Analysis of cluster randomized trials in primary care: a practical approach

Marion K Campbell\textsuperscript{a}, Jill Mollison\textsuperscript{b}, Nick Steen\textsuperscript{c}, Jeremy M Grimshaw\textsuperscript{a} and Martin Eccles\textsuperscript{c}


\textbf{Background.} Cluster randomized trials increasingly are being used in health services research and in primary care, yet the majority of these trials do not account appropriately for the clustering in their analysis.

\textbf{Objectives.} We review the main implications of adopting a cluster randomized design in primary care and highlight the practical application of appropriate analytical techniques.

\textbf{Methods.} The application of different analytical techniques is demonstrated through the use of empirical data from a primary care-based case study.

\textbf{Conclusion.} Inappropriate analysis of cluster trials can lead to the presentation of inaccurate results and hence potentially misleading conclusions. We have demonstrated that adjustment for clustering can be applied to real-life data and we encourage more routine adoption of appropriate analytical techniques.

\textbf{Keywords.} Analysis approaches, cluster randomized trials, primary care.

Introduction

The majority of randomized controlled trials (RCTs) in primary care to date have randomized individual patients to different interventions. As with other areas of healthcare, however, the use of the cluster randomized trial, where groups of patients (such as practices) rather than individual patients are randomized, is increasing.

It is widely recognized that the cluster randomized trial is more appropriate for the evaluation of a number of interventions such as family-based dietary interventions, community-based health promotion initiatives or educational interventions targeted at the health professional rather than the individual patient. The cluster randomized trial also provides protection against contamination across trial groups when trial patients are managed within the same setting.

Despite the growing literature on the appropriate methods for the design and analysis of cluster randomized trials,\textsuperscript{1,2} those trials that account appropriately for clustering remain in the minority.\textsuperscript{3,4,5} Inappropriate analysis and poor reporting of such trials can lead to the presentation of inaccurate results and hence potentially misleading conclusions. The majority of the publications on cluster methodology have been presented in the statistical and epidemiological literature, however, and it may be that, despite a few publications specifically within the primary care field,\textsuperscript{6,7,8} the transfer of the wider literature to more generic researchers has been slow.

The aim of this article, therefore, is to review briefly the main implications of adopting a cluster randomized design and to highlight the practical application of appropriate analytical techniques through the use of empirical data from a primary care-based case study.

Cluster randomized trials

The primary implication of adopting a cluster randomized design is that patients within any one cluster (such as a practice) are often more likely to respond in a similar manner, and thus can no longer be assumed to act independently. This lack of independence in turn leads to a loss of statistical power in comparison with a patient randomized trial. A statistical measure of this intracluster dependence is known as the ‘intracluster
correlation coefficient’ (ICC) and, to achieve the equivalent power of a patient randomized trial, standard sample size calculations (for a completely randomized design) need to be inflated by a factor:

\[ 1 + (n - 1) \rho \]

where \( n \) is the average cluster size and \( \rho \) is an estimate of the ICC. This inflation factor is often referred to as the ‘design effect’.\(^1\)

The ICC takes a value of between 0 and 1 and would be high if, for example, the management of patients within practices was very consistent; but, there was wide variation across different practices. A recent study of UK data sets relevant to implementation research showed that in primary care settings, the ICCs for process variables appear to be of an order of magnitude higher than those for outcome variables (estimates for process variables from primary care were of the order of 0.05–0.15), whereas ICCs for outcome variables were generally lower than 0.05.\(^9\) As both the ICC and the cluster size influence the calculation, as shown by the equation for the design effect, even small values of ICC can have a substantial impact on power.

The analysis of cluster randomized trials must also take into account the clustered nature of the data. Standard statistical techniques are no longer appropriate, unless an aggregated analysis is performed at the level of the cluster (see below), as they require data to be independent. If the clustering effect is ignored, many authors have highlighted that \( P \)-values will be artificially extreme, and confidence intervals will be over-narrow, increasing the chances of spuriously significant findings and misleading conclusions.\(^5,10\)

Analysis of cluster randomized trials

There are two main approaches to the analysis of cluster randomized trials: analysis at the cluster level or analysis at the patient level.

Traditionally, analysis has been focused at the cluster level; however, recent advances in statistics have led to the development of techniques which can incorporate the patient level data. Within each approach, simple analyses such as \( t \)-tests or more complex approaches such as regression analyses may be undertaken. Both allow the effect of the intervention to be tested; however, only complex analyses allow adjustment for potential covariates, such as baseline performance.

Analytical methods for each approach are described below, and worked examples using data from a particular primary care-based evaluation are presented. It should be noted that these methods are appropriate for completely randomized designs and \( P \)-values are quoted to increased levels of accuracy to highlight the differences between methods. Readers should refer to more detailed texts, e.g. Murray,\(^2\) for discussion of the appropriate methods to analyse stratified or matched designs.

Case study: the urological referral guidelines evaluation (URGE) study

The URGE study aimed to evaluate the effectiveness of a guideline-based open access ‘fast-track’ investigation service for two common urological problems, benign prostatic hyperplasia (BPH) and microscopic haematuria. General practices were allocated randomly to two groups; one group received guidelines for the appropriate referral of BPH patients for the open access ‘fast-track’ system whilst the other group acted as a control for BPH patients (but did receive guidelines for microscopic haematuria).

Data were collected on two cohorts of patients, one referred before (an indicator of baseline performance) and another referred after the introduction of the fast-track service. Data were collected on pre-referral general practice management, hospital and general practice follow referral, and patient outcome.

For the purposes of this article, we focus on the evaluation of the effectiveness of the intervention for BPH patients only. Data for a single outcome are used: waiting time from the date of patient referral to first appointment at hospital. Waiting time was measured in days and was found to have a skewed distribution that was log transformed to normality. Therefore, geometric means are quoted throughout; the effect sizes and the corresponding 95% confidence intervals (CIs) relate to the ratio of mean waiting time in the intervention group compared with the control group. Data were available on 513 patients (211 before and 312 after the introduction of the fast-track service) referred from 54 general practices from the North East of Scotland.

Cluster level analysis

The traditional approach to the analysis of cluster randomized trials has been to calculate a summary measure for each cluster, such as a cluster mean or proportion. Because each cluster then provides only one data point, the data can be considered to be independent, allowing standard statistical tests to be used.

For example, within the URGE trial, the mean waiting times post-intervention for each general practice could be calculated (when different patients are included pre- and post-, only post data comparisons can be made using simple analyses) (see Table 1). The overall group means can then be compared using a standard \( t \)-test resulting in a significance of \( t_{48} = 3.99, P = 0.0003 \). This results in an effect size of 0.65 (95% CI: 0.53–0.81); in other words, the waiting time was on average 35% less in the guideline group (Table 2). When the size of the clusters varies widely, it is preferable to carry out a weighted \( t \)-test,
adjustments can now be made to simple statistical tests to account for the clustering effect. For example, test statistics based on chi-squared or $F$-tests should be divided by the design effect (as described earlier), while test statistics based on the $t$-test or the $z$-test should be divided by the square root of the design effect. Adjustments for these and other tests such as non-parametric tests are discussed by Donner and Klar.

In the URGE study, the mean waiting time per patient post-intervention in the guideline group was 39.4 days and 60.6 days in the control group. If the clustering effect had been ignored and a standard $t$-test performed, the analysis would have resulted in a $t$-value for the difference between groups of 5.11 (with a highly significant $P$-value of 0.000001 based on 310 degrees of freedom), and the resulting effect size would have been 0.65 (95% CI, 0.55–0.77) (Table 2).

The design effect for the time to first appointment outcome within the URGE trial was 1.56; hence the revised $t$-value adjusting for clustering is calculated:

$$t_{\text{value}} = \frac{5.11}{\sqrt{(1.56)}} = 4.09$$

resulting in a revised significance level of 0.000006. The 95% confidence interval can also be adjusted for clustering. The revised 95% confidence interval is 0.52–0.80.

Despite a highly significant difference in waiting times between the groups, this example illustrates the impact of clustering on the significance of trial results. If clustering had been ignored, the analysis would have returned a spuriously low $P$-value and overly narrow confidence intervals, over- emphasising the impact of the intervention.

Similarly, there have been advances in the development and use of new modelling techniques to incorporate patient level data such as mixed linear models, hierarchical linear modelling and generalized estimating equations. These modelling techniques allow the inherent correlation within clusters to be modelled explicitly, and thus a ‘correct’ model can be obtained.

The aim of statistical modelling is to identify the main factors that explain variation in the outcome. In the URGE study, factors other than the intervention might also explain variation in the waiting time, e.g. patient and practice characteristics. When analysing guideline implementation trials, such as the URGE study, the primary aim of modelling is to adjust for the effect of such covariates before the effect of the intervention is tested rather than to maximize the proportion of variation explained.

An analysis plan or strategy should be developed before any analysis is undertaken to ensure that the modelling is hypothesis-led rather than data-driven. The a priori model-fitting analysis strategy should identify:

- the covariates which are to be considered for inclusion in any modelling approach to analysis.

### Table 1: Post-intervention mean waiting times* (days) per practice

<table>
<thead>
<tr>
<th>Practice</th>
<th>Intervention</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practice A</td>
<td>43.4</td>
<td>0.0</td>
</tr>
<tr>
<td>Practice B</td>
<td>61.6</td>
<td>0.0</td>
</tr>
<tr>
<td>Practice C</td>
<td>0.0</td>
<td>83.9</td>
</tr>
<tr>
<td>Practice D</td>
<td>0.0</td>
<td>68.7</td>
</tr>
<tr>
<td>Overall mean</td>
<td>39.4</td>
<td>60.6</td>
</tr>
</tbody>
</table>

* Waiting times were log transformed.

### Table 2: Comparison of waiting times* between intervention and control groups

<table>
<thead>
<tr>
<th>Statistical test</th>
<th>Test statistic</th>
<th>$P$-value</th>
<th>Effect size</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregated analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t$-test</td>
<td>3.99</td>
<td>0.0003</td>
<td>0.65</td>
<td>0.53, 0.81</td>
</tr>
<tr>
<td>Weighted $t$-test</td>
<td>4.72</td>
<td>0.0003</td>
<td>0.65</td>
<td>0.54, 0.78</td>
</tr>
<tr>
<td>Individual patient</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted $t$-test</td>
<td>5.11</td>
<td>0.000001</td>
<td>0.65</td>
<td>0.55, 0.77</td>
</tr>
<tr>
<td>Adjusted $t$-test</td>
<td>4.09</td>
<td>0.000006</td>
<td>0.65</td>
<td>0.52, 0.80</td>
</tr>
<tr>
<td>Multilevel modelling</td>
<td>4.08</td>
<td>0.0001</td>
<td>0.66</td>
<td>0.54, 0.81</td>
</tr>
<tr>
<td>Multilevel modelling$^b$</td>
<td>2.71</td>
<td>0.01</td>
<td>0.70</td>
<td>0.55, 0.91</td>
</tr>
</tbody>
</table>

* Waiting times were log transformed.
$^b$ The analysis was conducted on all patients (pre- and post-intervention cohorts) and the model contained a correction for baseline, intervention and intervention × phase interaction.

Using cluster sizes as the weights. This weighted analysis returns an effect size of 0.65 (95% CI: 0.54–0.78), with a significance of $F_{38} = 4.72, P = 0.00003$.

Standard statistical techniques such as multiple regression can also be used when data have been summarized at a cluster level. These analyses, however, can only adjust for cluster level covariates directly, but can incorporate patient level covariates through a two-stage process.

Whilst these cluster level approaches overcome the problem of the non-independence of the data, they are in general not statistically efficient (except in the particular case of the analysis of continuous outcomes when there is no variation in cluster size).

**Patient level analysis**

Recent developments in the statistical field now allow all the patient level data to be utilized, whilst accounting for the intracluster correlation, thus increasing the statistical power of the analysis.
• the order in which confounding variables are to be considered for inclusion in the model with the intervention variable fitted last (or an ‘intervention × phase’ interaction if pre- and post-measurements have been taken).14

An example of a model-fitting analysis strategy which could have been used for the URGE data is displayed in Figure 1.

Order of model fit:

- design variables eg phase (pre/post)
- cluster/individual level covariates eg practice size
- intervention eg guideline/no guideline
- intervention × phase interaction

1 only if pre- and post- measurements have been taken

**Figure 1 Example of model-fitting strategy**

Multilevel modelling was undertaken for the URGE study using the software package MLWin, developed by the Institute of Education in London (Table 2). As outlined above, an *a priori* model-fitting analysis strategy was developed which identified the order in which covariates were to be included in the model. Only after all covariates were included in the model was the effect of the ‘intervention × phase’ interaction examined. After adjustment for the pre-identified covariates, the interaction remained significant. The effect size estimated from the multilevel model was 0.70 (95% CI: 0.55–0.91). The resulting t-ratio was $t = 2.71$, $P = 0.01$. This indicates that when all the data are used in the analysis, the waiting time was on average 30% less in the guideline group compared with the control group (Table 2).

An in-depth discussion of all the available modelling methods is beyond the scope of this article. Researchers should refer to specific texts such as Murray2 for a general introduction to possible methods, or to Kreft and de Leeuw15 for discussion of multilevel models. Similarly, a range of statistical software packages are available for the analysis of clustered data sets. A discussion of the more common packages can be found on the multilevel modelling web site: http://www.ioe.ac.uk/multilevel/. For a discussion of generalized estimating equations, readers should refer to Burton et al.16

These modelling techniques adjust well for clustering and allow adjustment for both cluster level and patient level covariates. These types of analyses are more computationally intensive, however, and require greater statistical expertise both in the execution of the procedures and in the interpretation of the results.

**Discussion**

With the increasing popularity of the cluster randomized trial, it is important that researchers be aware of the implications of adopting such a design. Cluster RCTs are more complex to undertake than patient randomized trials in that they require increased sample sizes, with associated recruitment issues, and the analysis of these trials is not so straightforward. Cluster trials are the gold standard design for some interventions, however, and it is important that researchers have the information to design and analyse them appropriately.

The majority of the methodological developments in the field of cluster RCTs have been published in the more specialized fields of statistics and epidemiology. While statisticians and epidemiologists have the greatest need for this information, it is important that generic health services and primary care researchers have access to the principal findings of this research if they are to plan and conduct cluster RCTs appropriately.

We have outlined here the primary implications of adopting a cluster design and have highlighted that methods, some of which are easy to apply, do exist whereby cluster RCTs can be analysed appropriately. While we have identified a range of plausible methods, however, the choice of method and its actual implementation and interpretation should not be considered lightly, and expert statistical advice should be sought early in the planning of such studies. It should also be noted that the analysis options described are only appropriate for a completely randomized trial design with a continuous outcome. While the general approach to the analysis of binary data will be similar, whether cluster or individual level, the specifics of the analysis will be different. Similarly, more complex designs, such as stratified or matched designs, will require more sophisticated analysis strategies.2

When planning a cluster RCT, it is important to think about the analysis strategy at the design phase as the choice of analysis approach may impact on the design of the trial. For example, to ensure that robust multilevel modelling can be undertaken, it is necessary that both sufficient clusters are recruited to the study and sufficient number of patients are available per cluster.17

Considerable debate surrounds the choice of unit of analysis in cluster randomized trials.2,18 Some authors stress that analysis should only be undertaken at the level of randomization; for example if a trial is randomized by practice, it should only be analysed by practice. Murray2 argues that this emphasis on the unit of analysis may be misplaced and that attention should be focused rather on the appropriate specification of the model for the
analysis, where the model selected should be well matched to the underlying structure of the data.

In conclusion, this study has demonstrated that adjustment for clustering can be applied to real-life data in a relatively straightforward manner, if advice and relevant software are available, and we encourage more routine adoption of appropriate analytical techniques.

Acknowledgements

We would like to thank Ruth Thomas, lead researcher on the URGE study, for access to the study data. The research was funded by the Changing Professional Practice in Europe Group (a concerted action funded from the EU BIOMED-2 programme). The Health Services Research Unit is funded by the Chief Scientist Office of the Scottish Executive Health Department and is part of the MRC Health Services Research Collaboration. The views expressed are not necessarily those of the funding bodies.

References