Analysing cluster randomised trials with an assessment of outcome at baseline

Richard Hooper, Reader;[1] Andrew Forbes, Professor;[2] Karla Hemming, Senior Lecturer;[3]
Andrea Takeda, Systematic Reviewer;[4] Lee Beresford, Statistician[1]

[1] Centre for Primary Care & Public Health, Queen Mary University of London, London, UK
[2] Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Australia
[3] Institute of Applied Health Research, University of Birmingham, Birmingham
[4] UCL Institute of Health Informatics, University College London, London, UK

Corresponding author:

Dr Richard Hooper,
Centre for Primary Care & Public Health,
Queen Mary University of London,
Yvonne Carter Building,
58 Turner Street,
Whitechapel,
London E1 2AB

r.l.hooper@qmul.ac.uk

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, a worldwide licence to the Publishers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to i) publish, reproduce, distribute, display and store the Contribution, ii) translate the Contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or, abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) licence any third party to do any or all of the above.

Summary points

- Clinical trials that are randomised in clusters often include an assessment of participants' outcomes in a baseline period.

- The analysis of cluster randomised trials is more complex than for individually randomised trials. A variety of methods have been suggested to allow for baseline assessments of outcome.

- We recommend either an analysis of covariance approach which takes account of cluster differences at baseline, or an analysis which treats assessments at baseline and follow-up as longitudinal but which recognises that there will not be any systematic differences between the randomised groups at baseline.

- Simply comparing the difference between outcomes at baseline and follow-up in the two randomised groups – that is, calculating the difference of differences – is not the best approach, and can be misleading.

In a cluster randomised clinical trial all the participants who belong to the same "cluster" (*e.g.* a local community, school or general practice) are randomised to receive the same treatment. People from the same cluster tend to be more similar than those from different clusters, and the analysis must allow for this.[1] A common enhancement is to add an assessment of participants' outcomes in a baseline period, *i.e.* before randomisation. Even if different participants are assessed at baseline and follow-up the fact that they are sampled from the same cluster allows some control for cluster differences. But the analysis of cluster randomised trials is more complex than for individually randomised trials, and it is not obvious how best to adjust for baseline assessments. A variety of methods of analysis have been suggested: here we describe different approaches and their flaws, and make recommendations.

Trial design

The decision to randomise a trial in clusters is usually a pragmatic one: the intervention might be delivered at cluster level, for example, or there might otherwise be a risk of people in the same cluster sharing their treatments, and thus attenuating treatment effects. Three illustrative cluster randomised trials which include a baseline assessment of participants' outcomes are described in Boxes 1-3. Coventry and colleagues (Box 1), for example, used a cluster randomised design with general practices as clusters to look at the effectiveness of an integrated collaborative care model for people with depression and long term physical conditions.[2]

The sample size required for a cluster randomised trial is larger than for an individually randomised trial: how much larger depends on a parameter called the intracluster correlation – the correlation between the outcomes of two individuals from the same cluster.[3] The higher the intracluster correlation, the more heterogeneity there is between clusters, and the greater the advantage in controlling for cluster differences, for example with a baseline assessment.[4] Researchers might choose to assess the same individuals at baseline and follow-up (a cohort design) or to take different samples from the same cluster on the two occasions (repeated cross-sections), and this may again be a pragmatic decision. Costantini and colleagues (Box 2) studied the quality of end-of-life care in patients with cancer, leading them to a design in which they recruited different samples from each hospital ward at baseline and follow-up.[5]

Approaches to analysis

Difference of differences

One method for estimating the effect of an intervention in a trial with clusters randomised to intervention and control groups, and assessments at baseline and follow-up, is with a longitudinal repeated measures analysis that tests for a statistical interaction between group and time. This is sometimes called a difference of differences analysis because it evaluates how much the groups differ in terms of the difference between outcomes at baseline and follow-up. This was the approach used by He and colleagues to evaluate a school-based education programme aimed at reducing salt intake in children and their families (Box 3), for example.[6]

In individually randomised trials the shortcomings of a difference of differences approach (more usually known in this case as a change score analysis) are well understood.[7,8] If the baseline assessment is only poorly correlated with the follow-up then subtracting the baseline outcome just adds random noise to the signal we are trying to detect. Worse still, if the two groups differ markedly (by chance) at baseline and then level themselves out at follow-up we might be misled by the difference of differences into thinking that a change had been effected in one group but not the

other (regression to the mean).[9] The problem is that the analysis does not use what we know: before randomisation outcomes in the two groups should, on average, be the same.

Analysis of covariance

The method usually recommended for baseline adjustment in an individually randomised trial is analysis of covariance, or ANCOVA.[7] Adjustment in cluster randomised trials is more complex. We consider repeated cross-section and cohort designs in turn.

In a repeated cross-section design, where there are different individuals at baseline and follow-up, a simple way to deal with baseline assessments of outcomes is to bundle them up in each cluster as a mean or other aggregate measure, to form a cluster-level covariate. Outcomes at follow-up can then be analysed either at aggregate cluster level or as individual outcomes, in either case adjusting for the cluster-level baseline covariate. If analysing individual outcomes at follow-up then the analysis must also allow for clustering using mixed regression or generalised estimating equations (GEE), as with any cluster randomised trial.[1] This was the method used by Costantini and colleagues (Box 2):[5] each patient's quality of care score at follow-up was adjusted for the mean quality of care score found in that hospital ward at baseline.

In a cohort design, where the same individuals are assessed at baseline and follow-up, an obvious approach to ANCOVA is to analyse each individual's outcome at follow-up adjusted for that individual's outcome at baseline, with mixed regression or GEE to allow for differences between clusters. This ANCOVA approach with individual-level baseline adjustment was the method used by Coventry and colleagues (Box 1):[2] each participant's depression scale score at follow-up was adjusted for his or her score at baseline, thus allowing for participant differences. In a methodological paper that deserves more attention, however, Klar and Darlington found that even more precise results could be obtained by adjusting an individual's outcome at follow-up both for the individual's baseline assessment and for the baseline cluster mean.[10] In the Coventry example this would be achieved by adjusting also for the mean of the baseline depression scores of all patients from the same practice. The baseline cluster mean captures more information about cluster differences than the individual baseline assessment.

Constrained baseline analysis

In individually randomised trials an alternative to ANCOVA is to treat outcomes collected at baseline and follow-up as longitudinal, and to use a repeated measures analysis to estimate the effect of the intervention being "switched on" in one of the randomised groups on the second of these occasions.[11] This is sometimes referred to as a constrained baseline analysis because, unlike a difference of differences analysis, it assumes there is no systematic difference between the groups at baseline. In cluster randomised trials this is a special case of the approach to analysis recommended more generally for stepped wedge trials (cluster randomised trials where outcomes are assessed at multiple time-points, and different clusters switch over to the intervention at different times).[12] Two ways of doing this have been described in the literature: one assumes that the correlation between two people from the same cluster is the same whether they are sampled in the same period or a different period,[13] while the other allows the correlation to be weaker between different periods.[14–16] The distinction is technical but important. A study using The Health Improvement Network (THIN) general practice database found that, for the health outcomes investigated (HbA1c, systolic and diastolic blood pressure, body mass index, total cholesterol, and HDL cholesterol), correlations between individuals from the same practice were between 12% and 51% smaller when those individuals were sampled from different 15-month periods,[17] motivating the use of the second, more flexible model. A constrained baseline analysis that lacks this flexibility

in the correlation structure is known to over-state the precision of the treatment effect, potentially leading to false-positive findings.[15]

In the cluster randomised case a constrained baseline analysis might be expected to produce similar results to ANCOVA, as in the individually randomised case.[11] The method is extremely flexible, is available in cohort or repeated cross-section forms, and allows an analysis based on individual-level data, with no aggregation needed either at baseline or at follow-up.

Some benefits of aggregating outcomes by cluster

When a cluster randomised trial involving relatively few clusters – fewer than 40, say – is analysed with mixed regression or GEE there is known to be an increased risk of a false positive finding (an inflated Type I error rate) unless an appropriate correction is made.[18] Corrections such as that of Kenward and Roger are increasingly accessible, and should be considered in such cases.[19] For repeated cross-section designs, one benefit of performing an ANCOVA entirely at the cluster level – aggregating follow-up outcomes by cluster and adjusting for the aggregate baseline outcome – is that a mixed regression or GEE approach is unnecessary. This greatly simplifies the analysis and keeps the risk of false positive findings under control. But could this also help with cohort designs? Theory shows that if the correlation between the baseline cluster mean and the mean at follow-up in the same cluster is known, it is immaterial whether it was the same participants who were assessed on each occasion or different participants: the precision of the treatment effect estimate will be the same.[14] This suggests that if we treated a cohort design as if it *were* a repeated cross-section design, and adjusted purely at the aggregate, cluster level rather than the individual level, the analysis would perform just as well. Indeed, in some situations cluster-level and individual-level analyses of cohort designs do give identical results (see online supplement).

Individual-level baseline adjustment also runs into difficulties if some participants in a cohort design have missing baseline assessments. ANCOVA with individual-level baseline adjustment requires us in this case either to impute the missing baseline assessments or to exclude those individuals from the analysis. Adjusting only for a baseline cluster mean offers a straightforward analysis of all available data without the need to impute or exclude data, but there are still difficulties in this case: participants who are assessed at baseline but have a missing follow-up assessment still warrant individual attention, as the baseline could offer a valuable clue to the unobserved outcome (and reasons for dropping out).[20] As observed above, a constrained baseline analysis offers an alternative using all available data, but at the individual level. Note finally that in some trials having a missing baseline assessment will mean a participant is not eligible to be followed up.

Implementation

The online supplement provides tutorials showing how to implement the analyses described above using the Stata package (Stata Corporation, College Station TX, USA). The supplement also includes results of large-scale simulations of trials in typical scenarios which illustrate some of the performance issues outlined. A constrained baseline analysis with a realistic correlation structure performed consistently well for both cohort and repeated cross-section designs, as did (more surprisingly) an ANCOVA performed entirely at the cluster level. A constrained baseline analysis with less flexible correlation structure risked an inflated Type I error rate in some scenarios, while a difference of differences analysis and ANCOVA with purely individual-level baseline adjustment did not always achieve the statistical power that was expected.

Further work, including simulation studies, is needed to quantify the performance of different methods in a broader range of circumstances than we have been able to consider here. We only

looked at balanced situations where equal numbers of participants are sampled from each cluster at each assessment. (Klar and Darlington also provided simulations of unbalanced situations.)[10] Methods outlined in this article can be applied to unbalanced designs, though lack of balance introduces further subtleties: aggregating outcomes at cluster level, for example, may not be the most efficient way to weight the contribution of clusters of different size. We have emphasised the importance of getting the right model for the correlation structure, but there may be complexities beyond those we have considered: outcomes at baseline and follow-up may correlate in different ways in the intervention and control groups, for example.

Recommendations

The analysis of a cluster randomised trial with a baseline assessment of outcome is not as straightforward as it might seem, but the advice is similar for cohort and for cross-sectional designs. ANCOVA should adjust for the baseline cluster mean, even in a cohort design where individual-level baseline adjustment is also possible. A good, all-round alternative to ANCOVA is a constrained baseline analysis with a suitably flexible model for the correlation between individuals from the same cluster. We do not recommend a difference of differences analysis for a cluster randomised trial. Any analysis using mixed regression or generalised estimating equations has an increased risk of a false positive finding when there are relatively few clusters, so analysts should apply a correction in this case if one is available, or consider aggregating results at cluster level.

Footnotes

Contributors: RH conceived the article and led the writing of the paper. He is the guarantor. RH, AT and LB performed simulations and conducted scoping reviews to identify example studies. AF and KH contributed further ideas. All authors contributed to the final version of the manuscript.

Competing interests: We have read and understood BMJ policy on declaration of interests and declare that we have no competing interests.

References

1. Donner, A., Klar, N., Design and Analysis of Cluster Randomization Trials in Health Research. 2000, London: Arnold.
2. Coventry, P., et al., Integrated primary care for patients with mental and physical multimorbidity: cluster randomised controlled trial of collaborative care for patients with depression comorbid with diabetes or cardiovascular disease. BMJ 2015;350:h638.
3. Kerry, S.M., Bland J.M., The intracluster correlation coefficient in cluster randomisation. BMJ 1998;316(7142):1455.
4. Teerenstra, S., et al., A simple sample size formula for analysis of covariance in cluster randomized trials. Stat Med 2012;31(20):2169-2178.
5. Costantini, M., et al., Liverpool Care Pathway for patients with cancer in hospital: a cluster randomised trial. Lancet 2014;383(9913):226–237.
6. He, F.J., et al., School based education programme to reduce salt intake in children and their families (School-EduSalt): cluster randomised controlled trial. BMJ 2015;350:h770.
7. Vickers, A.J., Altman, D.G., Analysing controlled trials with baseline and follow-up measurements. BMJ 2001; 323(7321):1123–1124.
8. Senn, S., Change from baseline and analysis of covariance revisited. Stat Med 2006;25(24):4334-4344.
9. Morton, V., Torgerson, D.J., Effect of regression to the mean on decision making in health care. BMJ 2003;326(7398):1083-1084.
10. Klar, N., Darlington, G., Methods for modelling change in cluster randomization trials. Stat Med 2004;23:2341-2357.
11. Coffman, C.J., Edelman, D., Woolson, R.F., To condition or not condition? Analysing 'change' in longitudinal randomised controlled trials. BMJ Open 2016;6(12):e013096.
12. Hemming, K., et al., The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ 2015;350:h391.
13. Hussey, M.A., Hughes, J.P., Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials 2007;28(2):182-91.
14. Hooper, R., Bourke, L., Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. BMJ 2015;350:h2925.
15. Hooper, R., et al., Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Stat Med 2016;35(26):4718-4728.
16. Girling, A.J., Hemming, K., Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. Stat Med 2016;35(13):2149-2166.
17. Martin, J., Girling, A., Nirantharakumar, K., Ryan, R., Marshall, T., Hemming, K., Intra-cluster and inter-period correlation coefficients for cross-sectional cluster randomised controlled trials for type-2 diabetes in UK primary care. Trials 2016;17:402
18. Kahan, B.C., et al., Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. Trials 2016;17:438
19. Kenward, M.G., Roger, J.H., Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 1997;53(3):983-997.
20. White I.R., Thompson S.G., Adjusting for partially missing baseline measurements in randomized trials. Stat Med 2005; 24:993-1007.

Box 1. Integrated primary care for patients with mental and physical multimorbidity: cluster randomised controlled trial of collaborative care for patients with depression comorbid with diabetes or cardiovascular disease. Coventry *et al*, *BMJ* 2015;350:h638

Objective: To test the effectiveness of an integrated collaborative care model for people with depression and long term physical conditions.

Primary Outcome: Symptoms of depression on the self-reported symptom checklist-13 depression scale (SCL-D13).

Planned sample Size: 15 general practices (clusters) per arm and 15 patients per cluster (n=450) to detect a difference between groups equivalent to a standardised effect size of 0.4, with 80% power (α=0.05; intraclass correlation coefficient 0.06), allowing for 20% attrition.

Design & Analysis: Patients registered at a participating practice who had a record of diabetes and/or coronary heart disease were screened for depressive symptoms over the telephone, and face-to-face two weeks later, to determine eligibility. Participants' depression scale scores were collected at baseline and again after 4 months follow-up (a cohort design). Outcomes at follow-up were compared between the intervention and control arms. Analysis was conducted at the individual level, adjusting for clustering and for baseline score.

Box 2. Liverpool Care Pathway for patients with cancer in hospital: a cluster randomised trial. Costantini *et al*, *Lancet* 2014;383:226–237

Objective: To assess the effectiveness of the Liverpool Care Pathway translated into an Italian context (LCP-I) in improving the quality of end-of-life care for patients with cancer in hospitals and for their families.

Primary Outcome: Mean score on the overall quality of care toolkit scale.

Planned sample Size: 10 hospital wards (clusters) per arm and 20 patients per cluster (n=400) to detect an absolute increase of 10 points on the toolkit scale, with 80% power (α=0.05; intraclass correlation coefficient 0.1).

Design & Analysis: In each ward, all patients who died in the 3 months before randomisation and in the 6 months after the conclusion of the LCP-I programme were identified (a repeated cross-section design) and their quality of end-of-life care was assessed. Outcomes at follow-up were compared between the intervention and control arms. Analysis was at the individual level, adjusting for clustering and for the mean baseline assessment of outcome in that ward.

Box 3. School based education programme to reduce salt intake in children and their families (School-EduSalt): cluster randomised controlled trial. He *et al*, *BMJ* 2015;350:h770

Objective: To determine whether an education programme targeted at schoolchildren could lower salt intake in children and their families.

Primary Outcome: Salt intake (measured as urinary sodium excretion, averaged over two consecutive 24-hour collections).

Planned sample size: 12 schools (clusters) per arm and 10 children per cluster (n=240) to detect a difference in salt intake of 1.0g a day, with 90% power (α=0.05; intraclass correlation coefficient 0.01).

Design & Analysis: One class was selected from each school, and 10 children were randomly selected from each class. Each child's urinary sodium was measured at baseline and again after 6 months follow-up (a cohort design). Analysis was at the individual level, and compared the change from baseline to follow-up between the intervention and control arms.