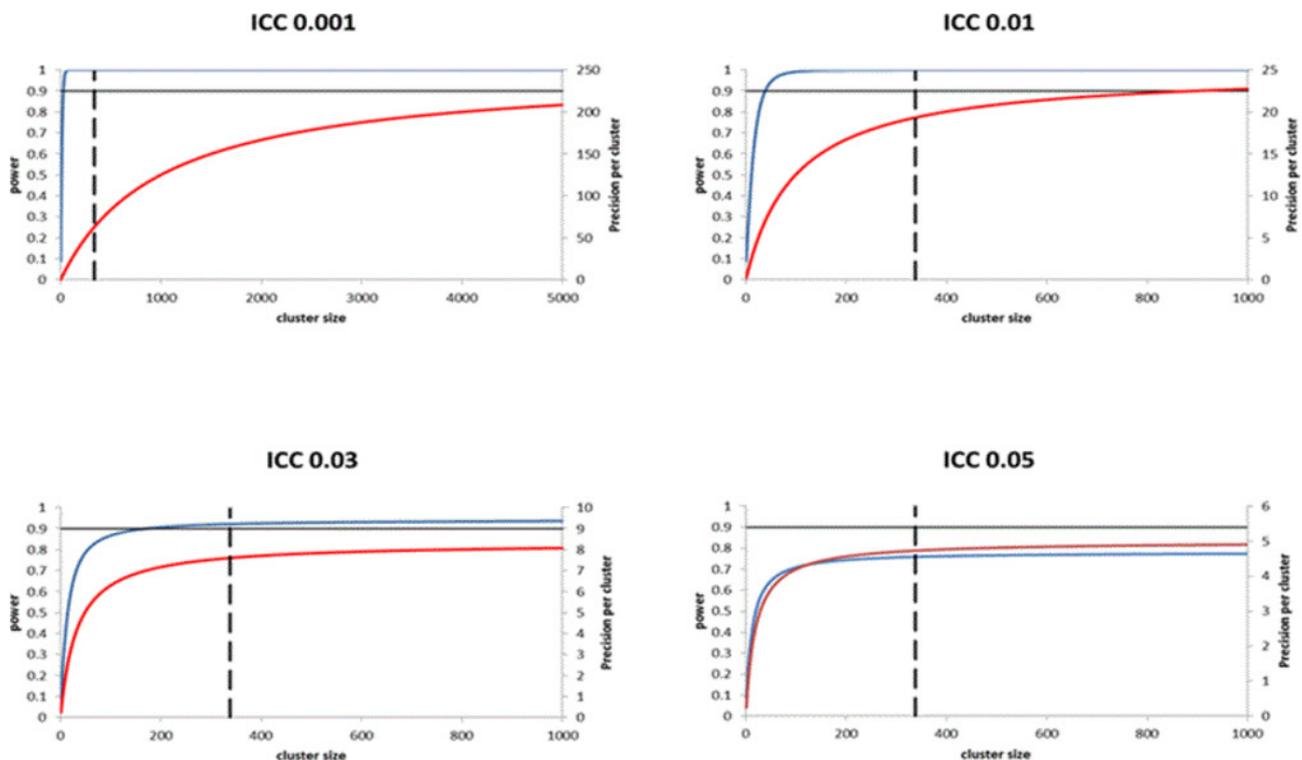
DIMINISHING RETURNS IN CRTS

It has been known for some time that the power in a CRT begins to reach a plateau as the cluster size is increased,[7–10] and a set of practical recommendations for designing cluster trials so that they make efficient use of their data has recently been published.[11] At the core of these recommendations is the fact that, in cluster trials, increasing cluster sizes, even by a vast amount, may result in negligible gains in power. This levelling-off is dependent on several design parameters,

including the intracluster correlation coefficient (ICC) (which measures the extent to which observations within a cluster are correlated) as well as the target effect size. Target effect sizes should be clearly justified and should ideally be based on the minimally important difference.[14 15]

How the increase in power begins to level off as the cluster sizes increase is illustrated in figure 1 using a series of power curves (for full details of calculations, see Hemming et al[11]). For example, when the ICC is 0.01 the power to detect a standardised effect size of 0.25 with 12 clusters per arm plateaus when the cluster size is about 100. This means that, when the standardised effect size of 0.25 is truly the minimally important clinical difference, participants above a threshold of 100 add little value to the analysis of the primary outcome of the study.

An additional complication is that while participants' contribution to the power may be negligible, participants might still make a valuable contribution to the study should the minimal important effect be smaller than the target effect assumed in the sample size calculation. For example, in figure 1, when the ICC is 0.01 the precision (ie, 1/variance of the treatment effect) continues to increase up to cluster sizes of about 1000. This means that while cluster sizes of about 100 would be sufficient to detect an effect size of 0.25, the CI width around smaller effect sizes will

Figure 1 Illustration of diminishing returns in precision as cluster size increases, for typical intracluster correlation coefficient (ICC) values. Curves show increases in power (blue line) and precision (red line) as cluster size increases. All power curves correspond to a cluster randomised trial with 12 clusters per arm, designed to detect a standardised effect size of 0.25. At a significance level of 5%, 338 individuals per arm are required to yield 90% power under individual randomisation. Precision curves are independent of effect size (assumed to be continuous outcome). Dashed lines represent required sample size per arm under individual randomisation.

continue to decrease up to cluster sizes of about 1000. This may be valuable if the actual effect size is smaller than the target difference but still clinically important.

It is important to bear in mind that these calculations depend on estimated parameters such as the ICC. Ideally, power and precision curves should take into account the uncertainty in these estimates and should be considered across a range of scenarios, as with any power calculation. All our calculations have assumed equal cluster sizes and large sample theory. Specific recommendations have been provided for inflating the sample size to accommodate varying cluster sizes using the coefficient of cluster size variation, and this equates to increasing the number of clusters per arm by 2 or 3.[16 17] There are other CRTs that are statistically more efficient, that is, deliver the same power for a substantially reduced total sample size. These include the cluster randomised cross-over trial (only a feasible design choice when the intervention can be withdrawn) and some trials with premeasurements in a baseline period.[18]

ETHICAL ISSUES IN STUDIES WITH EXCESSIVELY LARGE CLUSTER SIZE

A cornerstone of the ethics of medical research is that the risks to study participants must be reasonable when considering the potential benefits to them and to society. Researchers and research ethics committees must (1) minimise risks to participants consistent with sound scientific design and (2) ensure that knowledge benefits outweigh risks.[19 20] Studies that accrue needlessly large numbers of participants may violate one or both requirements by exposing some participants to research burdens and risks needlessly.

Risks to research participants may include exposure to a treatment or service that is inferior (eg, when a trial continues beyond the point at which the inferiority of an intervention could have been detected, thereby needlessly exposing intervention participants to an inferior treatment); being denied access in a timely manner to treatment (eg, when trial continues beyond the point at which the superiority of an intervention could have been detected, thereby needlessly exposing control participants to an inferior treatment); or exposure to the burdens associated with data collection or other non-therapeutic study procedures.

Burdens and risks to society also include being denied access to treatments or services when a trial continues beyond the point at which the superiority of an intervention could have been detected. Other burdens and risks to society include a financial cost and delay in answering the study question.

Appropriate identification of research participants is essential to conducting a proper analysis of benefits and harms: only those individuals who are research participants logically fall under the remit of the research ethics committee. Therefore, identification of research participants can help determine if excessive cluster sizes are consequential. The Ottawa Statement provides explicit criteria for identification of research participants in a CRT[20]: in brief, individuals are research participants when they are targeted by a study intervention and/or have their identifiable private information collected for research.

Implications in lower risk or low burden settings

Many CRTs evaluate interventions such as health promotion or knowledge translation interventions, which pose little burden or risk to trial participants. Burdens are minimal because outcomes might be obtained in de-identified form using routinely collected data sources. In this situation, the data might be considered available for 'free'. That is, once data linkage has been established, the number of observations included per cluster, whether from 500 or 5000 participants, is largely immaterial. If the intervention is low risk and routinely collected outcome data are available, there are few ethical implications of large cluster sizes. Some CRTs evaluate interventions targeted at health professionals but assess outcomes on patients. In such trials if patients are not interacted with for elicitation of outcome data they will not be research participants. However, healthcare professionals might still be research participants, although are frequently not identified as such.[21]

In case study 1, the ASCEND trial,[22] cluster sizes were set by the average number of invitations that each screening region posted each day. Figure 2 shows that a similar level of power could have been achieved with a much smaller cluster size (while retaining 50 clusters) for the anticipated ICC of 0.0002 (see online supplementary appendix 1 for more details on the calculations). However, the figure also demonstrates that all observations would make a material contribution to the precision of the treatment effect, if the effects of the intervention were smaller than that allowed for in the power calculation. Small absolute changes might be materially important to the screening programme, but if so should be specified in advance.[14]

If it is the case that smaller effects were not important, not all observations would have made a material contribution to the prespecified primary outcome analysis. Smaller cluster sizes might then have been achievable, for example, by composing half region-days instead of full region-days and running the trial for 10 half days instead of 10 full days (and so retaining 50 clusters). However, logistics may have meant that it was not practically possible to change the leaflet type midway through the day, or this may have added cost implications. In this situation, an acknowledgement and awareness of the negligible contribution of a substantial proportion of the data could have allowed important secondary outcomes (or safety outcomes, in other trials where applicable) to be included as fully powered analyses.

Case study 1 The ASCEND trial: evaluation of a low-risk intervention with excessively large sizes

Background

A cluster randomised trial published in *The Lancet* in 2016 evaluated an intervention to increase colorectal screening uptake and reduce the socioeconomic differences between uptake rates.[22] The intervention consisted of a leaflet designed specifically to communicate complex information to readers with low literacy (trial 1 in paper). The trial was powered for a primary analysis of testing whether the effect of the intervention differed across different socioeconomic status groups (ie, an interaction). This trial lasted 10 consecutive working days and was run across five UK screening regions. The outcome data were routinely collected. The unit of randomisation was a region-day (the clusters); thus, the trial included 50 clusters: 25 randomised to receive the intervention and 25 randomised to current practice. The cluster size was fixed by the number of invitations sent out (per region) in any given day.

Power calculation

The trial was designed with 90% power and 5% significance and required 13 500 observations per arm under individual randomisation. This was inflated to allow for clustering, assuming an intracluster correlation coefficient of 0.0002. The trial protocol states that the sample size was further inflated to increase generalisability and reduce internal validity issues associated with trials with a small number of clusters. The final number of clusters was 5 regions over 10 days (ie, 50 clusters).

Number of participants included

The trial enrolled 50 clusters and a total of 163 255 people contributed to the primary outcome, equating to an average cluster size of 3265.

In this trial, citizens were research participants: the study intervention is a leaflet that seeks to increase their participation in cancer screening. Researchers should consider carefully the implications of excessive cluster sizes on the study research participants. However, because the intervention likely poses negligible risks or burdens to participants and may even be beneficial for low socioeconomic groups, and because larger cluster sizes do not increase the costs or time required for the trial, the excessive cluster sizes are likely to be inconsequential. Nonetheless, the benefits to society may have been increased if the trial had made more efficient use of the data (eg, prespecified important secondary analyses).

Implications in higher risk or higher burden settings

Trial participation can pose some burden or risk for participants, particularly when evaluating higher risk interventions. A small but not insignificant minority of CRTs evaluate higher risk interventions, including medicinal products such as vaccines,[23] and what constitutes low risk may be contentious, as the second case study illustrates.[24–26] Moreover, even trials of low-risk interventions can pose some burden to participants, for example, the burden of questionnaire completion and other procedures to elicit outcomes (such as blood draws).[19] In these studies, individual patients or healthcare professionals are more likely to be human research participants since interventions might either be directly or indirectly targeted at them, and they may also be interacted with for elicitation of outcome data.

In case study 2, the Flexibility in Duty Hour Requirements for Surgical Trainees (FIRST) trial,[27] cluster sizes were set by the average number of patients admitted to each hospital over the duration of 1 year. Figure 3 demonstrates that for this design, cluster sizes above a threshold of about 1000 make minimal contribution to power (as per the effect size specified and under a non-inferiority
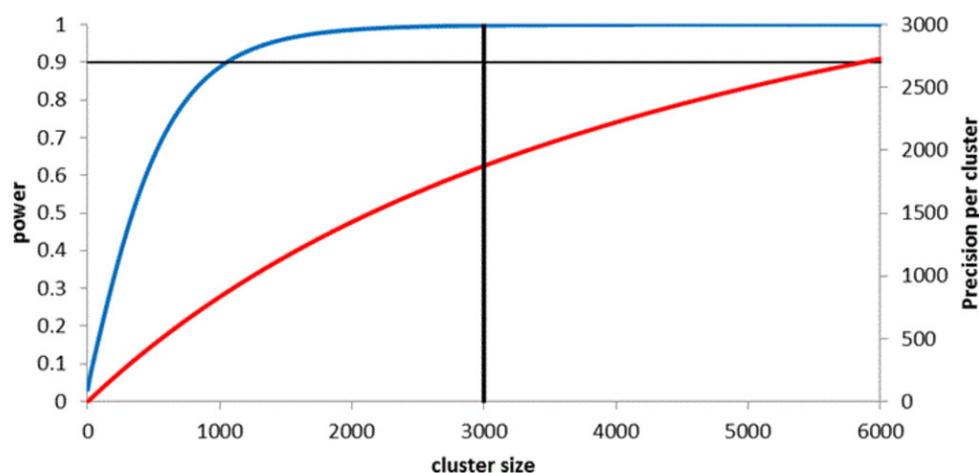
Figure 2 Power and precision curves for ASCEND trial. Curves show increases in power (blue line) and precision (red line) as cluster size increases, assuming a cluster randomised trial with 25 clusters per arm, designed to detect a standardised effect size of 0.04, at a significance level of 5% (which requires a sample size of 13 500 per arm under individual randomisation), and assuming an intracluster correlation coefficient of 0.0002. Vertical solid line represents cluster size in the study (3000). Full details of calculations are provided in online supplementary appendix 1.

Case study 2   The FIRST trial: evaluation of a higher risk intervention with excessively large cluster sizes

Background
The FIRST trial was a cluster randomised trial, published in *The New England Journal of Medicine* in 2016, evaluating the effect of flexible maximum shift length policies on patient outcomes (serious complication or death) and trainee doctors' well-being.[24] The cluster randomised trial enrolled 117 general surgery residency programmes (the clusters) in the USA and ran from 1 July 2014 to 30 June 2015. Each residency programme was randomly assigned to either the standard working hour policy or a more flexible policy with waived rules on shift lengths. Data on primary patient outcomes were ascertained from routinely collected data.

Power calculation
The trial had a non-inferiority design, powered at 80% on a composite of serious patient complications, which had an estimated baseline rate of 9.94%; with an absolute non-inferiority margin of 1.25%, equating to an upper bound of 11.19%; and an adjusted alpha value of 0.04 to allow for an interim analysis. To replicate this non-inferiority calculation, we double the alpha value (ie, 0.04×2) and retain a two-sided test. This requires a sample size of 7750 per arm under individual randomisation. The intracluster correlation coefficient was estimated to be 0.004.

Number of participants included
The trial enrolled 117 clusters and a total of 138 691 patients contributed to the primary outcome, equating to an average cluster size of 1185 patients.

design) or precision. Reducing the cluster sizes from about 1200 to 600 would have reduced power only marginally from about 85% to 79% (see online supplementary appendix 1 for full details of calculations).

The obvious way to reduce the cluster size in this study is to reduce the duration of the trial (as outcomes were obtained anonymously from routinely collected sources, selecting a random sample of outcomes would not be meaningful). For example, conducting the trial over half a year instead of a full year would have reduced the cluster sizes from about 1200 to 600. Such a change would have reduced power only marginally from about 85% to 79% (figure 3). In so doing the finding that the change to flexible working hour policies was no worse for patient outcomes or resident well-being might possibly have been reported after only 6 months instead of 1 year, allowing this non-inferior intervention to be rolled out to other hospitals sooner.

There may have been practical or scientific reasons for implementing the new resident duty hour policies over a 1-year period. For example, it may have been difficult to change resident duty hour policies halfway through an academic year. Yet, as in case study 1, an acknowledgement and awareness of the negligible contribution of a substantial proportion of the data could have allowed important secondary or safety outcomes to be included as fully powered important secondary analyses. These might include demonstrating sustainability of the effect over a prolonged period of time. But, again, such a justification needs to be prespecified, and the analysis formulated in such a way so as to demonstrate these effects.
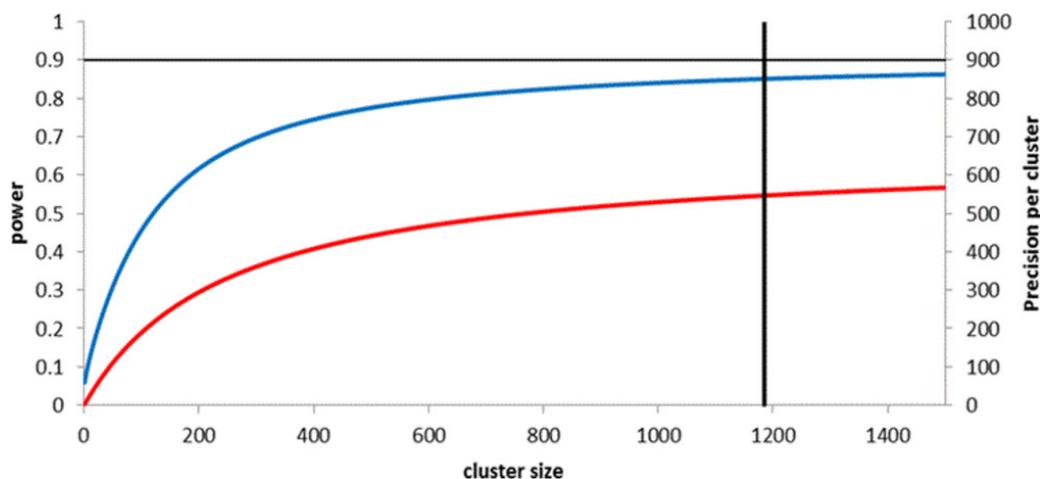
Figure 3   Power and precision curves for the FIRST trial. Curves show increases in power (blue line) and precision (red line) as cluster size increases, assuming a cluster randomised trial with 45 clusters per arm, designed to detect a difference from 9.94% to 11.19%, at a significance level of 8%, and assuming an intracluster correlation coefficient of 0.004. Vertical solid line represents cluster size in the study (1185). Full details of calculations are provided in online supplementary appendix 1.

In this trial, patients were not research participants since the study interventions were neither targeted at them; there were no additional burdens due to data collection; and all data that were used were non-identifiable.[18] However, trainee doctors are research participants, being the direct target of the intervention. The implications of excessive cluster sizes on the trainee doctors should therefore be carefully considered (AR Horn *et al*, under review). The trial received criticism from some patient and doctor groups, citing increased risks associated with the intervention.[24–26] Shortening the duration of the trial may have ameliorated some of the concerns of those criticising the trial, again increasing the social benefit of the trial.

This second case study is therefore an example where researchers (including those on funding and ethical review boards) could have benefited from the knowledge that a substantial proportion of the data made very minimal contribution to the primary analysis. While viewed by some as a low-risk intervention, there was much debate over the risks of the trial. The trial should be designed to answer the question by exposing the minimum number of participants as possible to these perceived/debated trial risks. While there may have been reasons as to why the trial was needed to run for 1 year (this was the cause of the large cluster sizes), these were not articulated, and because this fact was not identified by the researchers the 'extra' data were not put to good use (ie, a prespecified and fully powered secondary analysis).

CONCLUSIONS

Researchers, funders and research ethics committees need to carefully consider whether all participants in CRTs will contribute information to the study (box 1). The point beyond which further recruitment or inclusion of participants makes a trivial contribution to the treatment effect estimate can most easily be ascertained by the use of power or precision curves.[11] This issue is particularly important when the trial is evaluating an intervention that poses substantial burdens or risks to the participant, or when large cluster sizes increase social or financial burden.

For CRTs where individuals are not considered human research participants or CRTs evaluating relatively innocuous interventions that pose little or no risk to participants, excessive cluster sizes are unlikely to be ethically problematic (case study 1).[22] In these situations, when observations are 'free', researchers should consider powering the trial for important subgroup differences or safety outcomes so that data do not go to waste.[28] When there are trial or data collection costs, a more efficient trial design may improve the efficiency of the trial and its social value.

For CRTs evaluating interventions that are associated with some burden or risk to participants, or costs to society, any excessive cluster sizes must be justified (case study 2).[27] Possible justifications include a desire

---

Box 1 Recommendations to avoid undue exposure to burden or risks associated with participation in a cluster trial

► Due consideration must be given to the balance between recruiting large cluster sizes in pursuit of desired but notional levels of power and the burdens and risks to the individual participants. Proper identification of human research participants in cluster randomised trials can aid such considerations.

► Ethically designed trials might reduce the total number of participants being exposed to the burdens associated with trial participation and the possible risks associated with the intervention being evaluated. This might be achieved by recruiting a smaller number of participants from clusters or running trials for a shorter duration.

► When data are routinely collected and the intervention has minimal risk, then large cluster sizes may be acceptable as there will be minimal if any burden on participants. Furthermore, these large cluster sizes if prespecified may lead to improvement in power for secondary analyses or allow analyses showing sustainability of effect.

► When the intervention carries some risk or there are participant burdens related to data collection, then trials might be designed with slightly less than nominal levels of power to avoid excessive numbers of participants making a minimal or negligible contribution to the trials' primary objective.

---

to demonstrate sustainability, to estimate learning effects or to allow time for the intervention to become embedded in practice. Occasionally a net benefit might be provided to those participating in the trial, for example access to a known beneficial treatment or service that is not available outside of the study in a lower income or middle-income country, for example; or in vaccination studies in which there are both direct and indirect effects of the intervention. Otherwise excessive cluster sizes potentially expose individuals to needless risks while offering little or no additional social value.

REFERENCES

1 Guthrie B, Treweek S, Petrie D, *et al*. Protocol for the Effective Feedback to Improve Primary Care Prescribing Safety (EFIPPS) study: a cluster randomised controlled trial using ePrescribing data. *BMJ Open* 2012;2:e002359.

2 Palmer AC, Schulze KJ, Khatry SK, *et al*. Maternal vitamin A supplementation increases natural antibody concentrations of preadolescent offspring in rural Nepal. *Nutrition* 2015;31:813–9.

3 Lehtinen M, Apter D, Baussano I, *et al*. Characteristics of a cluster-randomized phase IV human papillomavirus vaccination effectiveness trial. *Vaccine* 2015;33:1284–90.

4 Palmu AA, Jokinen J, Borys D, *et al*. Effectiveness of the ten-valent pneumococcal Haemophilus influenzae protein D conjugate vaccine (PHiD-CV10) against invasive pneumococcal disease: a cluster randomised trial. *Lancet* 2013;381:214–22.

5 Button KS, Ioannidis JP, Mokrysz C, *et al*. Power failure: why small sample size undermines the reliability of neuroscience [correction appear in: *Nat Rev Neurosci* 2013;14:451]. *Nat Rev Neurosci* 2013;14:365–76.

6 Altman DG. Statistics and ethics in medical research: III How large a sample? *Br Med J* 1980;281:1336–8.

7 Allan D. Some aspects of the design and analysis of cluster randomization trials. *J R Stat Soc* 1998:95–113. Series C (Applied Statistics) 47.1.

8 Donner A, Klar N. Statistical considerations in the design and analysis of community intervention trials. *J Clin Epidemiol* 1996;49:435–9.

9 Guittet L, Giraudeau B, Ravaud P. A priori postulated and real power in cluster randomized trials: mind the gap. *BMC Med Res Methodol* 2005;5:25.

10 Hemming K, Girling AJ, Sitch AJ, *et al*. Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. *BMC Med Res Methodol* 2011;11:102.

11 Hemming K, Eldridge S, Forbes G, *et al*. How to design efficient cluster randomised trials. *BMJ* 2017;358:j3064.

12 Gulliford MC, van Staa TP, McDermott L, *et al*. Cluster randomized trials utilizing primary care electronic health records: methodological issues in design, conduct, and analysis (eCRT Study). *Trials* 2014;15:220.

13 van Staa TP, Dyson L, McCann G, *et al*. The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. *Health Technol Assess* 2014;18:1-146.

14 Cook JA, Hislop J, Altman DG, *et al*. Specifying the target difference in the primary outcome for a randomised controlled trial: guidance for researchers. *Trials* 2015;16:12.

15 Charles P, Giraudeau B, Dechartres A, *et al*. Reporting of sample size calculation in randomised controlled trials: review. *BMJ* 2009;338:b1732.

16 Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol* 2006;35:1292–300.

17 van Breukelen GJ, Candel MJ. Comments on 'Efficiency loss because of varying cluster size in cluster randomized trials is smaller than literature suggests'. *Stat Med* 2012;31:397–400.

18 Hemming K, Haines TP, Chilton PJ, *et al*. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ* 2015;350:h391.

19 Smalley JB, Merritt MW, Al-Khatib SM, *et al*. Ethical responsibilities toward indirect and collateral participants in pragmatic clinical trials. *Clin Trials* 2015;12:476–84.

20 Taljaard M, Weijer C, Grimshaw JM, *et al*. The Ottawa Statement on the ethical design and conduct of cluster randomised trials: precis for researchers and research ethics committees. *BMJ* 2013;346.

21 Taljaard M, Hemming K, Shah L, *et al*. Inadequacy of ethical conduct and reporting of stepped wedge cluster randomized trials: Results from a systematic review. *Clin Trials* 2017;14:333–41.

22 Wardle J, von Wagner C, Kralj-Hans I, *et al*. Effects of evidence-based strategies to reduce the socioeconomic gradient of uptake in the English NHS Bowel Cancer Screening Programme (ASCEND): four cluster-randomised controlled trials. *Lancet* 2016;387:751–9.

23 Taljaard M, McRae AD, Weijer C, *et al*. Inadequate reporting of research ethics review and informed consent in cluster randomised trials: review of random sample of published trials. *BMJ* 2011;342:d2496.

24 Rosenbaum L. Leaping without looking--duty hours, autonomy, and the risks of research and practice. *N Engl J Med* 2016;374:701–3.

25 Carome MA, Wolfe SM, Almashat S, *et al*. Letter to Jerry Menikoff, director, and Kristina Borror, director, Division of Compliance Oversight, Office for Human Research Protections, Department of Health and Human Services, regarding iCOMPARE trial. http://www. citizen.org/documents/ 2283. pdf (Nov 19 2015).

26 Carome MA, Wolfe SM, Almashat S, *et al*. Letter to Jerry Menikoff, director, and Kristina Borror, director, Division of Compliance Oversight, Office for Human Research Protections, Department of Health and Human Services, regarding FIRST trial.. (http://www. citizen. org/ documents/ 2284. pdf). (Nov 19 2015).

27 Bilimoria KY, Chung JW, Hedges LV, *et al*. National cluster-randomized trial of duty-hour flexibility in surgical training. *N Engl J Med* 2016;374:713–27.

28 Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. *J Clin Epidemiol* 1995;48:23–40.