The double jeopardy of clustered measurement and cluster randomisation

Michael S Kramer,[1 2] Richard M Martin,[3 4] Jonathan A C Sterne,[3] Stanley Shapiro,[2] Mourad Dahhou,[1] Robert W Platt[1 2]

Michael S Kramer and colleagues suggest that double clustering might explain the negative results of some cluster randomised trials and propose solutions

[1]Department of Pediatrics, McGill University Faculty of Medicine, Montreal, Canada
[2]Department of Epidemiology and Biostatistics, McGill University Faculty of Medicine
[3]Department of Social Medicine, University of Bristol, Bristol
[4]MRC Centre for Causal Analysis, University of Bristol
Correspondence to: M S Kramer
Montreal Children's Hospital, 2300 Tupper Street (Les Tourelles), Montreal, Quebec H3H 1P3
michael.kramer@mcgill.ca

Cite this as: *BMJ* 2009;339:b2900
doi: 10.1136/bmj.b2900

Cluster randomised trials have become popular for evaluating health service and public health interventions. The clusters are groups of individuals, such as families, schools, clinics, hospitals, or entire communities. Cluster randomised trials provide the rigours of randomisation, while reducing treatment "contamination"; contact between subjects randomised to two (or more) interventions may expose them to both interventions and thus reduce differences in outcome between the groups.[1 2] In addition, cluster randomisation is often more feasible than individual randomisation because group dynamics can make it easier to change practices or behaviours within an overall group than to change practices or behaviours among individuals within the same group.

But cluster randomisation also has some disadvantages. Primary among these is reduced statistical power due to within cluster correlation of outcomes. In other words, individuals within the same cluster are more likely to experience the same study outcome than those in other clusters, irrespective of treatment allocation. This within cluster correlation is usually assessed with the intraclass correlation coefficient (ICC). This coefficient is a measure of how much more similar the values of an outcome are within the same cluster than among different clusters randomised to the same treatment. It is formally defined as the ratio of the between cluster variance to the total variance. If all variation within each treatment group is "explained" by differences within clusters, and no variation is observed between clusters (that is, in the absence of clustering), the ICC=0.[3] Statistical power depends on the degree of clustering; the larger the ICC, the greater the reduction in statistical power. If ICC=0, a cluster randomised trial has the same statistical power as an individually randomised trial with the same number of participants; if ICC=1, the power is reduced to that of an individually randomised trial in which the sample size is equal to the number of clusters.

A second disadvantage of cluster randomisation can occur if the number of clusters is small. Despite proper randomisation, imbalance can occur in potentially confounding baseline factors that differ by chance across clusters. Such imbalance may require multivariable statistical adjustment, but adjustment cannot remove imbalance in factors that are unmeasured or imprecisely measured.

Although the advantages and limitations of cluster randomised trials are now well known, the consequences of clustered measurement have received far less attention. Observer level clustering of outcomes in individually randomised trials has been discussed,[4] but we recently encountered the "double jeopardy" that arises when clustered measurement occurs in trials.

SUMMARY POINTS

- Clustered measurement occurring in cluster randomised trials will reduce the precision of the results
- Random allocation of observers or a single observer will avoid clustered measurement but may be impossible for large, geographically dispersed clusters
- All studies should use standardised measurement techniques and ensure adequate training of observers
- Pilot studies and monitoring of initial data can identify difficulties in outcome measurement
- Despite these steps some systematic measurement differences may remain

Clustered measurement

In many studies, including both experimental (randomised) and observational studies, measurement of the outcome is naturally clustered. Measurement can be clustered because of either the observer (the person who measures the outcome) or the measuring instrument. The number of observers is often far lower than the number of participants in the study. For measurements susceptible to systematic (non-random) error, clustering among study participants measured by the same observer will occur if some observers tend to measure systematically higher or lower values than other observers, irrespective of the true value of the measurement. Such clustered measurement will lead to intracluster correlation, but the cluster is now defined as the group of individuals whose outcome is measured by the same observer.[4] This type of clustered measurement can also occur when several unstandardised measuring instruments are used for different participants, even with the same observer—for example, use of several inadequately calibrated sphygmomanometers for measuring blood pressure.

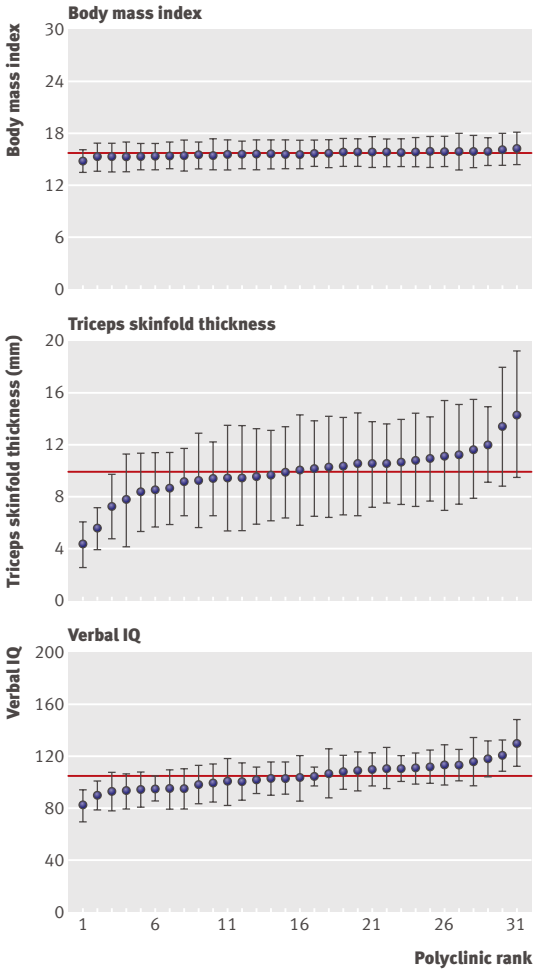
Combined clustering: "double jeopardy"

When measurements are clustered within the same groupings that serve as the units for cluster randomisa-

tion, however, a pernicious problem arises: the variation due to clustered measurement becomes inseparable from that due to clustered randomisation. Examples include a single teacher who obtains outcome measurements in a school where the school is the unit of randomisation or a clinician who is responsible for measuring outcome in a practice, clinic, or hospital where those sites are the units of randomisation. The conflation of clustered measurement with cluster randomisation can greatly increase the intraclass correlation and hence reduce statistical power. If the number of clusters is small, double clustering can also inflate or deflate true treatment differences if systematically higher measurements occur more frequently by chance in one treatment group than in the other.

Recent example

To show how measurement error and clustering can affect the precision of treatment effects in cluster randomised trials, we review our recent experience with the Promotion of Breastfeeding Intervention Trial, a cluster randomised trial of a breastfeeding promotion intervention carried out in the Republic of Belarus.[5] The units of randomisation were maternity hospitals and one affiliated polyclinic (outpatient clinic) per maternity hospital. These hospitals and clinics were spread across the country. The initial period of follow-up was for 12 months, with a subsequent follow-up at age 6.5 years for 13 889 (81.5%) of the 17 046 children originally randomised. The effects of the intervention on the 6.5 year outcomes have been reported.[6-10]



Mean (±1 SD) body mass index (top), triceps skinfold thickness (middle), and verbal IQ (bottom) in 31 participating polyclinics, in ascending order. Red horizontal lines depict the means of the 31 polyclinic means for each outcome

Here, contrast the results we obtained for three of these outcomes: body mass index (weight (kg)/(height (m)^2), triceps skinfold thickness, and verbal IQ score. The paediatricians were trained to measure all outcomes at a week long training session on a sample of school aged children living in a residential facility near Minsk. Each participating paediatrician was also given a training video (for the anthropometric measures) and detailed written instructions in Russian.[8] All anthropometric measurements were obtained in duplicate and averaged. Standard administration and scoring of the Wechsler Abbreviated Scale of Intelligence test was demonstrated by, and practised under the supervision of local child psychologists and psychiatrists with experience in IQ testing in children; during the training session, high interpaediatrician agreement was achieved on repeat testing of the same children.[10]

The figure shows the (crude) means of the three outcomes for each of the 31 clusters (polyclinics), in ascending order. The 31 means range from 14.7 to 16.2 for body mass index, 4.3 to 14.4 mm for triceps skinfold thickness, and 82 to 130 points for verbal IQ. The digital read out weight scale is the least susceptible to between clinic differences, and adequate attention to positioning the child and placing the horizontal stadiometer bar on the child's head can minimise systematic errors in measuring height. These features of measurement explain why mean body mass index does not vary greatly by polyclinic.

In contrast, the ranges in means for triceps skinfold thickness and verbal IQ were too wide to be explained by true geographic differences. It is not credible that average triceps skinfold thicknesses in 6.5 year old children would vary 3.5-fold among the 31 polyclinics (especially given the narrow observed range of body mass index) or that true average verbal IQ scores would vary by nearly 50 points. Instead, these differences are likely to reflect systematic measurement differences among the 31 polyclinics. Despite our efforts to standardise measurements across paediatricians and polyclinics, variability in technique for separating subcutaneous fat from muscle (for triceps skinfold thickness) and in acceptance of definitions of words and explanations of similarities between words (for verbal IQ) seems to have led to systematic differences between polyclinics.

The table shows the means in the experimental and control groups and the ICC for the same three outcome measurements. The ICC for body mass index was quite low, reflecting the consistency in measurement. The ICCs for triceps skinfold thickness and verbal IQ were both high, reflecting the large differences in means among the 31 polyclinics, although the ICC for triceps

skinfold was lower than for verbal IQ because of higher variation within polyclinics; the SD was about 40% of the mean for the triceps skinfold compared with 15% of the mean for verbal IQ. The mean values for body mass index and for triceps skinfold thickness were similar in the experimental and control groups, but because the ICC was much lower for body mass index the 95% confidence interval around the cluster adjusted difference in means was also much narrower. The cluster adjusted difference in mean verbal IQ scores was large (7.5 points higher in the experimental than in the control group), but because the ICC was high, the 95% CI was wide.

The effect of within polyclinic clustering on the precision (width of the confidence interval) of the estimated treatment differences can be shown by carrying out an intention to treat analysis without the cluster adjustment—that is, based on the individual as the unit of analysis. Such an analysis erroneously assumes that ICC=0. The estimated treatment differences are 0.1 (95% confidence interval 0.02 to 0.1) for body mass index (owing to rounding errors, this is larger than the crude difference), −0.1 (−0.2 to 0.1) mm for triceps skinfold thickness, and 10.0 (9.4 to 10.5) for verbal IQ. The confidence intervals are too narrow, providing overly precise estimates of the treatment effect, because they do not account for the clustered randomisation or measurement.

What can be done to minimise double clustering?

Some of the strategies we suggest for minimising double clustering can and should be incorporated into the design and conduct of all cluster randomised trials. Others, however, may be difficult or impossible to implement because of logistical obstacles.

One strategy is to randomly allocate observers across clusters. Such an approach may not be feasible, however, if observers and trial participants are geographically dispersed, as in our trial. Another potential solution is to use a single observer with proved measurement validity and precision to assess the outcome in all clusters. That approach is analogous to using a single, highly accurate laboratory to analyse blood or other biological samples obtained from multiple study sites. But in trials with large numbers of participants or wide geographical dispersion this may be difficult or impossible to achieve.

A third strategy is to standardise measurement techniques and ensure adequate training of observers. The trial's manual of procedures is an important training tool and reference guide, but for some types of measurement (such as triceps skinfold and verbal IQ in our study), systematic differences across clusters are likely to persist despite these efforts. A pilot study can identify difficulties in outcome measurement before starting the main trial. The pilot study can detect "outlier" observers and attempt to modify their behaviour, but this is unlikely to eliminate systematic differences for some types of measurement. Finally, initial data collection should always be monitored closely to identify observers who may require additional training and instruments that require repair or replacement. We incorporated this approach in our trial, and it should be feasible in all cluster randomised trials. It will, however, add to the costs and logistical difficulties of the trial when the clusters are numerous and geographically dispersed.

Results of intention to treat analysis for body mass index, triceps skinfold thickness, and verbal IQ at 6.5 year follow-up

Outcome	Mean (SD) value in experimental group	Mean (SD) value in control group	ICC	Mean (95% CI) cluster adjusted difference
Body mass index	15.6 (1.7)	15.6 (1.7)	0.03	0.1 (−0.2 to 0.3)
Triceps skinfold (mm)	9.9 (4.1)	10.0 (3.6)	0.18	−0.4 (−1.8 to 1.0)
Verbal IQ	108.7 (16.4)	98.7 (16.0)	0.31	7.5 (0.8 to 14.3)

ICC=Intraclass correlation coefficient.

Conclusion

We suspect that double clustering may have occurred more often than recognised in the past and could partly explain the negative results of some previous cluster randomised trials. Future CONSORT statements for cluster randomised trials[11] should recommend that reports contain text (or a table) summarising the distributions of the cluster means for each study outcome and describe design features (if any) used to reduce clustered measurement. Investigators should be aware of the potential for double clustering and implement study procedures that minimise its risk.

Contributors: MSK, SS, and RWP contributed to obtaining funding for this project and to the design, analysis, interpretation, and writing or revision of the manuscript. RMM and JACS had major roles in the interpretation of the analysis and in the writing and revision of the manuscript. MD performed the statistical analysis and contributed to its interpretation. MSK is guarantor.

Funding: Supported by a grant from the Canadian Institutes of Health Research. RWP is a career investigator (chercheur-boursier) of the Fonds de la recherche en santé du Québec. RMM was supported by grant number FOOD-DT-2005-007036 from the European Union's project on early nutrition programming: long-term efficacy and safety of trials.

Competing interests: None declared.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Donner A, Klar N. *Design and analysis of cluster randomisation trials in health research*. London: Arnold, 2000.
2 Ukoumunne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ. Methods for evaluating area-wide and organisation based interventions in health and health care: a systematic review. *Health Technol Assess* 1999;3:iii-92.
3 Kirkwood BR, Sterne JAC. *Essential medical statistics*. 2nd ed. Oxford: Blackwell Science, 2003.
4 Lee KJ, Thompson SG. Clustering by health professional in individually randomised trials. *BMJ* 2005;330:142-4.
5 Kramer MS, Chalmers B, Hodnett ED, Sevkovskaya Z, Dzikovich I, Shapiro S, et al. Promotion of breastfeeding intervention trial (PROBIT): a randomized trial in the Republic of Belarus. *JAMA* 2001;285:413-20.
6 Kramer MS, Matush L, Vanilovich I, Platt RW, Bogdanovich N, Sevkovskaya Z, et al. Does prolonged and exclusive breastfeeding reduce the risk of allergy and asthma? New evidence from a large randomised trial. *BMJ* 2007;335:815-20.
7 Kramer MS, Vanilovich I, Matush L, Bogdanovich N, Zhang X, Shishko G, et al. The effect of prolonged and exclusive breastfeeding on dental caries in early school-age children: new evidence from a large randomized trial. *Caries Res* 2007;41:484-8.
8 Kramer MS, Matush L, Vanilovich I, Platt RW, Bogdanovich N, Sevkovskaya Z, et al. Effects of prolonged and exclusive breastfeeding on child height, weight, adiposity, and blood pressure at age 6.5 years: new evidence from a large randomized trial. *Am J Clin Nutr* 2007;86:1717-21.
9 Kramer MS, Fombonne E, Igumnov S, Vanilovich I, Matush L, Mironova E, et al. Effects of prolonged and exclusive breastfeeding on child behavior and maternal adjustment: evidence from a large randomized trial. *Pediatrics* 2008;121:e435-440.
10 Kramer MS, Aboud F, Mironova E, Vanilovich I, Platt RW, Matush L, et al. Breastfeeding and child cognitive development: new evidence from a large randomized trial. *Arch Gen Psychiatry* 2008;65:578-84.
11 Campbell MK, Elbourne DR, Altman DG, for the CONSORT Group. CONSORT statement: extension to cluster randomised trials. *BMJ* 2004;328:702-8.

Accepted: 9 March 2009