**Research Article**

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

# Analysis of multicentre trials with continuous outcomes: when and how should centre effects be accounted for?

**Brennan C Kahan[a]*, Tim P Morris[a,b]**

**In multicentre trials, randomisation is often done using permuted blocks stratified by centre. It has previously been shown that stratification variables used in the randomisation process should be adjusted for in the analysis in order to obtain correct inference. For continuous outcomes, the two primary methods of accounting for centres are fixed-effects and random-effects models. We discuss the differences in interpretation between these two models, and the implications that each pose for analysis. We then perform a large simulation study comparing the performance of these analysis methods in a variety of situations. In total, 378 scenarios were assessed. We found that random centre effects performed as well or better than fixed-effects models in all scenarios. Random centre effects models led to increases in power and precision when the number of patients per centre was small (e.g. 10 patients or less), and in some scenarios when there was an imbalance between treatments within centres, either due to the randomisation method or to the distribution of patients across centres. With small samples sizes, random-effects models maintained nominal coverage rates when a degrees of freedom correction was used. We assessed the robustness of random-effects models when assumptions regarding the distribution of the centre-effects were incorrect, and found this had no impact on results. We conclude that random-effects models offer many advantages over fixed-effects models in certain situations, and should be used more often in practice. Copyright © 0000 John Wiley & Sons, Ltd.**

**Keywords:** Multicentre trials, continuous outcomes, fixed effects, random effects, randomised controlled trials

## 1. Introduction

Many randomised controlled trials (RCTs) recruit patients to multiple centres or hospitals, rather than to a single centre. This is because it may be difficult or impossible to recruit enough patients to a single centre to fulfil the sample size requirements in a reasonable time frame. Additionally, multicentre trials allow treatments to be tested in a variety of different settings, allowing the study results to be more generaliseable than results from a single centre trial [1].

In multicentre trials, patients in the same centre tend to have correlated outcomes, meaning they are more similar to other patients from the same centre than to patients from other centres. This is measured by the intraclass correlation coefficient (ICC), which measures the proportion of the total variability explained by the between-centre variance. Larger ICC values indicate a higher level of correlation between individuals in the same centre. This correlation can arise because of either

[a]*MRC Clinical Trials Unit, Aviation House, 125 Kingsway, London WC2B 6NH, UK*
[b]*MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 0SR, UK*
**Correspondence to: MRC Clinical Trials Unit, Aviation House, 125 Kingsway, London WC2B 6NH. E-mail: brk@ctu.mrc.ac.uk*

*Contract/grant sponsor: Tim Morris was supported by a UK Medical Research Council studentship MC_US_A737_0012*

differences between centres or differences between patients who present to different centres. For example, centres may have different clinical procedures or the quality of staff may vary between centres, both of which could lead to correlation between patient outcomes at the same centre. Conversely, patients presenting to one centre may be older and/or sicker than patients presenting to another centre, and so will have poorer outcomes.

For these reasons, RCTs with multiple centres often use balanced randomisation (most commonly permuted blocks within centre) to ensure similar proportions of treatment assignments across centres. It has previously been shown that when balanced randomisation has been used, all balancing factors should be adjusted for in the analysis, otherwise the standard errors for treatment effect will be biased upwards [2]. Therefore, if the randomisation process has balanced on centre, it is generally recommended that centre-effects should be accounted for in the analysis. However, this can be problematic in situations where there are few patients per centre, as this can involve estimating a large number of parameters compared to the overall sample size.

For continuous outcomes, the two primary methods of adjusting for centre in the analysis are models using either fixed centre-effects (FCE) or random centre-effects (RCE). FCE is the most common method of adjustment [3], however Pickering and Weatherall have shown when many centres have very few patients, RCE can increase power [4].

The outline of this paper is as follows: Section 2 explores issues surrounding unadjusted analyses after balancing on centre. Section 3 describes the statistical differences and differences in interpretation between FCE and RCE. Sections 4 and 5 present a simulation study comparing FCE and RCE in a variety of situations. Section 6 is a literature review of recent trials and assesses how frequently multicentre trials were performed, the average number of patients in each centre, and how often FCE and RCE were used in practice. Section 7 is a discussion. We restrict our attention to continuous outcomes, and do not consider the issue of treatment-by-centre interaction.

## 2. Adjusted vs unadjusted analyses after balanced randomisation

When centres or prognostic factors have been balanced between treatment groups during randomisation, it is necessary to adjust for these factors in the analysis in order to obtain valid results (i.e. correct coverage and type I error rates) [2, 5]. If an unadjusted analysis is performed after balanced randomisation, the estimate of treatment effect will be unbiased, however the standard errors (SEs) will be biased upwards by a factor of $\sqrt{1/(1-\mathrm{ICC})}$, leading to coverage and type I error rates that are incorrect, and a reduction in power.

The amount of the bias depends on the ICC: if the ICC is small it is unlikely that an unadjusted analysis will have much impact; for an ICC of 0.01 the SE will be biased upwards by only 1%, leading to a coverage rate of 95.2%, rather than the nominal 95%. However, higher ICCs will have a larger impact. ICCs of 0.05 and 0.10 will lead to coverage rates of 96.1% and 97.1% respectively. Parzen et al [5] gave an example where ignoring centre-effects after balanced randomisation led to a SE that was biased upwards by 22%, resulting in a large difference in p-values (0.027 unadjusted vs 0.006 adjusted)[5]. Therefore, in deciding whether to adjust for centre-effects, it is necessary to consider the size of the ICC. For a small ICC, the centre-effects can be ignored with little impact on the analysis. However, determining whether an ICC is small enough to ignore is difficult. There are two potential approaches to the issue.

The first approach involves estimating the ICC during the analysis of the trial, and deciding whether an adjusted analysis is necessary based on this estimate. A major concern for this approach is that it allows trialists to choose between two potential models for presentation (where they may choose the most favourable result). Even if the method of model selection is pre-specified (e.g. adjusting for centre-effects only if the estimated ICC is above a certain threshold), problems may still occur, as there will inevitably be errors which result in an unadjusted analysis when the centre-effects are large enough to cause bias. This is particularly a concern as the power to detect whether the ICC is greater than 0 will generally be low. Therefore, basing the analysis method on results from the trial data is not recommended.

The second approach involves using prior data to estimate the likely size of the ICC. Authors have previously catalogued ICC estimates in certain areas to help inform future studies. Adams et al [6] surveyed 1039 outcomes from 31 primary care studies, and found the median ICC was only 0.01, which would generally be small enough to ignore in an analysis. However, 40% of outcomes had an ICC between 0.01 and 0.055, 10% an ICC of 0.055 or higher, 5% an ICC of 0.095 or higher, and 1% an ICC of 0.27 or higher. Cook et al [7] presented 48 ICCs from multicentre surgery trials, and found 33% were less than 0.01, 23% were between 0.01 and 0.05, 15% were between 0.05 and 0.10, and 29% were greater than 0.10. Over 10% of ICCs from Adams et al [6] and 44% of ICCs from Cook et al [7] were greater than 0.05, which would be large enough to cause biased SEs, and to significantly affect coverage rates.

It has previously been noted that ICC estimates are highly variable [7]. Therefore, the approach of estimating the ICC value based on previous data can be problematic, as results could potentially be misleading. This issue is further explored in the next section.

Although this paper primarily focuses on continuous outcomes, the issues of adjusted vs unadjusted analyses apply to all outcome types, and consequently the examples in the following section uses binary endpoints.

### *2.1. A case study: the MIST trials*

In order to explore this issue of basing the analysis method (adjusted vs unadjusted for centre-effects) on ICC estimates from previous data, we examine ICC estimates from two previous trials.

The MIST1 and MIST2 trials [8, 9] assessed different treatments for patients with pleural effusions. MIST1 recruited 454 patients from 61 centres, and MIST2 recruited 210 patients from 11 centres. The two trials used the same inclusion/exclusion criteria, and all centres used in MIST2 were also used in MIST1; therefore, it would seem reasonable to use the ICC estimates from MIST1 to help inform the analysis method for MIST2. For example, one could look at the ICC of an outcome in MIST1 to decide whether adjustment for centre-effects is necessary in MIST2. Table 1 shows ICC estimates and 95% confidence intervals for both trials. Four outcomes are considered: mortality at 3 months, overall mortality, referral for surgery at 3 months, and overall referral for surgery.

It is apparent from Table 1 that for both trials the ICC estimates are imprecisely estimated. The 95% CI for the ICC for mortality at 3 months in the MIST1 trial is $<0.001$ to 0.88 for example. Even the most precisely estimated ICC has a CI wide enough to be of no help: the 95% CI for surgery at 3 months was 0.01 to 0.34.

In addition, ICC estimates varied not only between the two trials (where the ICC for overall mortality was 0.13 and $<0.001$ in MIST1 and 2 respectively), but also within a trial depending on the time at which an outcome was defined (e.g. within the MIST1 trial, the ICCs for overall mortality and mortality at 3 months were 0.13 and 0.02 respectively).

This demonstrates that choosing the analysis method based on previous data can be misleading. For example, if the situations were reversed and the ICCs for overall mortality in MIST1 and 2 were $<0.001$ and 0.13 respectively, we would likely not adjust for centre-effects in MIST2. However, ignoring centre-effects when the ICC was 0.13 would lead to a SE that was biased upwards by 15%.

Given how imprecise ICC estimates are, and the fact they can vary both between trials, and even within trials depending on when they are measured, we do not recommend basing the decision of whether to adjust for centre-effects on ICC estimates from previous data. Instead we recommend that when centre has been balanced on during randomisation, it is prespecified in the protocol or statistical analysis plan that centre-effects will be accounted for in the analysis.

## 3. Fixed and Random centre-effects

### *3.1. Differences in interpretation*

The two primary methods of adjusting for centre with a continuous outcome are FCE and RCE. A major difference between them is the way in which they treat the centres. FCE treats centres as fixed factors, no different to other covariates such as age or gender. RCE however assumes the centre-effects are random variables that follow some distribution (generally a normal distribution).

The chosen method of analysis (FCE vs RCE) can be important if the centre-effects themselves are of interest (e.g. testing one specific centre against another vs estimating the overall variability of the centre-effects would lead to different analysis choices). However, in the context of RCTs, centre-effects can be regarded as a nuisance parameter, as the primary goal is to estimate the treatment effect. From this perspective, either FCE or RCE are adequate choices.

It has previously been argued that one advantage of RCE is they allow the results to be generalised to centres not involved in the trial [4, 10] (as opposed to FCE where the results would only apply to those centres that took part in the trial). However this argument assumes that trial centres were randomly sampled from a wider population of centres. This will rarely, if ever be the case in RCTs, as centres are carefully selected on the basis of their ability to recruit patients and to adhere to the trial protocol.

It is interesting to note that both FCE and RCE can be used to account for individual patient-effects in crossover trials with a continuous outcome. However, the idea that RCE could generalise the results to a wider population of patients than

those included in the trial has, to our knowledge, not been raised [11, 12], and the estimated treatment effects from both models are regarded as giving the same interpretation.

We would therefore argue that using RCE does not allow the treatment effect to be generalised to centres outside of the trial, meaning that estimated treatment effects from both FCE and RCE analyses have the same interpretation. Any generalisation to patients or centres outside the trial should be done on the basis of external validity, rather than on the basis of a particular statistical model.

### *3.2. Differences in analysis*

The primary difference in estimating a treatment effect between the two methods is FCE rely solely on within-centre comparisons (a treatment effect is estimated within each centre separately, and the results are then combined), whereas RCE combine within and between-centre estimates.

From Senn [13] and Jones et al [14], the formula for estimating a treatment effect using FCE is:

$$\hat{\beta}_{\text{within}} = \sum_{i=1}^{c} w_i d_i \tag{1}$$

where $c$ is the number of centres, $w_i = \frac{n_i}{\sum_{i=1}^{c} n_i}$ , $d_i$ is the mean treatment difference in centre $i$, and $n_i$ is the number of patients recruited to centre $i$. The above formula assumes that the standard deviation (SD) of the outcome is the same across all centres.

RCE use both the within-centre estimate (as above), and the between-centre estimate. From Rabe-Hesketh and Skrondal [15], the between-centre treatment effect is estimated by first calculating the mean response for each centre. A regression model is run with each centre being treated as an observation, the mean response from each centre as the outcome, and the proportion of patients on the treatment of interest in each centre as a covariate. The resulting regression coefficient is the between-centre treatment effect.

The formula for calculating the treatment effect using RCE is:

$$\hat{\beta}_{\text{random}} = (1-m)\hat{\beta}_{\text{between}} + m\hat{\beta}_{\text{within}} \tag{2}$$

where

$$m = \frac{\hat{SE}(\hat{\beta}_{\text{between}})^2}{\hat{SE}(\hat{\beta}_{\text{between}})^2 + \hat{SE}(\hat{\beta}_{\text{within}})^2} \tag{3}$$

This combines the within and between-centre estimators of treatment effect.

### *3.3. Implications of each analysis*

Because FCE use a strictly within-centre comparison, centres where all patients are assigned to only one treatment arm by chance are not used in the estimate of the treatment effect. However, these centres do contribute to the estimate of the residual SD, provided they recruit more than one patient. Centres that recruit only one patient do not contribute to either the estimate of treatment effect or the residual SD (i.e. the patient is dropped from the analysis entirely). In contrast, centres where all patients are assigned to only one treatment arm contribute to both the estimate of treatment effect and residual SD when using RCE, even if only one patient is recruited to the centre.

Because FCE rely on within-centre comparisons, the distribution of the centre-effects has no impact on the analysis. The estimated treatment effect and SE for treatment effect will be exactly the same regardless of the centre-effects (including the case of centre-effect outliers). When using RCE, the variance of the centre-effects is used in the inference and estimation of the treatment effect, so different centre-effects could lead to different results.

Inference for FCE relies on the t-distribution, so it is reliable even in small sample situations (provided the patient level residuals are normally distributed). Inference for RCE relies on asymptotic results, and may not perform well in small sample situations. In particular, RCE use the normal distribution for inference rather than the t-distribution, which may result in confidence intervals that are too narrow and p-values that are too small, leading to type I error rates that are too high. However, it is unclear whether this is an issue for multicentre trials, many of which recruit at least 100 patients.

A number of modifications to the inference for RCE are available. These typically involve using the t-distribution for inference rather than the normal distribution. There are several different methods available for calculating the degrees

  

of freedom for the t-distribution. A simple method involves using the same residual degrees of freedom as when using FCE. This is calculated as the number of patients minus the number of parameters (including the number of centres). For example, a trial with 100 patients across 10 centres that fit only a treatment effect in the model would have 89 degrees of freedom. SAS has several different methods of calculating the degrees of freedom for RCE, however, most other major software packages do not offer any degree of freedom corrections, and instead rely on the normal distribution, which assumes either the variance is known or the sample size is large (e.g. Stata, R, and MLwiN).

## 4. Simulation study

A simulation study was performed to compare FCE and RCE in a variety of situations. Three primary situations were considered: (1) when the sample size is small compared to the number of centres (i.e. when many centres contain very few patients); (2) when the sample size is large compared to the number of centres (i.e. when most centres contain a large number of patients); and (3) when there are very few centres. These scenarios are described further in sections 4.1–4.3.

We generated data from the following model:

$$y_{ij} = \alpha + \beta t_i + u_j + \varepsilon_{ij} \tag{4}$$

where $\beta$ is the treatment effect, $t_i$ is an indicator of the treatment assignment for the $i$th individual, $u_j$ is the centre effect for centre $j$, and follows a normal distribution with mean 0 and standard deviation $\sigma_j$, and $\varepsilon_{ij}$ is the residual for patient $i$ in centre $j$, and follows a normal distribution with mean 0 and standard deviation $\sigma_i$. $u_j$ and $\varepsilon_{ij}$ were generated independently. The treatment effect was set to give 80% power based on the sample size and the residual SD (ignoring centre-effects).

For each scenario, we varied the following parameters:

- The ICC: values of 0.01, 0.05, and 0.10 were used. The residual SD ($\sigma_i$) was held constant at 1, and $\sigma_j$ was varied to give the desired ICC.
- The overall number of patients and number of centres (this is described further in sections 4.1–4.3, and in the online appendix).
- The method of randomisation. Three different methods were used: (1) permuted blocks within centre with a block size of 4; (2) permuted blocks within centre with a block size of 16; and (3) simple randomisation.
- The distribution of patients across centres, i.e. whether most patients were concentrated in a select few centres, and the remaining centres had relatively few patients (skewed patient distribution), or there was a relatively even number of patients in all centres (even patient distribution). More information can be found in the online appendix.

We used three methods of analysis: (1) FCE; (2) RCE with inference based on the normal distribution (using a Wald test); and (3) RCE with inference based on the t-distribution, using the same degrees of freedom (DF) as in a FCE analysis. All RCE models were estimated using restricted maximum likelihood.

Different analyses were compared in terms of (1) % bias in the SEs [16]; (2) coverage (proportion of times the 95% CIs contained the true treatment effect); (3) power; and (4) relative efficiency of RCE compared to FCE (calculated as the empirical SE of FCE divided by the empirical SE of RCE). 8000 replications were used for all scenarios in order to give a SE of less than 0.25% when estimating the coverage, assuming the true coverage is 95%.

Additional results for each set of simulations is available in the online appendix. All simulations were performed using Stata 12.1. There were no convergence issues for either FCE or RCE in any scenario.

### *4.1. Few patients per centre*

We simulated data based on the following six patient-centre combinations (where $c$ denotes the number of centres, and $n$ denotes the total sample size); 1) $c = 50$, $n = 100$; 2) $c = 100$, $n = 200$; 3) $c = 50$, $n = 150$; 4) $c = 100$, $n = 300$; 5) $c = 50$, $n = 200$; and 6) $c = 100$, $n = 400$. Scenarios 1–2 had on average two patients per centre, scenarios 3–4 had on average three patients per centre, and scenarios 5–6 had on average four patients per centre. The number of patients in each centre for each scenario can be found in the online appendix.

*4.1.1. Extension to three treatment arms* Many trials compare three or more treatments. In this setting, the important consideration is the number of patients available for each treatment arm per centre, rather than the number of patients per centre. For example, an average of six patients per centre may seem adequate. However, it is conceivable that in some centres all patients are randomised to two treatment arms and none to the third, particularly if large block sizes are used. These centres would then not contribute to any estimate of treatment effect involving the third treatment arm in a FCE analysis, and so in situations where the number of patients per treatment arm in each centre is small, a RCE analysis may be preferable.

We simulated data based on the following six patient-centre combinations; 1) $c = 50$, $n = 150$; 2) $c = 100$, $n = 300$; 3) $c = 50$, $n = 300$; 4) $c = 100$, $n = 600$; 5) $c = 50$, $n = 450$; and 6) $c = 100$, $n = 900$. Scenarios 1–2 have an average of one patient per treatment per centre, scenarios 3–4 have an average of two patients per treatment per centre, and scenarios 5–6 have an average of three patients per treatment per centre. The number of patients in each centre for each scenario can be found in the online appendix.

### *4.2. Many patients per centre*

We simulated data based on the following six patient-centre combinations; 1) $c = 25$, $n = 250$; 2) $c = 50$, $n = 500$; 3) $c = 100$, $n = 1000$; 4) $c = 15$, $n = 375$; 5) $c = 25$, $n = 625$; and 6) $c = 50$, $n = 1250$. Scenarios 1–3 have on average 10 patients per centre, and scenarios 4–6 have on average 25 patients per centre. See the online appendix for the number of patients in each centre for each scenario.

### *4.3. Few overall centres*

We simulated data based on the following six patient-centre combinations; 1) $c = 5$, $n = 100$; 2) $c = 10$, $n = 200$; 3) $c = 5$, $n = 250$; 4) $c = 10$, $n = 500$; 5) $c = 5$, $n = 375$; and 6) $c = 10$, $n = 750$. Scenarios 1–2 have an average of 20 patients per centre, scenarios 3–4 have an average of 50 patients per centre, and scenarios 5–6 had an average of 75 patients per centre. Only a balanced distribution of patients across centres was used. See the appendix for the number of patients in each centre for each scenario.

### *4.4. Simple randomisation*

When simple randomisation (or any other method of randomisation that does not balance on centre) is used, valid inference can be obtained even if centre-effects are not accounted for in the analysis. However, when the centre-effects are associated with outcome (i.e. a non-zero ICC), an adjusted analysis will generally result in more power and precision. We therefore compared FCE, RCE, and unadjusted analyses after simple randomisation was used in order to determine which methods of analysis were preferable. These analysis methods were compared using all scenarios described in sections 4.1–4.3.

### *4.5. Sensitivity analyses*

*4.5.1. Very large number of patients per centre* We performed a set of simulations to compare FCE and RCE when there was a very large number of patients per centre. Simulations were performed as above. We used the following scenarios: 1) $c = 25$, $n = 2500$; and 2) $c = 100$, $n = 10,000$. Both scenarios had an average of 100 patients per centre. We used both balanced and skewed patient distributions for each scenario, and set the ICC to 0.05 for all simulations.

*4.5.2. Non-normal centre effects and centre outliers* We performed another set of simulations to determine how robust RCE are to departures from assumptions concerning the distribution of the centre-effects. Non-normal centre-effects for $u_j$ were simulated using a t-distribution with 3 degrees of freedom or a Chi-square distribution with 1 degree of freedom. Centre outliers were simulated using a normal distribution, and replacing the value of one of the centre-effects with five times the between-centre SD.

All simulations were generated in the same way as in section 4, except when generating non-normal centre-effects, the between-centre SD ($\sigma_j$) was held constant (based on the variance of the chosen distribution) and the patient level residual SD ($\sigma_i$) was chosen to give the desired ICC.

Only a subset of the previous centre-sample size combinations were used, and all ICCs were set to 0.05. We used the following scenarios: 1) $c = 50$, $n = 100$; 2) $c = 100$, $n = 400$; 3) $c = 25$, $n = 250$; 4) $c = 50$, $n = 1250$; 5) $c = 5$, $n = 100$;

  

and 6) $c = 10$, $n = 750$. We used both balanced and skewed patient distributions for all scenarios, apart from 5 and 6 where only a balanced distribution was used.

*4.5.3. When all centres have at least one patient on each treatment arm* In our simulations involving a small number of patients per centre, some centres would not recruit patients to both treatment arms, and would therefore be excluded from a FCE analysis. We performed another set of simulations to compare FCE and RCE when all centres recruited at least one patient to each treatment arm. We used 1) $c = 25$, $n = 100$; and 2) $c = 25$, $n = 250$. For scenarios 1 and 2, we set each centre to recruit 4 and 10 patients respectively. The block sizes were set to 6 and 18 respectively so that each centre would always recruit at least one patient from each arm. An ICC of 0.05 was used for all simulations.

*4.5.4. Simulations based on FCE* Until now, we generated centre-effects based on a RCE model. We performed a set of simulations to determine whether RCE models were robust to centre-effects based on a FCE model. Centre-effects were based on the MIST2 trial [9]. We generated 210 patients across 11 centres. The within-centre SD was set to 15.8. The number of patients in each centre can be found in the online appendix.

## 5. Simulation study results

### *5.1. Few patients per centre*

Eighteen scenarios (accounting for different ICCs, randomisation methods, and patient distributions across centres) were considered for all six combination of centres and patients (108 scenarios in all). Figure 1 shows the median, minimum, and maximum estimates for power, coverage, % bias in the SEs, and relative efficiency from these 18 scenarios for each centre-sample size combination.

The % bias in the estimated SE was small for both FCE and RCE, although RCE was biased downwards by approximately 2-3% in some scenarios. Coverage rates for FCE were nominal across all scenarios (range 94.3 to 95.5). RCE gave slightly lower than nominal coverage with two patients per centre (e.g. for c=50, n=100, the median coverage across 18 scenarios was 94.4, range 93.9 to 94.8). However, RCE with a DF correction provided closer to nominal coverage in these scenarios (e.g. for c=50, n=100, the median coverage across 18 scenarios was 95.0, range 94.5 to 95.4). With larger samples, RCE gave close to nominal coverage. Coverage rates were not affected by ICC, randomisation method, or patient distribution across centres.

Figure 2 shows power results across different scenarios for an ICC of 0.05. Using RCE (with or without a DF correction) gave greatly increased power compared to FCE in most situations. The median increase in power using RCE compared to FCE across all 108 scenarios was 13.9% (IQR 9.4 to 20.7%; range 2.1 to 30.2%). Differences were less extreme with larger sample sizes, an even patient distribution, and a block size of 4, though still favoured RCE. Differences in power between RCE with and without a DF correction were minimal. RCE was more efficient than FCE in all situations, although the difference was small with larger numbers of patients per centre, an even patient distribution, and a block size of four (median relative efficiency 1.16, range 1.02 to 1.43).

### *5.2. Few patients per centre (3 treatment arms)*

Figure 3 shows the median, minimum, and maximum estimates for power, coverage, % bias in SEs, and relative efficiency from across the 18 scenarios used for each centre-sample size combination.

FCE and RCE both gave unbiased estimates of the SE across all scenarios. Coverage rates were close to nominal for all analysis methods, although RCE with a DF correction did generally give slightly better results than RCE in scenarios with only one patient per treatment per centre.

Power results across different scenarios with an ICC of 0.05 are shown in Figure 4. RCE gave increased power compared to FCE (median increase across all 108 scenarios 5.3%; IQR 3.1 to 12.5%; range 0.4 to 18.4% ), although the difference was small for scenarios with a block size of four, an even patient distribution across centres, and larger numbers of patients per treatment per centre. Results for relative efficiency were similar to results for power, in that RCE gave better results in all scenarios, though the difference was negligible in some cases (median 1.06, range 1.01 to 1.32).

*5.3. Many patients per centre*

Figure 5 shows the median, minimum, and maximum estimates for power, coverage, % bias in the SEs, and relative efficiency from across the 18 scenarios used for each centre-sample size combination.

Both FCE and RCE gave unbiased estimates of the SE across all scenarios, and coverage results were nominal using all three analysis methods (FCE, RCE, and RCE with a DF correction).

RCE provided small gains in power (median increase 1.5%; range -0.1 to 4.7%), particularly with only 10 patients per centre (median 2.3%; IQR 1.6 to 3.2%). With 25 patients per centre the difference in power was minimal (median 0.5%; IQR 0.2 to 0.8%).

With 10 patients per centre, RCE were upwards of 5% more efficient than FCE, although with a small block size and even patient distribution the difference is negligible. With 25 patients per centre, RCE were approximately 2.5% more efficient, although in most scenarios the difference was negligible. There were no occurrences of FCE being more efficient than RCE.

*5.4. Few overall centres*

Both FCE and RCE gave unbiased estimates of the SE across all scenarios, and coverage was close to nominal using all three analysis methods, although a DF correction gave slightly improved coverage rates for RE. Power results were similar for all analyses methods, although RCE gave very slight improvements. Efficiency was similar for both FCE and RCE.

*5.5. Simple randomisation*

Each method of analysis gave unbiased SEs and correct coverage. RCE gave higher power compared to FCE and unadjusted analyses in all scenarios. Compared to FCE, RCE increased power by a median of 1.9% (IQR 1.1 to 3.4%) in trials with many patients per centre, and 0.8% (IQR 0.3 to 1.2%) in trials with a small number of centres. With a small number of patients per centre, results were similar to those in sections 5.2 and 5.1.

When the ICC was low (i.e. 0.01), the difference in power between RCE and an unadjusted analysis was negligible. In trials with many patients per centre, RCE led to a median increase in power of 1.2% (IQR 1.1 to 1.3%) with an ICC of 0.05, and 2.9% (IQR 2.3 to 3.4%) with an ICC of 0.10. Similar results were seen in all other scenarios.

*5.6. Sensitivity analyses*

*5.6.1. Very large number of patients per centre* There was no difference between FCE and RCE in terms of % bias in SEs, coverage, power, or efficiency.

*5.6.2. Non-normal centre effects and centre outliers* RCE were robust to both non-normal centre-effects and centre outliers. For a t-distribution with three degrees of freedom, SEs were unbiased across all 30 scenarios. Coverage rates were close to nominal, however improved when a DF correction was used (range 94.0 to 95.6% for RCE, and 94.5 to 95.7% for RCE with a DF correction). Likewise power and relative efficiency were unaffected (76.0 to 81.1% for RCE with a DF correction for power; 1.00 to 1.38 for relative efficiency).

Similar results were seen with a Chi-square distribution and for centre outliers (results not shown), indicating RCE are robust to assumptions concerning the distribution of centre-effects.

*5.6.3. When all centres have at least one patient on each treatment arm* Both FCE and RCE gave nominal coverage rates and unbiased SEs. RCE led to higher power than FCE in both scenarios (5.1% for 4 patients per centre, and 1.4% for 10 patients per centre.

*5.6.4. Simulations based on FCE* Both FCE and RCE gave good results for coverage and estimated SEs, although power was slightly higher with RCE (between 0.5 and 1.4% depending on the randomisation method).

## 6. Literature review

We carried out a literature review to assess how frequently multicentre trials are performed and the average number of patients per centre. When more than two treatment groups were used we summarised the average number of patients per

treatment per centre. For trials with a continuous primary outcome, we also assessed whether randomisation was balanced on centre, and if so, whether centre was adjusted for in the analysis and whether FCE or RCE models were used. We hand-searched The Lancet, BMJ, NEJM, and JAMA between January and December 2010 for reports of parallel group, individually randomised trials. We excluded cluster randomised, crossover, non-randomised, single-arm, and phase I or II trials. Articles reporting single centre studies, secondary analyses, interim analyses, or results that had been previously published in 2010 were also excluded, as were articles that did not state the number of centres involved. Although we recorded whether the primary outcome was continuous or not, we also assessed trials whose primary outcome was not continuous as we felt most of these trials would likely have continuous secondary outcomes.

After excluding single-centre trials ($n = 22$) and trials that did not report the number of centres ($n = 30$), we identified 176 eligible trials. The majority of trials had only two treatment groups, however 31 trials (18%) used three or more treatment groups.

The median number of centres was 20.5 (IQR 8 to 71). Twenty four trials (14%) had fewer than 5 centres, 32 (18%) had between 5 and 10 centres, 38 (22%) had 11–25 centres, 27 (15%) had 26–50 centres, 24 (14%) had 51–100 centres, and 31 (18%) had more than 100 centres.

The median number of patients per trial was 596 (interquartile range 300 to 1563), and only 5 trials (3%) had fewer than 100 patients. For trials with two treatment groups, the median number of patients per centre was 26.6 (interquartile range 11.1 to 80.0). Only three trials (2%) had an average of four or fewer patients per centre, although 31 trials (21%) had between 5 and 10 patients per centre. Twenty six trials (18%) had 11–20 patients per centre, 31 (21%) had 21–50, and the remaining 54 (37%) more than 51.

For trials with three or more treatment groups, the median number of patients per treatment group per centre was 3.9 (interquartile range 1.8 to 17.3). Thirteen trials (42%) had an average of three or fewer patients per treatment per centre. Six trials (19%) had 5–10 patients per treatment per centre, 5 (16%) had 11–20, 4 (13%) had 21–50, and the remaining 3 (10%) had 51 or more.

Forty six trials had a continuous primary outcome, 21 (46%) of which used centre as a balancing factor in the randomisation. The majority of trials that used centre as a balancing factor reported adjusting for centre in the analysis (13/21; 62%). Of the 13 trials that used an adjusted analysis, 10 (77%) reported using a FCE, one (8%) reported using RCE, one used a stratified rank-sum test, and one did not report the method of adjustment.

## 7. Discussion

Multicentre trials are common in practice, and pose unique challenges to the analysis of trials. If balanced randomisation is used to ensure treatment assignments are balanced within centres, then the centre-effects should be accounted for in the analysis. Failure to do so can result in SEs for treatment effect that are biased upwards, leading to confidence intervals that are too wide, and a reduction in power [2]. In many trials the ICC is small, and so an unadjusted analysis may lead to valid results, however this will not generally be known prior to trial commencement. Additionally, previous reviews [6, 7] have shown that a significant proportion of ICCs are large enough to affect coverage rates. Therefore, when centre is balanced on during randomisation, the protocol and statistical analysis plan should pre-specify that centre-effects will be accounted for in the analysis. If they are not accounted for in the analysis, this decision should be justified. We found that 38% of trials which used centre as a balancing factor in the randomisation and had a continuous outcome did not adjust for the centre-effects in the analysis, meaning that the results from these trials may be overly conservative.

We have used simulation to compare FCE and RCE as methods of adjusting for centre-effects in a variety of situations, and have accounted for different numbers of centres, different numbers of patients per centre, different ICCs, and differing patient distribution across centres. In total, we considered 378 scenarios.

In all scenarios, RCE were either superior to FCE, or the two methods were equivalent. There were no scenarios where FCE were preferable. The primary advantages of RCE was an increase in power and efficiency compared to FCE. These advantages were substantial in trials with a small number of patients per centre (e.g. reductions in power from 80% to 50-60% with 50 centres and 100 patients using FCE). RCE also performed better in scenarios where there was imbalance between treatment assignments within centres, which was the case with a skewed patient distribution or with a block size of 16. The increase in power and efficiency from RCE was in part due to FCE analyses excluding patients from centres that only recruited to one treatment arm. However, even when all centres recruited patients to both arms, RCE still increased power, likely due to the incorporation of between-centre information to the analysis.

Despite these advantages, the coverage with RCE was too low in scenarios with a small number of patients per centre. However, using RCE with a DF correction provided nominal coverage, and maintained advantages in power and efficiency compared to FCE. Therefore, when using RCE with small sample sizes, or with a small number of patients per centre, we recommend using a DF correction to ensure nominal coverage. Further research is required to determine which DF correction is best; although the simple correction used in our simulations is easy to implement and gave good empirical results, there is (to our knowledge) no theoretical basis to show it will give good results in scenarios outside of those studied in this paper. One procedure which does have a theoretical basis is the Kenward-Roger DF correction, which not only corrects the DF used, but also adjusts the estimated SE. Given we found estimated SEs to be too low for RCE with very few patients per centre, the Kenward-Roger DF correction may be ideal in these circumstances, particularly in trials with fewer than 100 patients. This DF correction is however only currently available in SAS.

Although RCE were most beneficial with a small number of patients per centre, these scenarios were relatively rare for trials with only two treatment arms. Our literature review showed that only 2% of trials with two treatment arms had four or fewer patients per centre. However, for trials with three or more treatment arms, the median number of patients per treatment in each centre was 3.9, and 42% of trials had three or fewer patients per treatment in each centre. Our simulations indicate that RCE may provide substantial benefits in these trials, and in trials with up to 10 patients per centre in the two-arm setting. We found that 21% of trials had between 5–10 patients per centre, and so RCE could also be beneficial in these scenarios.

RCE also offer advantages with skewed patient distributions across centres, as this will inevitably lead to more unbalanced treatment assignments within smaller centres. It is unclear how often balanced or skewed patient distributions occur in practice, and it is likely that in many trials the patient distribution lies somewhere between the two scenarios we have used in our simulations. However, in our experience, trials with a large number of centres tend more towards the skewed patient distribution. There are several potential reasons for this: (1) centres usually start recruiting at different times throughout the trial, so some will have a longer time frame for recruitment; (2) centres vary in size, and some centre may recieve more eligible patients; and (3) some centres may be more enthuasiastic about the trial than others, and may put more effort into recruitment.

One concern about RCE is it makes additional assumptions regarding the distribution of centre-effects. However, we found RCE to be robust to non-normal centre-effects and to centre outliers. Neither extreme skewness or heavy tails in the centre-effects had any impact on results.

When there was a small number of centres (i.e. 5 or 10) or a large number of patients per centre (25 or more), RCE and FCE performed equally well in terms of coverage, power, and efficiency, regardless of patient distribution, block size, or ICC. Fifty eight per cent of trials had more than 20 patients per centre, indicating that in the majority of trials, either RCE or FCE are adequate choices.

When centre has not been balanced on during randomisation (e.g. if simple randomisation or permuted blocks ignoring centre has been used), ignoring centre-effects in the analysis will still lead to valid results. However, when the ICC is non-zero, an analysis that adjusts for the centre-effects will be more powerful, and as the size of the ICC increases, so will the amount of power that can be gained from an adjusted analysis. RCE are superior to FCE and unadjusted analyses in most scenarios, however when the ICC is very low, unadjusted analyses and RCE give similar results. Therefore, if the ICC is expected to be moderate or large, we recommend adjusting for centre-effects using RCE. If the ICC is expected to be small, either RCE or an unadjusted analysis can be used. This decision should be pre-specified prior to analysis.

One interesting issue is the implication of FCE when some centres contain only one patient. These patients will be dropped from the analysis entirely which goes against the intention-to-treat principle, which dictates that all patients are included in the analysis. Additionally, the ethical implications of this are unclear. It is recognised that patients who enter a trial do not necessarily benefit themselves (as the treatment they receive could be less beneficial than standard care, or even harmful), but their information will be used to benefit future patients by helping to inform patient care. Asking a patient to participate in a study that could prove harmful, but at the same time potentially denying them the opportunity for their experience to benefit future patients if they happen to present to a centre where they are the only patient recruited, is difficult to justify. RCE avoid this issue, as all patients are included in the analysis, regardless of whether they were the only patient recruited to their centre.

Our simulation study only considered the case where there was one follow-up measurement per patient. However, when continuous outcomes are used, the patient often has several follow-up measurements. A mixed-effects model is generally used to take into account the correlation between measurements from the same patient. This method of analysis

  

is particularly useful in the case where patients may have outcomes observed at some, but not all, timepoints. Accounting for centre-effects using RCE would require a three-level random-effects model, which could in some circumstances be computationally difficult. It is unclear in this scenario whether RCE or FCE is superior.

We have not considered the issue of treatment-by-centre interaction in our simulations. ICH E9 [1] suggests that questions of treatment-by-centre interaction should not be included in the primary analysis, but should rather be regarded as exploratory, as with other subgroup analyses. We have therefore focused on scenarios which reflect the primary analysis and have not compared FCE and RCE models when accounting for treatment-by-centre interactions.

Given the wide variety of scenarios in which RCE provide substantial benefits over FCE, we recommend they be used more frequently in practice. In particular, RCE should be the default option when the number of patients per centre is small (i.e. $<10$), or when there are likely to be imbalances in treatment assignments within centres, either due to large block sizes, or to a skewed patient distribution where many centres have a small number of patients.

## Acknowledgements

We thank the MIST1 and MIST2 trial teams for the use of their data. We are grateful to Dan Bratton, Rachel Jinks, and Sunita Rehal for their helpful comments on the manuscript. We would also like to thank two anonymous reviewers whose comments helped to improve the article.

## References

1. ICH E9 Expert Working Group. Statistical principles for clinical trials: ICH harmonised tripartite guideline. *Statistics in Medicine* 1999; **18**:1905–1942, doi:10.1002/(SICI)1097-0258(19990815)18:15%3C1903::AID-SIM188%3E3.0.CO;2-F. URL http://dx.doi.org/10.1002/(SICI)1097-0258(19990815)18:15%3C1903::AID-SIM188%3E3.0.CO;2-F.
2. Kahan BC, Morris TP. Improper analysis of trials randomised using stratified blocks or minimisation. *Statist. Med.* 2012; **31**(4):328–340, doi:10.1002/sim.4431. URL http://dx.doi.org/10.1002/sim.4431.
3. Tangri N, Kitsios GD, Su SH, Kent DM. Accounting for Center Effects in Multicenter Trials. *Epidemiology* Nov 2010; **21**(6):912–913, doi:10.1097/EDE.0b013e3181f56fc0. URL http://dx.doi.org/10.1097/EDE.0b013e3181f56fc0.
4. Pickering RM, Weatherall M. The analysis of continuous outcomes in multi-centre trials with small centre sizes. *Statist. Med.* Dec 2007; **26**(30):5445–5456, doi:10.1002/sim.3068. URL http://dx.doi.org/10.1002/sim.3068.
5. Parzen M, Lipsitz SR, Dear KBG. Does clustering affect the usual test statistics of no treatment effect in a randomized clinical trial? *Biometrical Journal* 1998; **40**(4):385–402, doi:10.1002/(SICI)1521-4036(199808)40:4%3C385::AID-BIMJ385%3E3.0.CO;2-%23. URL http://dx.doi.org/10.1002/(SICI)1521-4036(199808)40:4%3C385::AID-BIMJ385%3E3.0.CO;2-%23.
6. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology* Aug 2004; **57**(8):785–794, doi:10.1016/j.jclinepi.2003.12.013. URL http://dx.doi.org/10.1016/j.jclinepi.2003.12.013.
7. Cook J, Bruckner T, MacLennan G, Seiler C. Clustering in surgical trials - database of intracluster correlations. *Trials* Jan 2012; **13**(1):2+, doi:10.1186/1745-6215-13-2. URL http://dx.doi.org/10.1186/1745-6215-13-2.
8. Maskell NA, Davies CW, Nunn AJ, Hedley EL, Gleeson FV, Miller R, Gabe R, Rees GL, Peto TE, Woodhead MA, *et al.*. U.K. Controlled trial of intrapleural streptokinase for pleural infection. *The New England journal of medicine* Mar 2005; **352**(9):865–874, doi:10.1056/NEJMoa042473. URL http://dx.doi.org/10.1056/NEJMoa042473.
9. Rahman NM, Maskell NA, West A, Teoh R, Arnold A, Mackinlay C, Peckham D, Davies CWH, Ali N, Kinnear W, *et al.*. Intrapleural Use of Tissue Plasminogen Activator and DNase in Pleural Infection. *N Engl J Med* Aug 2011; **365**(6):518–526, doi:10.1056/NEJMoa1012740. URL http://dx.doi.org/10.1056/NEJMoa1012740.
10. Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for Center in Multicenter Studies: An Overview. *Annals of Internal Medicine* Jul 2001; **135**(2):112–123. URL http://www.annals.org/content/135/2/112.abstract.
11. Senn S. *Cross-over Trials in Clinical Research*. Wiley, Chichester, 1993.
12. Jones B, Kenward MG. *Design and Analysis of Cross-Over Trials*. 2 edn., Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Chapman & Hall/CRC, 2003. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0412606402.
13. Senn S. Some controversies in planning and analysing multi-centre trials. *Statist. Med.* 1998; **17**(15-16):1753–1765, doi:10.1002/(SICI)1097-0258(19980815/30)17:15/16%3C1753::AID-SIM977%3E3.0.CO;2-X. URL http://dx.doi.org/10.1002/(SICI)1097-0258(19980815/30)17:15/16%3C1753::AID-SIM977%3E3.0.CO;2-X.
14. Jones B, Teather D, Wang J, Lewis JA. A comparison of various estimators of a treatment difference for a multi-centre clinical trial. *Statist. Med.* 1998; **17**(15-16):1767–1777, doi:10.1002/(SICI)1097-0258(19980815/30)17:15/16%3C1767::AID-SIM978%3E3.0.CO;2-H. URL http://dx.doi.org/10.1002/(SICI)1097-0258(19980815/30)17:15/16%3C1767::AID-SIM978%3E3.0.CO;2-H.
15. Rabe-Hesketh S, Skrondal A. *Multilevel and Longitudinal Modeling Using Stata*. 2 edn., Stata Press, 2008. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/1597180408.

16. White IR. simsum: Analyses of simulation studies including Monte Carlo error. *Stata Journal* 2010; **10**(3):369–385.

**Table 1.** ICC estimates and 95% CIs from the MIST 1 and 2 trials

| Outcome | MIST1 | MIST2 |
|---|---|---|
| Mortality at 3 months | 0.02 (<0.001 to 0.88) | <0.001 (<0.001 to 1) |
| Overall mortality | 0.13 (0.01 to 0.58) | <0.001 (<0.001 to 1) |
| Surgery at 3 months | 0.05 (0.01 to 0.34) | 0.03 (<0.001 to 0.73) |
| Overall surgery | 0.02 (0.001 to 0.36) | 0.02 (0.001 to 0.52) |