Choosing appropriate analysis methods for cluster randomised cross-over trials with a binary outcome

Katy E. Morgan*,[1], Andrew B. Forbes[2], Ruth H. Keogh[1], Vipul Jairath[3] and Brennan C. Kahan[4]

Abstract

In cluster randomised cross-over (CRXO) trials, clusters receive multiple treatments in a randomised sequence over time. In such trials there is usually correlation between patients in the same cluster. In addition, within a cluster, patients in the same period may be more similar to each other than to patients in other periods. We demonstrate that it is necessary to account for these correlations in the analysis to obtain correct Type I error rates. We then use simulation to compare different methods of analysing a binary outcome from a two-period CRXO design. Our simulations demonstrated that hierarchical models without random effects for period-within-cluster, which do not account for any extra within-period correlation, performed poorly with greatly inflated Type I errors in many scenarios. In scenarios where extra within-period correlation was present, a hierarchical model with random effects for cluster and period-within-cluster only had correct Type I errors when there were large numbers of clusters; with small numbers of clusters the error rate was inflated. We also found that generalised estimating equations did not give correct error rates in any scenarios considered. An unweighted cluster-level summary regression performed best overall, maintaining an error rate close to 5% for all scenarios, although it lost power when extra within-period correlation was present, especially for small numbers of clusters. Results from our simulation study show that it is important to model both levels of clustering in CRXO trials, and that any extra within-period correlation should be accounted for.

1 Introduction

Cluster randomised trials are used in a number of situations including when an intervention is aimed at the cluster level, when a parallel group trial would be unfeasible, or for logistical reasons. However, cluster randomisation can lead to a substantial reduction in power compared to an individually randomised trial. Power in a cluster randomised trial can be increased by adding more clusters, or by recruiting more people from each cluster. However, there is a limit on the amount of power that can be gained using the latter approach [1], and the ability to increase the number of clusters in a trial may be limited by financial or logistical constraints. An alternative method that may increase the power of a cluster randomised trial is to add a cross-over element [2, 3]. Cluster randomised cross-over (CRXO) trials use a design in which

*Email: katy.morgan@lshtm.ac.uk
[1]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK
[2]School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia
[3]Department of Medicine, London Health Sciences Network & Western University, London, ON, Canada
[4]Pragmatic Clinical Trials Unit, Queen Mary University of London, London, UK
Accepted for publication by Statistics in Medicine

clusters receive multiple treatments in a random order. For example, in a CRXO trial with two treatments and two periods, clusters would be randomly assigned to receive either treatment A followed by treatment B, or B and then A. Including a cross-over element in a trial with cluster randomisation allows each cluster to act as its own control, which may reduce the number of clusters or patients needed to achieve the desired power [2, 3]. The CRXO design may also help to counteract imbalances in baseline patient characteristics that can occur in cluster randomised trials with a small number of clusters [3].

Despite the potential advantages of a CRXO study design, the cross-over element leads to a more complicated data structure, which in turn leads to a more complex analysis. If the analysis is not handled appropriately it could lead to incorrect or misleading results. However, a systematic review of CRXO trials [4, 5] deemed that only 10% (14/139) of analyses used potentially appropriate methods that accounted for both the cluster and cross-over components of the design. To date, much of the research regarding methods of analysis for CRXO trials has focused on continuous outcomes [6] or case studies of individual CRXO trials [7]. Forbes *et al.* [8] have considered the analysis of a binary outcome from a CRXO trial, but with an emphasis on cluster-level summary methods, in which cluster-level summaries are modelled instead of individual-level data. They also performed a limited assessment of individual-level models using generalised estimating equations (GEEs) with an identity link. It is currently unclear which individual-level methods of analysis are most appropriate for binary outcomes from CRXO trials, and in particular whether hierarchical models might be a useful approach.

In this paper we use simulation to compare methods of estimating a treatment effect for a binary outcome. We consider both models that estimate a treatment odds ratio (OR), including hierarchical models and GEEs, and also cluster-level summary methods of analysis that estimate the treatment effect as a difference in proportions. Our simulation study incorporates a wide range of scenarios and uses a standard statistical software package, Stata 13 [9], to implement the methods. We only consider a CRXO design with two time periods with a different set of patients in each period.

We start by outlining the structure of the data that are obtained from such a trial design in Section 2. In Sections 3 and 4 we outline the methods of analysis examined, and describe the structure of our simulation study. In Sections 5-7 we present results from our simulation study. We discuss the application of the CRXO design to a particular trial, TRIGGER2, in Section 8. Discussions and conclusions are given in Sections 9 and 10.

2 The structure of CRXO data

In cluster randomised trials outcomes from individuals in the same cluster are frequently more similar to each other than they are to outcomes from individuals in different clusters. This correlation between outcomes in the same cluster, called the intra-cluster correlation coefficient (ICC), violates the usual assumption that all patients are independent, and therefore requires that the clustering be taken into account in the analysis [1, 10–12].

CRXO trials have a more complicated structure. In addition to outcomes being correlated within clusters, there may be clustering within the same cluster-period (any given period within a single cluster): within each cluster, outcomes in one period may be more similar to each other than they are to outcomes in another period. Equivalently, there are two potential sources of variation — between clusters, and between periods within a cluster. For example, this may

occur in a trial where the members of clinical staff differ between periods. Different teams of clinical staff may have differing approaches to concomitant care that may in turn produce differing health outcomes of their patients. Two ICCs are therefore required to describe the data. In this paper we use one ICC to represent the correlation between two outcomes in the same cluster-period, and one ICC to represent the extra correlation between two outcomes in the same cluster-period compared with outcomes in different periods.

In Figure 1 we show two different ways of modelling how the mean outcome changes in each cluster and period for a simple CRXO trial with three clusters and two periods. The mean outcome could be the proportion of successful events, or the odds of a successful event, in the case of a binary outcome. Here, we use hierarchical models to model the log odds of a successful event. These models are discussed further in Section 3.1. We consider these models in the absence of a treatment effect for clarity. The first model, shown in panel A, assumes the mean outcome varies across clusters, and across time periods; however, the variation over time is assumed to be the same for all clusters:

$$\text{logit}(\Pr\{Y_{ijk} = 1|c_i\}) = \mu + \beta_\pi \pi_j + c_i \,, \tag{1}$$

where Y_{ijk} is the outcome for person k in period j of cluster i, $Y_{ijk} = 1$ is considered to be a "successful" event, μ is an intercept, c_i is a cluster effect that could be modelled either as a fixed or a random effect, β_π is a fixed period effect and π_j is an indicator variable that is 1 for the second period and 0 for the first. This model assumes that all observations in a cluster are equally correlated regardless of whether they belong to the same period, once the fixed effect of period has been taken into account.

The second model (panel B) assumes that the mean outcome varies across clusters and over time, and that the variation across time periods is different for each cluster:

$$\text{logit}(\Pr\{Y_{ijk} = 1|c_i, p_{ij}\}) = \mu + \beta_\pi \pi_j + c_i + p_{ij} \,, \tag{2}$$

where p_{ij} is a normally distributed random effect for cluster-period (*i.e.* a random effect for cluster-by-period interaction) with mean 0 and variance component σ_p^2. This model assumes that all observations within a cluster are correlated, and that within a cluster two observations in the same period are more correlated than two observations from different periods.

2.1 Impact of ignoring extra correlation within cluster-periods

The model in Panel A of Figure 1 is sometimes used to analyse CRXO trials [5]. These models assume that outcomes from all patients within a cluster are equally correlated, regardless of which period they are in. If this assumption is incorrect, and patient outcomes in the same period are more similar to each other than to outcomes in other periods, then cluster-period will be a source of non-ignorable clustering, a concept that is discussed in Ref. [13]. Briefly, a source of clustering is non-ignorable if both the outcomes and treatment assignments within the cluster are correlated. Treatment assignments within a cluster-period are correlated, as all patients in the cluster-period receive the same treatment; hence the correlation is equal to one. Outcomes within a cluster-period will also be correlated if patients are more similar to other patients in the same cluster-period than to patients in other periods within the cluster.

When clustering is non-ignorable it must be included in the trial analysis to obtain correct

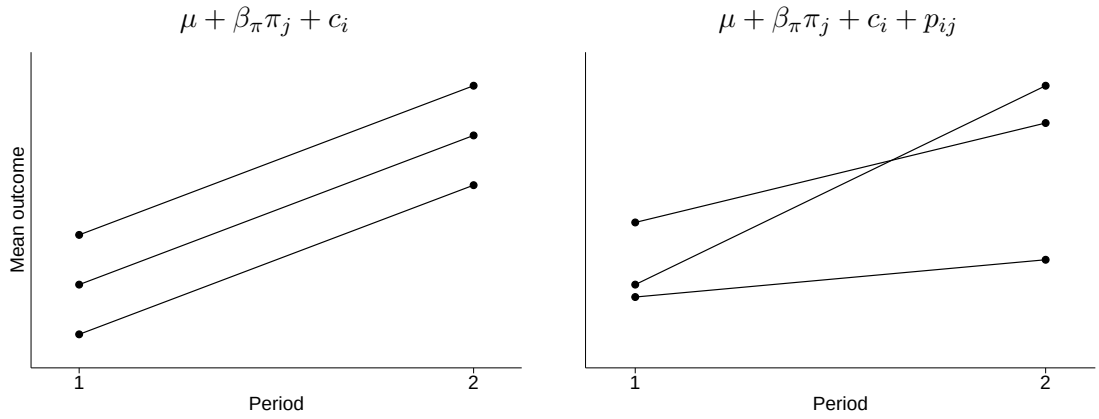

Figure 1: A sketch to illustrate how mean outcomes might be modelled in a CRXO trial. Each line represents a cluster, and each dot a cluster-period. Panel A shows a model where the mean outcome varies between the three clusters and by period, but the period effect is assumed to be the same for all clusters. Panel B allows the outcomes to change differently over time depending on the cluster.

Type I error rates. For CRXO trials, ignoring extra cluster-period correlation in the analysis may lead to inflated Type I error rates, increasing the chance of a false-positive result. Therefore, if there is extra correlation within cluster-period, the model in Panel A will give incorrect and potentially misleading results. Hence, a model such as that in Panel B should be used instead.

2.2 CRXO trials and intra-cluster correlation coefficients

As discussed in Section 2, in a CRXO trial there are two possible types of correlation — within a cluster, and additional correlation within a cluster-period — and two ICCs are therefore required to describe the data. We denote the first as ρ_c [6], the correlation between two outcomes in the same cluster-period, and define it on the logistic scale (see *e.g.* Refs. [14, 15] for a discussion of ICC definitions for binary outcomes on the logistic scale) as:

$$\rho_c = \frac{\sigma_c^2 + \sigma_p^2}{\sigma_c^2 + \sigma_p^2 + \pi^2/3}\,, \tag{3}$$

where σ_c^2 is the variance between the cluster means, σ_p^2 is the variance between the cluster-period means conditional on the cluster mean, and π is the mathematical constant. This ICC is based on an underlying linear model for a latent continuous outcome [16]. The $\pi^2/3$ term represents the residual variance of a standard logistic model. The continuous outcome is then dichotomised to produce a binary outcome following a logistic model. ρ_c represents the correlation between two outcomes in the same cluster-period on the original continuous (logistic) scale.

We denote our second ICC ρ_p and define it as:

$$\rho_p = \frac{\sigma_p^2}{\sigma_c^2 + \sigma_p^2 + \pi^2/3}\,. \tag{4}$$

It represents the additional correlation between two outcomes in the same cluster-period compared with two outcomes from different periods in the same cluster:

$$\rho_p = \text{Corr}(Y^*_{ijk}, Y^*_{ijk'}) - \text{Corr}(Y^*_{ijk}, Y^*_{ij'k'}) \,, \tag{5}$$

where:

$$\text{Corr}(Y^*_{ijk}, Y^*_{ijk'}) = \frac{\sigma_c^2 + \sigma_p^2}{\sigma_c^2 + \sigma_p^2 + \pi^2/3} = \rho_c \,, \tag{6}$$

is the correlation between two latent continuous outcomes Y^*_{ijk}, $Y^*_{ijk'}$ from cluster i, period j and subjects k and k', and:

$$\text{Corr}(Y^*_{ijk}, Y^*_{ij'k'}) = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_p^2 + \pi^2/3} = \eta \,, \tag{7}$$

is the correlation between two outcomes in cluster i but different periods j and j'.

3 Methods of analysis

In this section we start by outlining a number of methods of analysis which we then evaluate in Sections 5-7 via a simulation study. We use the subscript $i = 1, \ldots, C$ to denote cluster, where C is the total number of clusters, $j = 1, 2$ to denote period, and $k = 1, \ldots, n_{ij}$ to denote individual, with n_{ij} individuals in cluster i during period j. The number of events in each cluster-period will be denoted $Y_{ij} = \sum_{k=1}^{n_{ij}} Y_{ijk}$, where Y_{ijk} is the binary outcome for person k in period j and cluster i.

3.1 Individual-level methods

Hierarchical models

Hierarchical models can be used to describe data with a multi-level structure [16]. Variables that describe the structure of the data, such as cluster, can be included either as fixed effects, each with their own regression coefficient, or as random effects, which follow a distribution. Hierarchical models allow complex correlation structures to be modelled in a way that utilises the power available from individual-level data. They are also easily extended to account for baseline covariates. [16]

In this study we consider four different hierarchical models. The first model has fixed effects for cluster:

$$\text{logit}(\Pr\{Y_{ijk} = 1\}) = \mu + \beta_\tau \tau_{ij} + \beta_\pi \pi_j + \gamma_i \,, \tag{8}$$

where μ is an overall mean, β_τ is the treatment effect, τ_{ij} and π_j are indicator variables for treatment and period respectively, β_π is a fixed period effect, and γ_i are fixed effects for cluster. Note that clusters can be modelled as fixed effects for a CRXO trial because each cluster receives both the intervention and control, and such models utilise within-cluster information alone to estimate treatment effects. We refer to this model as "Fixed" as it has fixed effects for cluster.

The second model has a random effect for cluster rather than fixed effects:

$$\text{logit}(\Pr\{Y_{ijk} = 1 | c_i\}) = \mu + \beta_\tau \tau_{ij} + \beta_\pi \pi_j + c_i \,, \tag{9}$$

where c_i is a normally distributed random effect with mean zero and variance σ_c^2. We refer to this model as "Random" as it uses random effects for the clusters.

These two models allow the mean outcome to change going from period 1 to period 2, but all clusters change by the same amount, as in panel A of Figure 1. As discussed in Section 2.1, if there is extra correlation within cluster-period then these models will be mis-specified and may give incorrect and misleading results.

We also consider two further models. The first has fixed effects for cluster and a random effect for cluster-period:

$$\text{logit}(\Pr\{Y_{ijk} = 1|p_{ij}\}) = \mu + \beta_\tau \tau_{ij} + \beta_\pi \pi_j + \gamma_i + p_{ij}\,, \tag{10}$$

where p_{ij} is a normally distributed random effect with variance σ_p^2, and other terms are defined as above. We refer to this model as "Fixed-random".

Lastly we consider a model with random effects for both cluster and for period within cluster, which we refer to as "Random-random":

$$\text{logit}(\Pr\{Y_{ijk} = 1|c_i, p_{ij}\}) = \mu + \beta_\tau \tau_{ij} + \beta_\pi \pi_j + c_i + p_{ij}\,, \tag{11}$$

The variance components σ_c^2 and σ_p^2 from this model are used in the ICC definitions given in Section 2.2.

The Fixed-random and Random-random models allow the variation seen in panel B of Figure 1, where mean outcome in each cluster can change over time in a different way.

Generalised estimating equations

An alternative to likelihood based methods such as hierarchical models is to use generalised estimating equations (GEEs) [16–18]. In GEEs marginal probabilities are modelled and the resulting odds ratios (ORs) are population averaged [17, 18]. Unlike the ORs from the Random-random hierarchical model that compares the odds of outcome from two people chosen at random within the same cluster-period, population averaged ORs compare the odds of outcome of two people picked at random from the entire study population regardless of which cluster or cluster-period they belong to. Cluster and cluster-period are averaged over, and hence GEEs model marginal probabilities rather than probabilities that are conditional on cluster-period. The two types of OR are often very similar in practice, but they estimate two different population parameters. The extent to which the two ORs differ will depend on the size of the variance components in the Random-random hierarchical model [1].

In GEEs clustering is not accounted for by adding terms to the model, but the correlations are modelled explicitly in a working correlation matrix [16]. Standard errors (SEs) can be calculated using a robust sandwich estimator which can account for mis-specification of the working correlation matrix, although this relies on a sufficient number of clusters being available [16, 17].

The basic marginal model we consider for the GEEs is:

$$\text{logit}(\Pr\{Y_{ijk} = 1\}) = \mu + \beta_\tau \tau_{ij} + \beta_\pi \pi_j\,, \tag{12}$$

where $\Pr\{Y_{ijk} = 1\}$ is now a marginal probability and β_τ is a population averaged treatment effect.

This basic marginal model can be used in conjunction with different working correlation matrices, both with and without a robust sandwich estimator. To our knowledge, it is not

currently possible to specify a working correlation matrix that captures two different correlations for cluster and cluster-period in the Stata command `xtgee` [19], and manual coding of such a working correlation matrix and its inverse in the GEE algorithm would be required. We therefore chose to look at working correlation matrices that model the higher level of clustering (an exchangeable correlation matrix that captures a constant correlation within clusters but ignores any clustering at the cluster-period level) and that model the lower level of clustering (an "exchangeable in cluster-period" matrix that captures correlation within cluster-period but ignores additional correlation within clusters). The exchangeable correlation matrix assumes that there is a common correlation ρ between all observations in a cluster, regardless of which period the observation belongs to, corresponding to $\sigma_p^2 = 0$ in the Random-random hierarchical model. The "exchangeable in cluster-period" correlation matrix assumes that observations within the same period are exchangeable with a common correlation ρ, but observations from different periods in the same cluster are uncorrelated. This is equivalent to the assumption that the variance σ_c^2 is zero and that there is only correlation between outcomes in the same cluster-period, as would be the case in a parallel group cluster randomised trial. This working correlation assumes $\rho_c = \rho_p$, or $\eta = 0$, and therefore does not exploit the correlation between individuals in different cluster-periods that allows the cross-over element to improve efficiency over a cluster randomised design. We also consider an independent working correlation matrix.

Estimates of the SE from all GEEs were considered both with and without the use of a robust sandwich estimator.

3.2 Linear regressions on summary measures

Cluster-level methods can be used to model summary statistics for each cluster-period. The main advantages of cluster-level methods are their robustness and ease of implementation, while a potential disadvantage is that they do not make full use of the data which may lead to a loss in power [10].

We can fit a linear regression to model the proportion of events in each cluster-period. Since we are analysing cluster-period summaries we no longer need to account for correlations within cluster-period, but we do still need to account for cluster effects. Defining $P_{ij} = Y_{ij}/n_{ij}$ as the proportion of events in period j of cluster i, we can fit the following linear regression model with fixed cluster effects:

$$P_{ij} = \alpha + \beta_{\tau,L}\tau_{ij} + \beta_\pi \pi_j + \gamma_i + \epsilon_{ij}\,, \tag{13}$$

where α is an overall mean, $\beta_{\tau,L}$ is the treatment effect and τ_{ij} is an indicator variable for treatment of interest, β_π is a fixed period effect and π_j is an indicator variable for period, γ_i are fixed cluster effects, and ϵ_{ij} is a normally distributed residual error term with mean zero and variance σ_e^2. We add an L subscript to $\beta_{\tau,L}$ to highlight that the treatment effect given by this model is measured on the linear scale, *i.e.* it corresponds to a difference in proportions rather than an OR as in the individual-level models. Note that using fixed effects and random effects for cluster in this model will give identical results [20]. This model is equivalent to the cluster-level method used by Turner *et al.* [6].

An unweighted linear regression assumes that all data points have the same error variance. It is also possible to use weights in conjunction with this regression model, which may help to increase efficiency if this assumption does not hold. We therefore chose to look at the following three weightings in our simulation study, in addition to the unweighted regression. The first

weights by size of cluster, with the weight for cluster i set proportional to [6, 8]:

$$\left(\frac{1}{n_{i1}} + \frac{1}{n_{i2}}\right)^{-1}, \tag{14}$$

where n_{i1} and n_{i2} are the number of patients in cluster i in the first and second periods respectively. This weighting assumes that $\rho_c = \eta$ or equivalently that $\rho_p = 0$ (there is no extra cluster-period correlation), and that the variances for the two treatments are equal and the same for all clusters.

The second set of weights are based on the inverse of the variance of $P_{i2} - P_{i1}$ and are a combination of cluster-period size and estimated ICC. The weights for each cluster are set proportional to [6, 8]:

$$\left(\frac{1 + (n_{i1} - 1)\hat{\rho}_c}{n_{i1}} + \frac{1 + (n_{i2} - 1)\hat{\rho}_c}{n_{i2}} - 2\hat{\eta}\right)^{-1}, \tag{15}$$

where $\hat{\rho}_c$ is a sample estimate of ρ_c and $\hat{\eta}$ is a sample estimate of the correlation between two outcomes in different periods in the same cluster (and is equal to the difference between ρ_c and ρ_p — see Section 2.2). These weights relax the assumption that $\rho_c = \eta$, although still assume that the variances are the same for both treatments. For a full derivation and description of assumptions for these weights see Ref. [8].

The third set of weights uses the inverse of the binomial variance for each cluster-period:

$$\left(\frac{p_{ij}(1 - p_{ij})}{n_{ij}}\right)^{-1}. \tag{16}$$

A zero-cell correction, a technique used in meta-analysis [21] in which 0.5 is added to each p_{ij} and $1 - p_{ij}$ for any p_{ij} equal to zero or one, was used to avoid undefined weights.

In addition to these three methods of weighting, we also considered a variety of other weightings which gave results that were very similar to the size and ICC weights specified above — see Sections 1 and 6 of the online appendix for more details.

Some of these weights require estimates of the ICCs on the linear scale. For ρ_c we use an ANOVA estimator that is defined in references such as [10, 15, 22, 23]. For the correlation η we use the pairwise estimator that is given in Donner *et al.* in [23]. Both of these definitions are given in Section 2 of the online appendix.

4 Simulation study: data generation

In our simulation study we started by conducting two small initial simulation studies to identify which methods of analysis appeared to perform well (Section 5). This was followed by a full factorial simulation study on the subset of methods that performed well in the initial simulation study (Sections 6 and 7). In this section we describe the data generation process used in our simulation study. The parameters used in our simulations are outlined in Sections 5-7.

We generated data sets using Equation (11). Simulations were carried out in Stata 13 [9]. Data sets were generated such that the number of subjects varied across clusters and cluster-periods. The number of patients in each cluster-period was generated in the following way:

- Let m represent the average number of patients per cluster-period across the study. A

value of m was chosen for each scenario (see Section 6 for further details).

- A value for the average size of each cluster in the study was sampled from a negative binomial distribution with mean m and standard deviation (SD) $0.65 \times m$. A value of 0.65 was chosen as this is the coefficient of variation found in general practice list size by Eldridge, Ashby and Kerry [24] (see Section 4 of the online appendix for further discussion). Any zeros were redrawn. This gives a number m_i for each cluster, with all m_i greater than zero, representing the average size of a particular cluster.
- For each cluster, we selected the number of individuals in each cluster-period by sampling from a normal distribution with mean m_i and SD $0.65/100 \times m_i$. A coefficient of variation of 0.65/100 was chosen to ensure a smaller variation in size between periods than variation across clusters. Numbers less than or equal to zero were redrawn.

For each scenario we compared the different methods of analysis in terms of the mean estimated treatment effect, bias in the estimated SEs, failure rates, power, and Type I error rate across 5000 simulated data sets. This number of replications gives a Monte Carlo error of 0.3% when estimating the Type I error rate, assuming a true rate of 5% [25]. Methods were classed as failing to run when applied to a specific simulated data set if they did not produce parameter and standard error estimates, for example if the model did not converge.

Bias in the estimated SEs was calculated as the ratio of the model based SE to the empirical SE; empirical SEs were calculated as the SD of the individual treatment effects, and model based SEs were calculated as the square root of the mean of the treatment estimate variances. All summary measures were calculated only for scenarios where the analysis method did not fail to run (*e.g.* due to non-convergence).

5 Initial simulation study

We started by conducting two small simulation studies to rule out any methods of analysis that did not perform well across any scenarios. The first set of simulations varied the number of clusters while keeping other parameters fixed; the second set of simulations varied the size of ρ_p while keeping other parameters fixed. For each set of simulations, we set the treatment effect to zero to evaluate the Type I error rate. We used an event rate of 15% in the control arm during the first period, which corresponds to $\mu = \log\left(\frac{0.15}{1-0.15}\right) = -1.735$ in Equation (11), with a fixed period effect OR of 0.85 ($\beta_\pi = \log(0.85) = -0.163$ in Equation (11)), and a ρ_c of 0.062.

For the scenarios looking at increasing the number of clusters we used a ρ_p of 0.023, which corresponds to variance components of ($\sigma_c^2 = 0.137$, $\sigma_p^2 = 0.081$). The following values were used for the number and size of clusters:

- 6 clusters, with an average of 200 patients per cluster-period,
- 12 clusters, with an average of 60 patients per cluster-period,
- 20 clusters, with an average of 34 patients per cluster-period,
- 30 clusters, with an average of 22 patients per cluster-period,
- 50 clusters, with an average of 14 patients per cluster-period,
- 80 clusters, with an average of 8 patients per cluster-period.

The average numbers of patients per cluster-period were found by simulation as those required to give 80% power using an unweighted cluster-summary level regression for ($\rho_c = 0.062, \rho_p = 0$). Note that, since $\rho_p = 0$ for this ICC combination, $\rho_c = \eta$, *i.e.* the correlation between two observations in a cluster is the same regardless of period.

For the scenarios increasing ρ_p we used 6 or 30 clusters, with their respective average sizes of 200 and 22 patients per cluster-period. A ρ_c of 0.062 was used, as before, and ρ_p was allowed to take the following values: 0.001, 0.005, 0.01 and 0.05. These ICCs correspond to the following variance component combinations: ($\sigma_c^2 = 0.214$, $\sigma_p^2 = 0.003$), (0.200, 0.017), (0.182, 0.035) and (0.042, 0.176).

A discussion of our reasons for particular parameter choices is given in Section 4 of the online appendix.

In these initial simulations we looked at the methods outlined in Section 3; consisting of an unweighted linear regression on the cluster-level summaries plus six different types of weighting, four hierarchical models and GEEs implemented in three different ways, each with and without robust SEs. Wald test statistics for all hierarchical models and GEEs were based on the normal distribution. Test statistics for the cluster-level summary regressions were based on a t-distribution with $C - 2$ degrees of freedom. In the online appendix we also include results using a t-distribution with $C - 2$ degrees of freedom for the hierarchical models and GEEs, to enable direct comparison with the results from the cluster-level summary regressions.

5.1 Results of initial simulations and discussion

The effect of increasing number of clusters on Type I error is displayed in Figure 2. The effect of increasing ρ_p on Type I error is shown in Figure 3. We also looked at some weightings and GEEs that are not presented in the figures but that displayed very similar behaviours to other models. The full results for all methods can be found in the online appendix in Tables 1-7. The Fixed-random hierarchical model in Equation (10) that uses fixed effects for cluster and a random effect for cluster-period was found to have a very high failure rate, for as few as 12 clusters (see Table 1 of the online appendix). We therefore chose not to include this method in any further scenarios. All GEEs displayed in the figures use robust SEs.

None of the methods showed any bias in the treatment effect estimates. However, we found that many of the methods have Type I errors considerably above 5%. The hierarchical models had Type I errors that were inflated to over 10% for scenarios with only a few clusters and a non-zero ρ_p. This was also the case for most of the GEEs, while the GEE using an exchangeable in cluster-period correlation matrix had Type I errors that were generally too conservative, falling to below 4% in several of the scenarios. This is because this working correlation matrix does not exploit the correlation between individuals in different cluster-periods.

The unweighted linear regression gave a Type I error rate of between 4.2 and 5.0% across all scenarios considered in this initial study. We therefore chose to study this model further in our factorial simulation study. The linear regression that is weighted by size has inflated Type I errors (above 5.6%) for all numbers of clusters when ρ_p is 0.023. For 6 clusters the Type I error is inflated for ρ_p values of 0.005 and above. However, we decided to look further at this method across a wider range of scenarios since it is a commonly used weighting scheme in parallel group cluster randomised trials.

The ICC weighted regression suffers from a relatively high failure rate (up to 4% — see Table 6 of the online appendix) in some scenarios, in addition to having a slightly inflated Type I error. Note that if the ICCs estimated from the data gave negative weights to some clusters, which led to those clusters being excluded from the analysis, we classed these data sets as failing to run. We therefore did not take this method forward. The binomial variance weighting gave

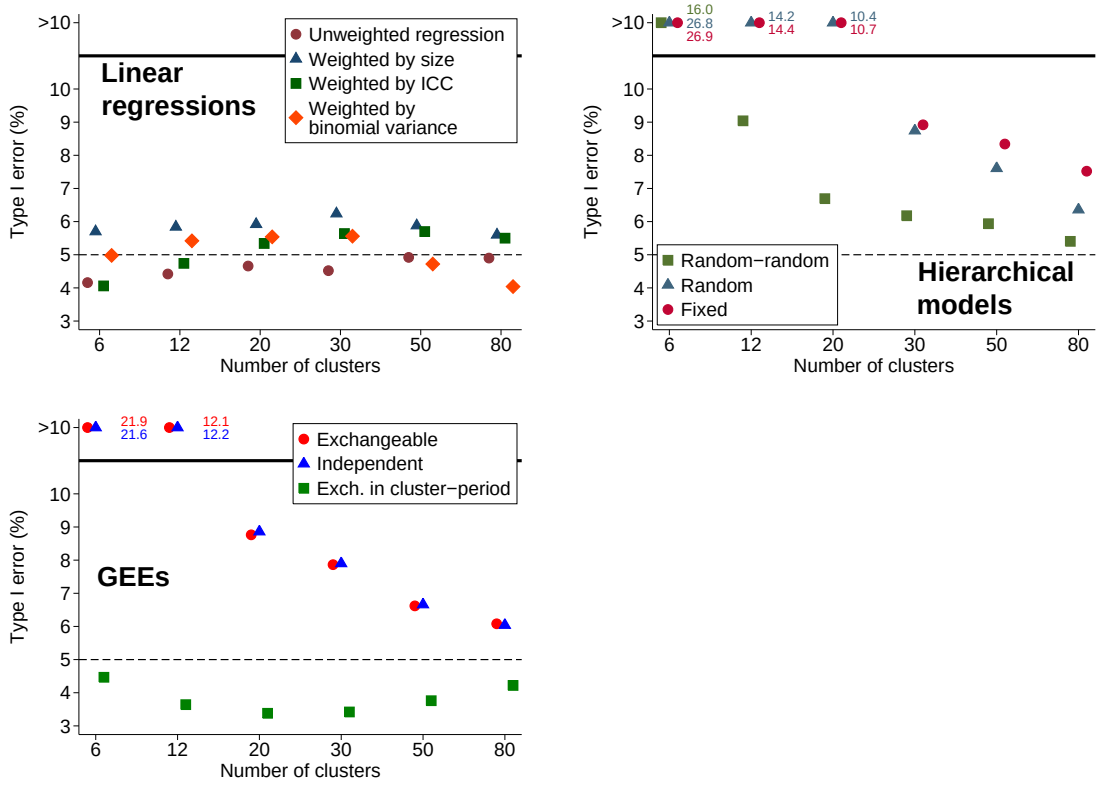

Figure 2: Type I errors across different numbers of clusters: the top left panel shows the linear regression methods, the top right panel shows the hierarchical models and the bottom left panel shows the GEEs. Random-random is a hierarchical model with random effects for cluster and cluster-period. Random and Fixed have random and fixed effects for cluster respectively. GEEs are labelled by their working correlation matrix, and all use robust SEs. Simulation parameters are set to an event rate of 15%, no treatment effect, fixed period effect OR of 0.85, $\rho_c = 0.062$ and $\rho_p = 0.023$. For clarity, Type I errors of greater than 10% have been plotted together and labelled individually.

appropriate Type I errors in the initial simulation study that increased number of clusters, but gave an inflated Type I error for larger values of ρ_p in the second initial simulation study. We therefore did not take this method forward.

Importantly, the Random hierarchical model, which has random effects for cluster but ignores any extra correlation within cluster-periods, did not perform well in these initial scenarios. It has an inflated Type I error when ($\rho_c = 0.062, \rho_p = 0.023$) even for 80 clusters, where the Type I error is 6.4%. Although the Type I error is close to the nominal value of 5% for smaller ρ_p values (up to $\rho_p = 0.005$) when there are 30 clusters, the false positive rate is too high for all values of ρ_p considered when there are only 6 clusters. Aside from a small number of scenarios (30 clusters, $\rho_p \leq 0.005$), the false positive rate for this method is inflated, in the worst cases to over 40% (6 clusters, $\rho_p = 0.05$). The results for this method, and for the Fixed model, were so poor that they were not carried forward to the full factorial simulation.

Although the Random-random hierarchical model does have inflated error rates for small numbers of clusters there is a suggestion that this behaviour improves for larger numbers of clusters — the Type I error rate is only 5.4% for 80 clusters. Although for six clusters the Type I error is 5.7% even for a very small ρ_p of 0.001, when the number of clusters is increased to

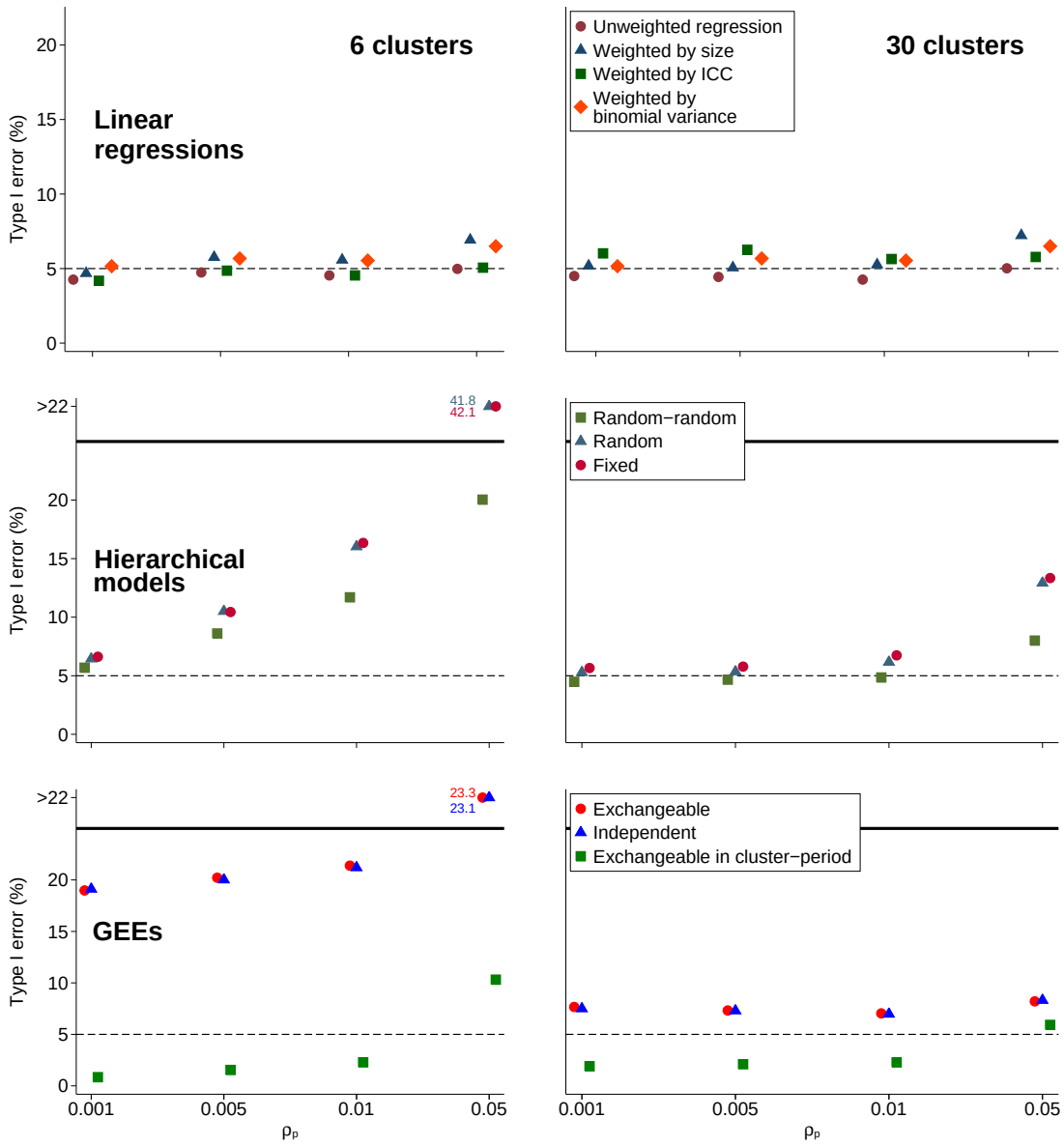

Figure 3: Type I errors across different values of ρ_p: the top panel shows the linear regression methods, the middle panel shows the hierarchical models and the bottom panel shows the GEEs. The left hand column is for 6 clusters, and the right is for 30 clusters. Other simulation parameters are set to an event rate of 15%, no treatment effect, fixed period effect OR of 0.85, and $\rho_c = 0.062$.

30 the Type I error is close to nominal until ρ_p is raised to above 0.01. The Random-random model also performs consistently better than any of the other hierarchical models. We therefore decided to take forward this model to the factorial simulation study.

6 Factorial simulation study

As described in Section 5.1, we took forward three methods to a fully factorial simulation study: an unweighted linear regression, a linear regression weighted by size and the Random-random hierarchical model. We have provided the Stata [9] code for these methods in Section 3 of the

online appendix.

The simulation parameters we varied in this study were:

- Event rate in the control arm during the first period: 15% or 45% ($\mu = \log\left(\frac{0.15}{1-0.15}\right) = -1.735$ or $\mu = \log\left(\frac{0.45}{1-0.45}\right) = -0.201$ in Equation (11)).
- Treatment effect: no treatment effect or a non-zero treatment effect (OR 0.5 for an event rate of 15%, corresponding to a decrease to an event rate of around 8%, $\beta_\tau = \log(0.5) = -0.693$ in Equation (11); OR 0.75 for an event rate of 45%, corresponding to an event rate on treatment of 38%, $\beta_\tau = \log(0.75) = -0.288$).
- Number of clusters: 6 or 30.
- ICC combinations: ($\rho_c = 0.023$, $\rho_p = 0$), (0.062, 0), (0.023, 0.01) and (0.062, 0.023), with values quoted on the logistic (underlying latent continuous variable) scale — see Section 4 of the online appendix for a discussion of these choices. These combinations correspond to variance components of ($\sigma_c^2 = 0.077$, $\sigma_p^2 = 0$), (0.217, 0), (0.044, 0.034) and (0.137, 0.081), where σ_c^2 and σ_p^2 are defined in Equation (11). Note that scenarios with $\rho_p = 0$ will correspond to the variation seen in Panel A of Figure 1. When ρ_p is non-zero the variation will be as in Panel B.
- Power/number of patients per cluster-period: for each scenario, we chose the average number of patients per cluster-period, m, such that an unweighted linear regression model gave the desired power (based on a t-distribution with $C-2$ degrees of freedom) using values for ρ_c and ρ_p of 0.062 and 0 respectively. This was done by using simulation. The event rate was set as for the particular scenario. We used numbers of patients that were needed to give either 80% or 90% power. This corresponds to 200 (80% power) and 330 (90% power) for 6 clusters and 15% event rate, 22 (80% power) and 31 (90% power) for 30 clusters and 15% event rate. For 45% event rate, averages of 400 and 600 patients per cluster-period were used for 6 clusters, and 55 and 75 for 30 clusters.

In addition a fixed period effect was generated in each data set (OR 0.85 for an event rate of 15%, corresponding to a decrease of around 2% in event rate to about 13% in the second period, $\beta_\pi = \log(0.85) = -0.163$ in Equation (11); OR 0.92 for an event rate of 45%, corresponding to the same absolute decrease of around 2%, $\beta_\tau = \log(0.92) = -0.083$).

6.1 Results

In this Section we present the results of our factorial simulation study. Full tabulated results can be found in the Section 6 of the online appendix.

6.1.1 Scenarios with an event rate of 15%

Figure 4 shows how the Type I error varies across the ICC combinations for the three methods for 6 clusters (left-hand column), and an event rate of 15%. For 6 clusters with an average of 200 patients per cluster-period (top left panel), the Type I error for the unweighted linear regression is consistently below 5% for all ICC combinations (range: 4.2 to 4.4%). However, the power drops from around 80% to below 60% for a ρ_p of 0.023, as can be seen in the top left panel of Figure 5.

For scenarios with zero ρ_p the size-weighted regression has Type I errors close to or below 5% and gives better power than the unweighted linear regression. This is as expected since the

variance in these scenarios depends on cluster size only and not ρ_p. However, when there is a non-zero ρ_p the Type I error for this method rises to 5.7% for 6 clusters and 200 patients. This behaviour is also observed for the Random-random hierarchical model, with the Type I error rising to over 10% for non-zero ρ_p combinations. The power for this model is the highest, but for scenarios with non-zero ρ_p this is not a valid comparison given the greatly inflated false positive rates.

We used a normal distribution to calculate p-values for the Random-random model. Using a t-distribution with $C - 2$ degrees of freedom can help to reduce inflated Type I errors, but when the Type I error is already close to nominal the use of the t-distribution produces overly conservative results (see tables in the online appendix). For example, for 6 clusters, an average of 200 patients per cluster-period, 15% event rate, and ($\rho_c = 0.062$, $\rho_p = 0.023$), use of the t-distribution gives a Type I error rate of 5.6% compared with 16.0% using a normal distribution. However for the same number of clusters and patients but with ($\rho_c = 0.023$, $\rho_p = 0$) the Type I error rate is 0.4% using a t-distribution, compared with 4.2% for a normal distribution. In addition, use of the t-distribution is not always sufficient to reduce the Type I error to the nominal rate. For example, for 6 clusters, 330 patients per cluster-period, 15% event rate, and ($\rho_c = 0.062$, $\rho_p = 0.023$), use of the t-distribution gives a Type I error rate of 7.1% which is still slightly inflated. It may therefore be beneficial to use a more sophisticated degree-of-freedom correction, such as the Kenward-Roger method [26, 27]. However, more advanced degree-of-freedom corrections are not always routinely available in standard statistical software packages, and further research would be needed to assess any benefits of using such a correction.

Increasing the average number of patients per cluster-period to 330, shown in the bottom left panel of Figures 4 and 5, results in a similar pattern of Type I errors and power for the three models.

For 30 clusters, shown in the right-hand column of Figure 4, the unweighted linear regression again gives appropriate Type I errors across all ICC combinations. The decrease in power for the non-zero ρ_p combinations is much less for 30 clusters, as seen in the right hand column of Figure 5. Both the size-weighted regression and the hierarchical model display inflated Type I errors for some of the scenarios with larger ICCs, particularly when ρ_p is non-zero, although to a lesser extent than with 6 clusters. They again both provide more power than the unweighted regression, but sometimes at the cost of an inflated Type I error rate.

6.1.2 Scenarios with an event rate of 45%

Results for an event rate of 45% were qualitatively similar to those for an event rate of 15%, and can be found in Section 5 of the online appendix. Unlike the scenarios with a 15% event rate, the failure rate of the Random-random hierarchical model was found to be high for some parameter values. For an event rate of 15% the failure rate was found to be lower than 0.5% in all scenarios. For an event rate of 45% and 30 clusters the failure rate was similarly low, but for only 6 clusters it varied between 1% and 9.3%, with the highest failure rates in scenarios with large numbers of patients per period per cluster and high ICCs.

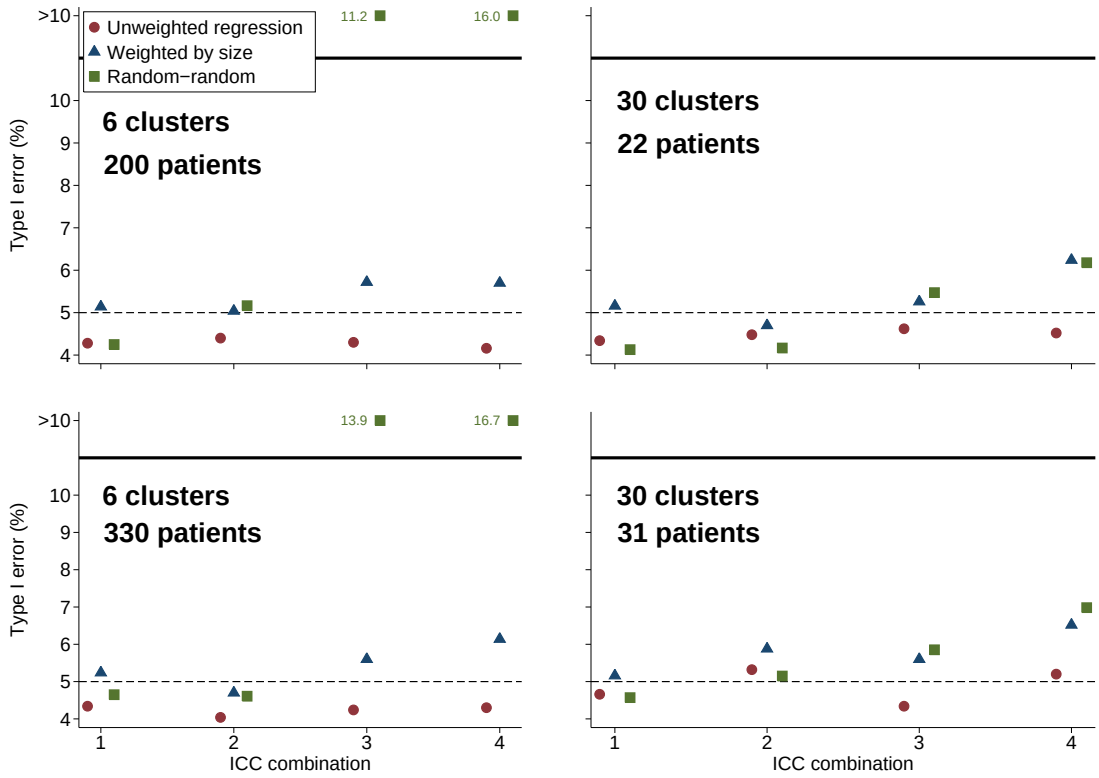

Figure 4: Type I errors across different ICC combinations: combination one is (ρ_c=0.023, ρ_p=0), combination two is (0.062, 0), combination three is (0.023, 0.01) and combination four is (0.062, 0.023). Graphs are labelled by number of clusters and average number of patients per cluster-period. Other simulation parameters are set to an event rate of 15%, no treatment effect, and a fixed period effect OR of 0.85.

7 Further exploration of the Random-random hierarchical model

Given the poor performance of the Random-random hierarchical model for certain ICC values and numbers of clusters, we explored this model further over a wider range of simulation parameters. We ran further simulations using just the Random-random model as an analysis method for the following parameters:

- Event rate: 15 %.
- Treatment effect: no treatment effect or a treatment OR of 0.5.
- Numbers of clusters: 6, 12, 20, 30, 50, 80 and 100.
- ICC combinations: ($\rho_c = 0.023$, $\rho_p = 0$), (0.062, 0), (0.023, 0.01) and (0.062, 0.023).
- Power: numbers of patients to give 80% power for an ICC combination (0.062, 0). This corresponds to 200 patients per cluster-period for 6 clusters, 60 for 12 clusters, 34 for 20 clusters, 22 for 30 clusters, 14 for 50 clusters, 8 for 80 clusters and 6 for 100 clusters.

A fixed period OR of 0.85 was used.

7.1 Results

Figure 6 shows how the Type I error and power vary across the ICC combinations and numbers of clusters for the Random-random model. Full tabulated results can be found in Section 6 of the online appendix.

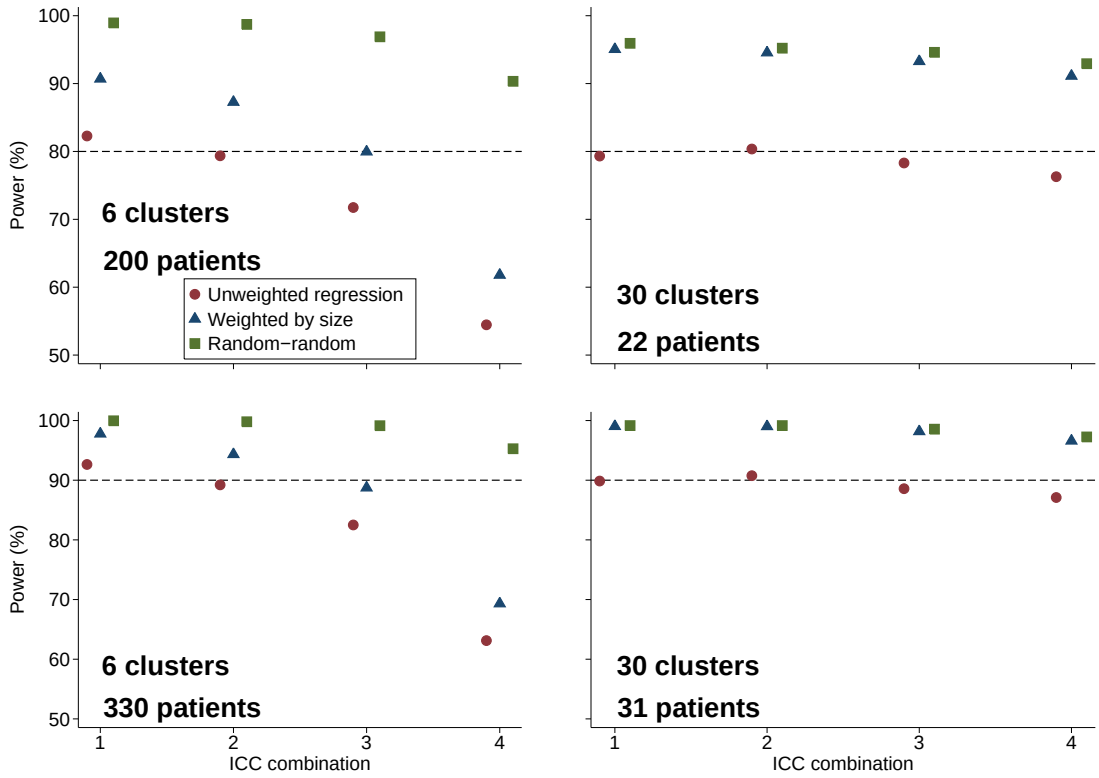

Figure 5: Power across different ICC combinations. Simulation parameters and ICC combinations are as in Figure 4 except that the treatment OR is 0.5.

For ICC combinations 1 and 2, *i.e.* those with $\rho_p = 0$, the Type I error remains close to nominal levels for all numbers of clusters studied. For non-zero ρ_p, the Type I error grows as the number of clusters is decreased. At least 50 clusters are needed for $(\rho_c = 0.023, \rho_p = 0.01)$ to get a Type I error close to 5%, and this rises to at least 80 clusters for the largest ICC combination of $(\rho_c = 0.062, \rho_p = 0.023)$.

The right hand panel of Figure 6 shows that the power remains high for all scenarios considered, although this comparison is not truly valid because of the elevated Type I error rates for scenarios with high ICCs and small numbers of clusters.

These results show that if there is extra correlation within a cluster-period, it is necessary to have a large number of clusters for the Random-random model to give Type I errors close to 5%.

8 Application to TRIGGER2

We now demonstrate how the results from the simulation study can be applied to inform the design and analysis of a CRXO trial. TRIGGER1 [28–30] was a parallel group, cluster randomised feasibility trial which compared two different haemoglobin thresholds for red blood cell transfusions for patients with acute upper gastrointestinal bleeding. One of the primary aims of TRIGGER1 was to inform the design and feasibility of a phase 3 trial, TRIGGER2.

TRIGGER1 took place in 6 UK hospitals, each of which recruited for a fixed period of 6 months. It is anticipated that TRIGGER2 will take place in between 20 and 40 hospitals, and that the primary outcome will be all-cause mortality. Due to the limited number of clusters

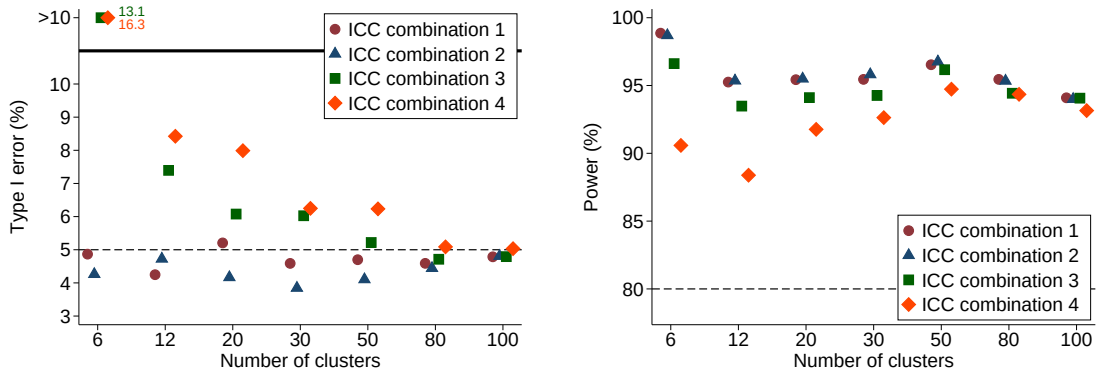

Figure 6: Type I errors (left hand panel) and power (right hand panel) for the Random-random hierarchical model, increasing the numbers of clusters across different ICC combinations: combination one is (ρ_c=0.023, ρ_p=0), combination two is (0.062, 0), combination three is (0.023, 0.01) and combination four is (0.062, 0.023). The average number of patients per cluster-period is 200 for 6 clusters, 60 for 12 clusters, 34 for 20 clusters, 22 for 30 clusters, 14 for 50 clusters, 8 for 80 clusters and 6 for 100 clusters. Other simulation parameters are set to an event rate of 15%, a fixed period effect OR of 0.85, and either no treatment effect (left hand panel) or a treatment OR of 0.5 (right hand panel).

available, as well as the likelihood that the intervention would result in a relatively small (though still clinically important) treatment difference, TRIGGER2 may be designed as a CRXO trial to increase power and reduce the number of patients required in each cluster.

Our simulation study has demonstrated that it is important to consider not only ρ_c but also ρ_p when calculating the sample size or choosing the analysis method for a CRXO trial. Because TRIGGER1 was a parallel group cluster randomised trial, estimating ρ_p is difficult. However, a crude estimate can be obtained by splitting the follow-up period into two halves; 0-3 months, and 4-6 months. From this, we estimated ρ_p as 0.012.

This ICC estimate is large enough that it should be accounted for in both the sample size estimate and the analysis. The sample size could be calculated using an analytical formula that allows for between cluster-period variation [8, 31] or by using simulation [32]. The simulation package created by Reich *et al.* [32] assumes that ρ_p is zero and would therefore not be suitable for use in situations where it is suspected to be non-zero. Because of the relatively small number of clusters, an individual-level analysis based on a hierarchical model is not likely to perform well. Therefore, an unweighted cluster-level summary analysis may be preferred.

It should be noted that our estimate of ρ_p from TRIGGER1 may not be a reliable estimate. We have assumed that each period lasts for three months, but the ρ_p estimate may not be appropriate for different period lengths. Additionally, this estimate is based on only 6 clusters and will therefore have a large error associated with it.

This demonstrates the challenges in trying to estimate ρ_p to help inform the design and analysis of CRXO trials. Estimates are unlikely to be routinely available from previously reported CRXO trials, and estimates from existing datasets may face similar issues as TRIGGER1.

9 Discussion

Cluster randomised cross-over trials may be useful in settings where recruiting larger numbers of clusters is not possible and carry-over of an intervention will not be a problem. The potential for carry-over should be carefully considered before using a cross-over design since any residual effects of treatment that are present in later periods will bias estimates of treatment effect [20]. If there is no carry-over, using a cross-over element in the design may increase the power and help counteract imbalance between the arms if there are only a small number of clusters [3]. However, it adds another level of clustering, periods within clusters, which may complicate the sample size calculation and analysis.

We have demonstrated that it is necessary to model ρ_p in the analysis, and results from our simulation study indicate that Type I errors can be substantially inflated (to over 20% in some cases) if this is not done appropriately. Although it may be tempting to model only the highest level of clustering for simplicity, our results show that this leads to inflated Type I errors. Using a hierarchical model without a random effect for cluster-period results in higher Type I errors than a model which does include such a term. In addition using the Random-random model when ρ_p is zero does not result in inappropriate Type I errors. It is therefore important to model all levels of clustering, not just the highest. At present, this does not seem to be generally acknowledged when analysing CRXO trials. The systematic review conducted by Arnup *et al.* [4, 5] deemed that out of 127 analyses performed at the individual-level, only four used potentially appropriate methods that account for both levels of clustering of the CRXO design. Fifty-four of the individual-level analyses did not account for either the clustering or cross-over, and no analyses used a random effect for cluster-period.

If ρ_p is zero then the unweighted and size-weighted cluster-level summary methods and the Random-random hierarchical model all appear to perform well. When ρ_p is non-zero, our results demonstrate that the number of clusters in a CRXO trial is very important in differentiating between methods. Choosing an appropriate analysis for a small number of clusters becomes very difficult. This is especially concerning given that not being able to recruit a large number of clusters may be a common reason for conducting a CRXO trial. Arnup *et al.*'s review of 91 CRXO trials found the median number of clusters to be 9 (IQR 4-21) [5].

We found that an unweighted cluster-level regression method is robust across all scenarios considered, but that this method can lose power when ρ_p is non-zero, especially for small numbers of clusters. These results agree with those in Forbes *et al.* [8], who found that cluster methods generally work well. Given this loss of power, it might be tempting to use a size-weighted linear regression. However, we found that this method did not work well in a wide variety of different scenarios.

Despite our results showing the robustness of an unweighted cluster-level regression method, the review by Arnup *et al.* [5] found that only 9% (12/139) of analyses were performed at the cluster-level. We have demonstrated that care needs to be taken as to whether it is appropriate to use an individual-level method of analysis such as the Random-random hierarchical model with random effects for cluster and cluster-period. For small numbers of clusters, as may be the case in many CRXO trials in practice, the Type I error is inflated for this model. Adopting a degree-of-freedom correction such as the Kenward-Roger method may reduce the Type I error rate in these scenarios, although further study is required to verify this. For such study, we note that implementation of Kenward-Roger degrees of freedom in generalised linear mixed models

is available in the SAS software, but we are not aware currently of its implementation in Stata or R outside of linear mixed models.

As discussed in Section 8, it can be difficult to find a reliable estimate for ρ_p for use in a sample size calculation or when deciding on a method of analysis. In the absence of a good estimate for the ICC, we recommend assuming a non-zero effect and accounting for it in the analysis. This will help to ensure correct confidence intervals and p-values, and avoid inflating the Type I error rate. For example, Gireaudeau *et al.* recommend in some circumstances using ρ_p set to half the size of ρ_c [31]. A strategy such as this is preferable to alternatives such as assuming that ρ_p is 0 and ignoring it in the analysis (which could lead to substantially increased Type I error rates), or using data from the trial to estimate the ICC and choosing whether to account for clustering by period in the analysis based on this estimate (as this type of preliminary testing strategy has been shown to perform poorly in many situations [33–35]).

Given the particular issues of loss of power and small numbers of clusters that have been highlighted by our simulation study, it would also be important to consider these issues at the stage of planning the sample size to be used for a CRXO trial. Giraudeau *et al.* [31] and Forbes *et al.* [8] have both published some work on sample size calculations for CRXO trials. It may also be worth calculating sample size by simulation, in order to consider the effect of likely values of ρ_p and numbers of clusters, extending the work of Reich *et al.* [32] to the case of non-zero ρ_p. The risk of using too few clusters has also been discussed in Ref. [36].

Our study contained some limitations. For the hierarchical models, we considered only a logit link as to our knowledge it is not possible to specify other link functions with a Random-random model for a binary outcome in Stata. However, the Random-random model with other link functions, such as log or identity, could be easily specified in other software packages such as SAS; although further research is required to ensure these models perform adequately.

In our simulation study, we only considered CRXO trials with two time periods and different individuals in each period. The Arnup systematic review [4, 5] of 91 CRXO trials found that 58 trials (69% of those with number of periods available) had two periods only, and that only 27 trials (30%) included the same individuals in all periods, suggesting that our results will be relevant to many CRXO trials that have been conducted. The methods of analysis considered in this paper could be extended to account for more periods. The methods could also be extended to other trial designs with multiple periods and clustering, such as stepped wedge trials. Such trials may have non-zero ρ_p and, given our results, it would be important to account for this in the analysis. However, more research would be needed to evaluate how the methods perform in those scenarios. For example, autocorrelation may become an important factor with increasing numbers of periods.

We simulated data sets with unequal numbers of patients per cluster-period. Our results may not be generalisable to situations with different distributions of patients across cluster-periods, since loss of power for unequal cluster sizes versus equal cluster sizes depends on the cluster size distribution [37]. However, allowing cluster sizes to vary is more realistic than assuming equal cluster sizes and is likely to reflect what will happen in a CRXO trial in practice, so our results offer a pragmatic comparison of methods and show what happens to the power under one possible data generation method.

10 Conclusions

Ignoring ρ_p in a CRXO trial can lead to inflated Type I errors if ρ_p is non-zero. Given that it will be very difficult to completely rule out the existence of a non-zero ρ_p, an analysis method that accounts for this should generally be chosen.

However, accounting for ρ_p is difficult. A hierarchical model with random effects for cluster and cluster-period requires a very large number of clusters if there is additional correlation within a cluster-period. For values of the extra correlation considered in this study, at least 50-80 clusters were required for nominal Type I error rates, with more clusters required for larger differences in correlation. An unweighted cluster-level summary method can be used with a smaller number of clusters but may lose power. If using this method we therefore recommend that the sample size used is large enough to account for this potential loss of power. Sample size simulations which account for both levels of clustering may be of use to establish how large a trial is needed.

Acknowledgements

KEM was funded through an NIHR research methods fellowship (MET-12-16). This article presents independent research funded by the National Institute for Health Research (NIHR). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

References

[1] Eldridge S, Kerry S. A Practical Guide to Cluster Randomised Trials in Health Services Research. Wiley; 2012.

[2] Rietbergen C, Moerbeek M. The design of cluster randomized crossover trials. JEBS. 2011;36 (4):472–490.

[3] Reich NG, Milstone AM. Improving efficiency in cluster-randomized study design and implementation: taking advantage of a crossover. Open Access Journal of Clinical Trials. 2014;6:11–15.

[4] Arnup SJ, Forbes AB, Kahan BC, Morgan KE, McDonald S, McKenzie JE. The use of the cluster randomized crossover design in clinical trials: protocol for a systematic review. Syst Rev. 2014;3:86. doi: 10.1186/2046-4053-3-86.

[5] Arnup SJ, Forbes AB, Kahan BC, Morgan KE, McKenzie JE. Appropriate statistical methods were infrequently used in cluster-randomized crossover trials: Results from a systematic review. Journal of Clinical Epidemiology. 2016;74:40–50. http://dx.doi.org/10.1016/j.jclinepi.2015.11.013.

[6] Turner RM, White IR, Croudace T. Analysis of cluster randomized cross-over trial data: A comparison of methods. Statist Med. 2007;26:274–289.

[7] Parienti JJ, Kuss O. Cluster-crossover design: A method for limiting clusters level effect in community-intervention studies. Contemporary Clinical Trials. 2007;28:316–323.

[8] Forbes AB, Akram M, Pilcher D, Cooper J, Bellomo R. Cluster randomised crossover trials with binary data and unbalanced cluster sizes: Application to studies of near-universal interventions in intensive care. Clin Trials. 2015;12(1):34–44. doi: 10.1177/1740774514559610.

[9] StataCorp. Stata Statistical Software: Release 13.; 2013. College Station, TX: StataCorp LP.

[10] Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. Wiley; 2010.

[11] Hayes RJ, Moulton LH. Cluster Randomised Trials. Chapman and Hall; 2009.

[12] Murray DM. Design and Analysis of Group-Randomized Trials. Oxford University Press; 1998.

[13] Kahan BC, Morris TP. Assessing potential sources of clustering in individually randomised trials. BMC Med Res Meth. 2013;13:58. doi: 10.1186/1471-2288-13-58.

[14] Eldridge SM, Ukoumunne OC, Carlin JB. The Intra-Cluster Correlation Coefficient in Cluster Randomized Trials: A Review of Definitions. Int Statist Rev. 2009;77:378–394.

[15] Wu S, Crespi CM, Wong WK. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemporary Clinical Trials. 2012;33:869–880.

[16] Rabe-Hesketh S, Skrondal A. Multilevel and Longitudinal Modelling Using Stata (2nd ed.). Stata Press; 2008.

[17] Zeger SL, Liang KY, Albert PS. Models for Longitudinal Data: A Generalized Estimating Equation Approach. Biometrics. 1988;44:1049–1060.

[18] Neuhaus JM, Kalbfleisch JD, Hauck WW. A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data. International Statistical Review. 1991;59 (1):25–35.

[19] StataCorp . Stata Statistical Software;. http://www.stata.com/manuals13/xtxtgee.pdf.

[20] Jones B, Kenward MG. Design and Analysis of Cross-Over Trials (2nd ed.). Chapman and Hall/CRC; 2003.

[21] Higgins JPT, (editors) SG. Cochrane Handbook for Systematic Reviews of Interventions. The Cochrane Collaboration; 2011. Version 5.1.0 [updated March 2011]. Available from www.cochrane-handbook.org.

[22] Ridout MS, Demétrio CG, Firth D. Estimating intraclass correlation for binary data. Biometrics. 1999;55:137–48.

[23] Donner A, Klar N, Zou G. Methods for the statistical analysis of binary data in split-cluster designs. Biometrics. 2004;60:919–925.

[24] 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 Epi. 2006;35:1292–1300. doi:10.1093/ije/dyl129.

[25] White I. simsum: Analyses of simulation studies including Monte Carlo error. The Stata journal. 2010;10, Number 3:369–385.

[26] Kenward MG, Roger JH. Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. Biometrics. 1997;53 (3):983–997.

[27] Kenward MG, Roger JH. An improved approximation to the precision of fixed effects from restricted maximum likelihood. Computational Statistics & Data Analysis. 2009;53 (7):2583–2595. doi: 10.1016/j.csda.2008.12.013.

[28] Jairath V, Kahan BC, Gray A, Doré CJ, Mora A, Dyer C, et al. Restrictive vs Liberal Blood Transfusion for Acute Upper Gastrointestinal Bleeding: Rationale and protocol for a cluster randomized feasibility trial. Transfus Med Rev. 2013;27 (3):146–53.

[29] Kahan BC, Jairath V, Murphy MF, Doré CJ. Update on the transfusion in gastrointestinal bleeding (TRIGGER) trial: statistical analysis plan for a cluster-randomised feasibility trial. Trials. 2013;14:206. doi: 10.1186/1745-6215-14-206.

[30] Jairath V, Kahan BC, Dore C, et al. Restrictive versus liberal blood transfusion for acute upper gastrointestinal bleeding (TRIGGER): a pragmatic, open-label, cluster randomised feasibility trial. Lancet. 2015;386(9989):137–144. doi: 10.1016/S0140-6736(14)61999-1.

[31] Giraudeau B, Ravaud P, Donner A. Sample size calculation for cluster randomized cross-over trials. Statist Med. 2008;27:5578–5585.

[32] Reich NG, Myers JA, Obeng D, Milstone AM, Perl TM. Empirical power and sample size calculations for cluster-randomized and cluster-randomized crossover studies. PLoS ONE. 2012;7 (4):e35564. doi: 10.1371/journal.pone.0035564.

[33] Kahan BC. Bias in randomised factorial trials. Stat Med. 2013;32 (26):4540–4549. doi: 10.1002/sim.5869.

[34] Freeman PR. The performance of the two-stage analysis of two-treatment, two-period crossover trials. Stat Med. 1989;8 (12):1421–1432.

[35] Shuster JJ. Diagnostics for assumptions in moderate to large simple clinical trials: do they really help? Stat Med. 2005;24 (16):2431–2438. doi: 10.1002/sim.2175.

[36] Taljaard M, Teerenstra S, Ivers NM, Fergusson DA. Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. Clinical Trials. 2016;Doi: 10.1177/1740774516634316.

[37] van Breukelen GJ, Candel MJ, Berger MP. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. Stat Med. 2007;26(13):2589–2603.