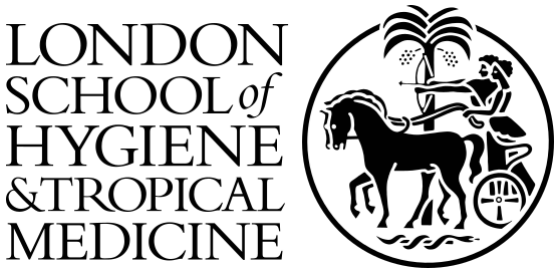

Bias and inference from misspecified mixed effect models in stepped wedge trial analysis

Jennifer A Thompson[a,b], Katherine L Fielding[a], Calum Davey[c], Alexander M Aiken[a], James R Hargreaves[c], Richard J Hayes[a]

[a] Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, U.K.

[b] MRC London Hub for Trials Methodology Research, London, UK

[c] Department of Social and Environmental Health Research, London School of Hygiene and Tropical Medicine, London, UK

Funded by MRC hub for trials methodology research.

Short title: Thompson et al

Abstract

Many stepped wedge trials (SWTs) are analysed using a mixed-effect model with a random intercept and fixed-effects for intervention and time-periods (referred to here as the standard model). However, it is not known whether this model is robust to misspecification.

We simulated SWTs with three groups of clusters and two time-periods; one group received the intervention during the first period, and two groups in the second period. We simulated period and intervention effects that were either common-to-all or varied-between clusters. Data were analysed with the standard model, or with additional random effects for period effect or intervention effect. In a second simulation study, we explored the weight given to within-cluster comparisons by simulating a larger intervention effect in the group of the trial that experienced both the control and intervention conditions and applying the three analysis models described previously.

Across 500 simulations we computed bias and confidence interval coverage of the estimated intervention effect.

We found up to 50% bias in intervention effect estimates when period or intervention effects varied between clusters and were treated as fixed effects in the analysis. All misspecified models showed under-coverage of 95% confidence intervals, particularly the standard model. A large weight was given to within-cluster comparisons in the standard model.

In the SWTs simulated here, mixed effect models were highly sensitive to departures from the model assumptions, which can be explained by the high dependence on within-cluster comparisons. Trialists should consider including a random effect for time-period in their SWT analysis model.

Keywords: Stepped Wedge Trials, Cluster Randomised Trials, Mixed effect model, model misspecification, simulation study

Introduction

Recent reanalysis of a high-profile stepped wedge trial (SWT) has brought into question methods commonly used to analyse these complex studies [1-3]. SWTs are often analysed using models that make strong assumptions about the clustering in the data [4]. It is currently unknown if estimates from these models are robust to deviations from these assumptions.

An SWT is a type of cluster randomised trial where clusters are randomised into groups. Each group begins to receive the intervention at a different time so that all clusters start the trial in the control condition and by the end of the trial all clusters are receiving the intervention.

The control and intervention conditions can, in principle, be compared in two directions known as the vertical and horizontal comparisons [4]. Vertical comparisons compare the outcomes of clusters in the intervention condition with the outcomes of clusters in the control condition within the same time-period; since the order of rollout is randomised, each of these comparisons is randomised. Horizontal comparisons compare outcomes from periods in the intervention condition with outcomes from periods in the control condition in the same cluster; these are non-randomised before-after comparisons that are confounded with time-period.

In practice, most analysis methods for SWTs incorporate information from both the vertical and horizontal comparisons in the intervention effect estimate and so need some way to adjust for period effects [4]. The most common analysis model (hereafter referred to as the standard model) is a mixed effect model with a random intercept to account for clustering and adjusting for period effects as a fixed categorical variable; this model is described by Hussey and Hughes [5]. Despite its wide use, guidance for using this analysis model is lacking. The model makes strong assumptions about the correlation structure of the data: the intervention effect and the period effects are assumed to be common to all clusters. It is not currently known whether the intervention effect estimate and its precision are robust to misspecifying these assumptions.

In the context of SWTs, we are most interested in estimation of the intervention effect and how robust this effect is to misspecification of the intervention effect itself as well as misspecification of the period effect. Previous research has found that misspecifying the random effects led to biased effect estimates as well as biased precision of estimates [6]. In parallel cluster randomised trials with baseline measurements and in cluster crossover randomised trials it has been shown that analyses with hierarchical models should include a random effect for period, sometimes referred to as a cluster-period interaction, to avoid residual confounding [7-10].

The importance of specifying the period effect correctly will depend on how much the horizontal comparisons contribute within the model. This has not been explored in the literature. If a large weight is given to this comparison, any residual confounding of the intervention effect by the period effects could lead to a biased estimate of the intervention effect.

In this paper, we will explore both issues with a simulation study comparing the standard model to other mixed effect models, focusing on a binary outcome with cross-sectional measurements. We then run a second set of simulations to explore the weight given to horizontal comparisons by each analysis model. Following the simulation studies, we explore the impact of misspecifying analysis models in our motivating example.

Motivating example

There has been much debate in recent literature about the results of a reanalysis of a highly cited SWT that investigated the effect on school attendance of a mass deworming intervention for school children in Kenya [1-3]. The trial included 75 schools (clusters) that were randomised into three groups and ran over two years. School attendance was measured as binary outcome with multiple observations for each individual child during each year. There was a geometric mean of 1,180 (interquartile range [IQR] 908.5, 1,864) observations in each school each year, with the attendance assessed on the same children in year two as year one. Children from schools in the first group began receiving the intervention at the start of the first year. Children from schools in the second group received no intervention during the first year and began receiving the intervention in the second year of the study. Children from schools in the third group did not receive the intervention during these two years (Figure 1).

In the reanalysis of this trial, it was found that the odds ratios for school attendance for year one and year two were both smaller when analysed individually (odds ratio (OR)=1.48 and 1.23 respectively) than the odds ratio given by the standard model when the data were pooled from both years (OR=1.82) [2]. We hypothesised that this could have been because the analysis model was misspecified and explored two potential types of misspecification:

1. The period effects varied between clusters. The standard model assumes that the period effects are common to all clusters. This could lead to a biased estimation of the intervention effect through biased estimation of the period effects.
2. The intervention effect varied between clusters. The standard model assumes that the intervention effect is common to all clusters. Treating an effect which truly varies as a fixed effect has been shown to lead to biased estimation of a covariate [6] and so the estimate of the intervention effect could be biased.

In this paper, we first used a simulation study based on the motivating example to explore the effect of ignoring variability between the clusters in the period effect and intervention effect in the analysis of SWTs. Secondly, we hypothesised that the effect of misspecification would by highly influenced by the weight given to horizontal comparisons in each analysis model and so also performed a further set of simulations to investigate this question. We then analysed the motivating example with different analysis models and compared the results in light of the findings of the simulation studies.

Simulation study methods

Simulation study 1

To investigate the impact of ignoring heterogeneity between clusters in the period effect and intervention effect, we compared analysis models which assumed these effects were common to all clusters (the standard model) to analysis models which allowed these effects to vary between clusters. We performed this with data in which the true underlying period effect and intervention effect were either common to all clusters or varied between clusters. A description of the scenarios we used to compare the analysis models is given, followed by the three analysis models we compared. A summary of the data scenarios simulated is given in Table 1.

We used the same trial design as our motivating example with clusters randomised into three groups and followed for two time-periods. During the first period, only the first group had received the intervention, and during the second time-period, the first and second groups had received the intervention. The third group never received the intervention. This trial design was chosen due to its simplicity; since there are only two time-periods, the period effect is simple to model. The horizontal comparison is only possible in one group; this allowed us to explore the weight given to this comparison. To mimic the motivating example and to avoid issues with small sample size, we assigned 25 clusters to each group and the number of observations in a cluster in each time-period was drawn from a log–normal distribution ($\mu = 6.9$, $\sigma = 0.74$); this gave a geometric mean number of observations in each cluster in each time-period of 1,027 (IQR 669, 1,798).

The cluster-level distribution of the outcome in the first period and the change from period one to period two (the period effect) was based on group 3 of the motivating example. This group was chosen because it did not receive the intervention. We modelled the log-odds in the first period and the log-odds ratio period effect from the motivating example as a bivariate normal distribution. This gave mean values for the log-odds in period one and log-odds ratio period effect, together with a 2x2 covariance matrix. This distribution described the outcome and how it varied between the clusters in each period. The mean values were used in all the simulation scenarios, but we manipulated the covariance matrix to create four scenarios of how the outcome varied between the clusters and periods (Figure 2). The mean odds in the first period was 6.61 (a proportion of 87%), and the mean odds ratio period effect between the second and first period was 0.32, which was equivalent to an odds of 2.12 (proportion of 68%) in the second period. The covariance matrices for each of the four scenarios are given in supporting information S1 and are described below:

(1) *Common period effect, high variability:*
The period effect was common to all clusters with between-cluster variance = 1.81. This was the amount of between-cluster variability observed in year one of the motivating example. This represents a simple scenario with a large intracluster correlation coefficient (ICC=0.20), where the standard model would have a correctly specified period effect.

(2) *Common period effect, low variability:*
The period effect was common to all clusters with between-cluster variance = 0.25. This was the amount of between-cluster variability observed in year two of the

motivating example. This represents a simple scenario with a lower ICC (ICC=0.05), where, again, the standard model would have a correctly specified period effect.

(3) *Varying period effect, decreasing variability:*
The period effect varied between clusters with the variability between the clusters decreasing from the first period to the second period. The initial between-cluster variance was 1.81 and the period effect variance was 1.89. The decrease in variability from period one to period two resulted from a negative covariance between the initial value and the period effect of -1.72. This complex scenario reflects the underlying trends seen in the motivating example. In this scenario, the standard model would have a misspecified period effect.

(4) *Varying period effect, stable variability:*
The period effect varied between the clusters but the between-cluster variance remained the same for both periods. Here the initial between-cluster variability and period effect variability remained the same as in scenario (3) but the covariance was reduced to -0.94. This scenario was chosen to assess the effect of a varying period effect without the additional complication of the between-cluster variation reducing in the second period. In this scenario, the standard model would have a misspecified period effect.

We simulated two scenarios for the intervention effect; these were not based on the motivating example:

A. An intervention effect that was common to all clusters. We simulated an intervention effect log(OR)=0.41 (equivalent to OR=1.5) for all clusters. We also simulated log(OR)=0 to calculate the type I error rate. In these scenarios, the standard model would have a correctly specified intervention effect.
B. An intervention effect that varied between clusters drawn from the distribution $log(OR) \sim N(0.41, 0.3)$. This gave a geometric mean OR=1.5 with an IQR=1.05-1.97. We also simulated a distribution $log(OR) \sim N(0, 0.3)$ to calculate the type I error rate. In these scenarios, the standard model would have a misspecified intervention effect.

The variation in the intervention effect was modelled as being independent of the underlying outcome and period effect between-cluster variability. This meant that the intervention effect varying between clusters would lead to increased variability between the clusters in period two as more clusters were receiving the intervention in this period.

Each scenario led to the odds of the outcome occurring in each cluster-period. From this, the observations within each cluster-period were sampled from a binomial distribution, assuming independence within each cluster-period. This assumes a cross-sectional design and is a deviation from the motivating example, where children were observed multiple times during the study, chosen for simplicity.

All combinations of these parameters were simulated.

Simulation study 2

Secondly, we hypothesised that the horizontal comparisons would depend on the model assumptions more heavily than the vertical comparisons. To aid interpretation of the results of simulation study 1, we sought to investigate the contribution of the horizontal comparisons to each analysis in each scenario.

In the trial design used for this paper, only group 2 contributed horizontal comparisons because groups 1 and 3 remained in the same condition for both periods of the study (see Figure 1). This meant that we could investigate the weights given to the horizontal and vertical comparisons by identifying how much weight was given to group 2 relative to groups 1 and 3.

To do this, we re-ran the simulations but with an intervention effect log(OR)=1.5 in group 2 of the trial but kept an intervention effect in group 1 of log(OR)=0.41. An unbiased intervention effect estimate from horizontal comparisons alone would have expectation $E\big(log(OR)\big) = 1.5$. An unbiased intervention effect estimate from vertical comparisons alone would have expectation $0.41 < E\big(log(OR)\big) < 1.5$ depending on the weights given to each cluster and to periods one and two of the trial. Comparing the intervention effect estimates of each model in each scenario to the horizontal comparison $E\big(log(OR)\big) = 1.5$ allowed us to see how much the horizontal comparisons contributed to the analysis compared to the vertical comparisons. Such a large imbalance in the intervention effect between groups is, of course, unlikely (although not impossible); this simulation study was designed to investigate the contributions of vertical and horizontal comparisons, rather than to explore a realistic scenario.

Analysis Models

Each simulated data set was analysed with three analysis models, each making different assumptions about the period effect and intervention effect.

Standard model:

First, we used the standard method of analysis [4, 5]: a mixed effect logistic regression with a random intercept and fixed effects for intervention effect and period effect:

$$y_{ijk} = \mu + \beta Z_j + \theta X_{ij} + u_i \qquad (1)$$

where y_{ijk} is the log odds of the outcome in cluster *i* in year *j* for observation *k*, μ is the mean log odds of the outcome in period one in the control condition, β is the period effect log odds ratio comparing the outcome in periods two and one, $Z_j$is an indicator of year; 0 for the first year and 1 for the second year, θ is the intervention effect log odds ratio, and X_{ij} is an indicator of whether cluster *i* received the intervention in year *j*, $u_i \sim N(0, \sigma_u^2)$ is a random intercept allowing for variability in the outcome between clusters.

This model assumes the period effect and the intervention effect are common to all clusters so is a misspecified model in scenarios where either the period effect or intervention effect varied between clusters.

Random period model:

Second, we added a random effect for period to the standard model:

$$y_{ijk} = \mu + (\beta + v_i)Z_j + \theta X_{ij} + u_i \quad (2)$$

where and $\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{u,v}^2 \\ \sigma_{u,v}^2 & \sigma_v^2 \end{pmatrix}\right)$ are a random intercept and random effect for period respectively.

This model assumes that the intervention effect is common to all clusters but allows the period effect to vary between clusters. It is a misspecified model in scenarios where the intervention effect varies between the clusters.

Sometimes, other literature has used a different model to allow the period effect to vary between the clusters [11, 12]. For details on how these models relate to one another, see supporting information S2

Random intervention model:

Third, we added a random effect for the intervention to the standard model:

$$y_{ijk} = \mu + \beta Z_j + (\theta + z_i)X_{ij} + u_i \quad (3)$$

where $\begin{pmatrix} u_i \\ z_i \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{u,z}^2 \\ \sigma_{u,z}^2 & \sigma_z^2 \end{pmatrix}\right)$ are a random intercept and random effect for intervention respectively.

This model assumes that the period effect is common to all clusters but allows the intervention effect to vary between clusters. The model is a misspecified model in scenarios where the period effect varies between the clusters.

Whilst the random period and random intervention models allow for variability in the period and intervention effect respectively, they can estimate a variability of close to zero if the effect is common to all clusters. The random period model is correctly specified in the scenario with common period effect, and likewise the random intervention model is correctly specified in the scenario with common intervention effect. Similarly, the random intervention model allows for a covariance between the intervention effect and the intercept ($\sigma_{u,z}^2$) but allows this covariance to be zero, as is the case in our simulation study.

Estimands and performance measures

We ran 500 simulations for each combination of parameters. This allowed us to estimate the intervention effect to within 5% accuracy assuming a variance estimate of 0.05. This variance is conservative as it is larger than the estimated variance we saw in the motivating example.

From the analysis models, we collected the estimated fixed effects, their standard errors, and the estimated between-cluster covariance matrix.

We calculated the mean, standard deviations, 95% confidence intervals, and the interquartile range of the intercept, intervention effect, and period effect estimates from the 500 simulations. We calculated percentage bias as:

$$percentage\ bias = \left(\frac{\bar{\hat{\beta}} - \beta}{\beta}\right)$$

Where β is the true effect and $\bar{\hat{\beta}}$ is the mean of the effect estimates.

We calculated the coverage of 95% confidence intervals as the proportion of simulations with the true effect contained within the 95% confidence interval of the estimate. We calculated the type 1 error rate as the proportion of simulations with true OR=1 with $p<0.05$ against a null of the intervention effect OR=1.

In the set of simulations with a different intervention effect in group 2 (Simulation study 2), we compared the mean of the intervention effect estimates to the horizontal intervention effect comparison of log(OR)=1.5.

Simulations were run in R version 3.2, the lme4 package was used for mixed effect models.

Results

Model convergence

The standard model converged in all simulations for both simulation studies. When either the period effect or the intervention effect varied between clusters the random period and random intervention models also converged in >99% of all simulations. However, when both period effect and intervention effect were common to all clusters, the random period model failed to converge in 3% to 9% of simulations, and the random intervention model failed to converge in 4% to 33% of simulations. Estimates from these models were excluded from performance statistics. Further details of convergence of the models is given in supporting information S3.

Simulation study 1 results

Bias of fixed effect estimates

Figure 3 gives the mean and IQR of intervention effect estimates for each scenario. A table of the mean values is given in supporting information S4.

Where there were common period and intervention effects, all three models performed similarly, with estimation of the intervention effect in line with the true underlying effect.

Where the period effect varied between the clusters, only the random period model gave unbiased estimates of the intervention effect. Depending on the scenario, the standard model had between -20% and -8% bias and the random intervention model between -51% and -8% bias. Bias was larger when the period effect varied with decreasing variability than with stable variability but was similar regardless of whether there was a common or varying intervention effect. We also observed bias in the period effect estimates and intercept

estimates from the standard model and random intervention model (supporting information S5 and S6).

Where the intervention effect varied between the clusters and there was a common period effect, the random intervention model and the random period model gave unbiased estimates of the intervention effect. Only the standard model intervention effect estimates had substantial bias (-9% and -16% bias for common period effects with high and low variability respectively).

Where either the period effect or intervention effect varied between clusters, the standard model intervention effect estimates had greater variability compared to the random period model or random intervention model. Differences were larger when the period effect varied between clusters than when the intervention effect varied between clusters. For example, the standard model intervention effect estimates were 3.6 times as variable as the random period model estimates when the period effect varied between clusters with decreasing variability with common intervention effect. Whereas, the standard model intervention effect estimates were 1.5 times as variable as the random intervention model estimates when the intervention effect varied between clusters with common period effect with high variability.

Standard errors, coverage, and type 1 error

In scenarios with a common period and intervention effects, 95% coverage was maintained regardless of the analysis model and the estimated standard errors were similar across analysis models (Figure 4 and supporting information S7 and S8).

When period effect or intervention effect varied between clusters, the standard model gave standard errors that were markedly smaller than the random period model and random intervention model. The mean intervention effect standard error from the standard model was less than 0.33 and 0.26 times the mean standard error of the random period model and random intervention model respectively.

The inappropriately small standard errors given by the standard model were in part explained by downward bias in the estimation of between-cluster variability (see supporting information S9). For example, when variability was stable over the two time-periods with a variance of 1.79 the standard model estimated the variance as 1.26.

The bias in estimates, standard errors, and increased variability in estimates led to under-coverage of the 95% confidence intervals of the intervention effect estimates (Figure 4). For the standard model, under-coverage was severe when either the intervention effect or the period effect varied between clusters (<25% coverage). Similarly, the random intervention model had under-coverage when the period effect varied between clusters (74% and 88% coverage for decreasing and stable variability respectively) regardless of intervention effect variability. Finally, the random period model had under-coverage of confidence intervals when the intervention effect varied between clusters with a common period effect (86% and 88% coverage for common period effect with high and low variability respectively).

Type 1 error rates followed the same patterns as coverage (supporting information S10).

Simulation study 2 results

Figure 5 gives the estimated log odds ratios for each scenario where the group 1 and 2 intervention effects differed ($log(OR) = 0.41$ in group 1 and $log(OR) = 1.5$ in group 2). All analysis models gave a mean estimated intervention effect close to the group 2 effect when there was a common period effect and a common intervention effect; this was the case in the high and low variability scenarios. This suggests that in these scenarios the intervention effect is largely estimated from horizontal within-cluster comparisons in group 2; groups 1 and 3 appeared to contribute to estimation of the period effect but had little influence on the intervention effect estimate.

The standard model estimates remained close to the group 2 intervention effect in all scenarios. The downward bias we observed in our first set of simulations suggests that at least some of the movement away from the group 2 effect is because of bias and not because of a reduction in the contribution of the horizontal comparisons. This implies that the standard model was continuing to estimate the intervention effect largely from horizontal comparisons in group 2.

In contrast, when the period effect varied between clusters, the random period model gave intervention effect estimates much further from the horizontal comparison estimates. This implies that the horizontal comparisons in group 2 could not contribute as much information to the analysis because there was less certainty about separating the period effect and intervention effect in these comparisons. This was similar in the scenarios where the intervention effect varied between clusters but the period effect was common for both the random period and random intervention models, but to a smaller degree.

Example

For our motivating example, we hypothesised that the standard model gave a larger intervention effect than either of the two time-periods analysed separately because the standard model was misspecified, either by ignoring variability in the period effect or the intervention effect. Our simulation study suggests that this is not the case because we would expect the standard model to underestimate the intervention effect with these types of misspecification, rather than overestimate the effect. However, we also found that the standard model gave a very large weight to the horizontal comparisons. This does help to explain the counterintuitive results seen in the motivating example [2].

We reanalysed the deworming trial using the three analysis models investigated in the simulation study, and additionally looked at year one and two separately using a mixed effect model with a fixed effect for intervention and a random intercept to attain estimates for the intervention effect from vertical comparisons. In line with the published reanalysis of this study, we ignored pupil-level clusters from multiple observations of the same pupils; this is in line with research suggesting it is sufficient to adjust for the highest level of clustering alone, known as passing the buck [13].

The results, in Table 2, are different from the published reanalysis as we have used a different version of the data (see supporting information S11 for details) and have not adjusted for covariates other than period [14].

We found that the standard model combining data from both years of the study gave a larger estimate of the intervention effect than either year analysed separately, which is as was found in the reanalysis [4].

Adjusting for variation between clusters in the period effect or intervention effect (i.e. using either the random period model or random intervention model) increased the intervention effect standard error and reduced the intervention effect towards the null. Both approaches gave an intervention effect estimate between the estimated effect in year one and year two. This suggests that the horizontal comparisons are contributing less to these analysis models than to the standard model; this is consistent with the findings of our second simulation study into the contribution of the horizontal comparisons.

The random period model found strong evidence of variability in the period effect ($p<0.001$), and the random intervention model found strong evidence of variability in the intervention effect ($p<0.001$). Since the period effect and intervention effect are confounded with one another, evidence of variability in the intervention effect could be caused by variability in the period effect or vice versa. The random period model estimated a between-cluster covariance matrix similar to the simulation study scenario with varying period effect with decreasing variability. The random intervention model estimated lower variability between clusters in the intervention condition than in the control condition because of the reduced variability in year two. This is a scenario that we did not consider in our simulation study where we only investigated a scenario with greater variability in the intervention condition. Inspection of the data suggests that the random period model is the most appropriate one. A mixed effect model with a random effect for period run on observations from group 3, which never received the intervention, finds strong evidence of variability in the period effect ($p<0.001$). But, a mixed effect model with a random effect for intervention run on observations from groups one and three, where the intervention effect is not confounded with the period effect, finds no evidence of variability in the intervention effect ($p=0.34$).

The random period model suggests that there is some evidence that the deworming intervention increased school attendance (OR=1.26, 95% CI 1.02, 1.57; $p=0.03$). The effect found using this model is weaker, both in terms of absolute size and level of statistical significance, than the effect found using the standard model. There are still limitations in these data and this analysis, on which further information has been published elsewhere [1-3].

Discussion

We found biased estimates and serious under-coverage of confidence intervals in the SWT scenarios we simulated when the analysis model ignored variability between clusters in the period effect or intervention effect. In these scenarios, results from the standard model were driven largely by the horizontal comparisons.

We have shown that, in the scenarios we considered, misspecifying the random effects of mixed effect models can result in biased intervention effect estimates. The standard model underestimated the intervention effect when either the period or the intervention effect varied between the clusters. The underestimation when the period effect varied may result from the standard model estimating an intervention effect averaged over the two periods, whereas the true effect for this scenario was a within-period intervention effect. This is analogous to the difference between the population-averaged effect and the cluster-specific effects that are given by different analysis methods. In the presence of intervention effect variability the standard model also gave biased estimates of the intervention effect. The random intervention model had even larger bias when it was misspecified than the standard model. Conversely, the random period model had only negligible bias in estimates in all scenarios we considered. These results are consistent with previous research into misspecifying mixed effect models in cluster randomised trials [7, 9]. We have built on this literature and shown that these results extend to SWTs. This highlights how sensitive mixed effect models can be to misspecification of model assumptions.

Caution is needed beyond estimation of the intervention effect itself. In our simulation study, the bias extended to standard errors and between-cluster variability. The latter has implications for reporting the ICC, as recommended by the CONSORT guidelines [15]. In addition to the implications for inference, the bias in standard errors has implications for determining the power and sample size of SWTs. Since the standard error from the standard model is used in most current methods of SWT sample size calculations [12, 16-18], they should not be applied when the period effect or intervention effect are expected to vary between clusters, at least in relation to the characteristics of the trial exemplar used in this paper. Instead, the method developed by Hooper et al may be more appropriate [11].

The result of these biases was under-coverage of confidence intervals for the intervention effect. If model assumptions do not hold, we risk being over-confident in our conclusions. We found particularly severe under-coverage when using the standard model. This has been seen in previous research into misspecified random effects [6, 19] and has recently been seen in the setting of SWTs [20]. This is reflected in our analysis of the motivating example; we see a large increase in the standard error of the intervention effect, and so confidence intervals are much wider when moving from the standard model to the random period model or random intervention model.

The results from our simulation study could be explained by the excessive weight given to the horizontal comparisons, even with a lower ICC=0.05. Because the horizontal comparisons are within-cluster comparisons, they avoid the additional variability of between-cluster variation. This means that if the period and intervention effects can be separated, the horizontal comparisons will be given more weight than the vertical comparisons by all the analysis models we considered. However, by making the stringent assumption that period and intervention effects are the same in every cluster, the standard model assumes too much certainty in separating the period and intervention effects. The reason that the standard model performed poorly in the simulation study was because of its reliance on the horizontal comparisons.

In the design we studied, the weight given to horizontal comparisons also meant that greater weight was given to some groups of clusters than others. The implications of this are not well understood. When there is a large difference in the weight given to each group, the intervention effect estimate no longer represents an average effect across the clusters and interpretation becomes more difficult. Further research is needed to explore this issue in more traditional SWT designs with more groups and when all clusters have observations in the control and intervention conditions, and so all clusters contribute through horizontal comparisons.

A criticism of the random intervention model and, to a lesser extent, the random period model is that they sometimes had problems with convergence. This occurred almost exclusively when both the period effect and the intervention effect were common to all clusters; the non-convergence resulted from the models attempting to estimate a true variance of zero, the boundary of the parameter. In this scenario, all the analysis models gave unbiased effect estimates and appropriate confidence interval coverage. We would suggest that an analysis plan gives an alternative, simpler model to use in case of convergence issues due to lack of variability. In our simulation study, this procedure gave good coverage and no bias in the scenarios with common period effect and intervention effect, where convergence was an issue (data not shown).

Given that the mixed effect model can be so sensitive to model assumptions, other analysis methods should be considered. This choice should be pre-specified and prior knowledge used to justify the assumptions made by the chosen analysis method. We found the random period model to be the most robust of the models considered, but there was still under-coverage of confidence intervals in some scenarios. Some have suggested using permutation tests on the standard model [20]. Although this will give correct inference, there is still a risk of biased intervention effect estimation. Alternative analysis methods which make fewer assumptions may be more appropriate. Generalised estimating equations have been suggested for the analysis of SWTs [21] and have been shown to be more robust to misspecification of the correlations in the data in other settings [22], but this robustness has yet to be assessed in the context of SWTs. Analysis methods which only make use of the vertical comparisons are desirable as they require no assumptions about period effects but there are no such methods currently published and these analyses are less efficient [23]. Sensitivity analysis could also be used to assess the robustness of results.

We have only considered a limited range of designs in this simulation study. We used a very simple SWT design to make the analyses as transparent as possible, this design only had two steps and not all clusters received the intervention in the course of the study. Further research is needed to confirm that our findings hold for other SWT designs. In more traditional SWTs, all clusters receive both the control and intervention conditions and so all clusters contribute horizontal comparisons. Since the problems we highlight arise from the horizontal comparisons, this might exacerbate the problems we identified. We have only considered two values for the ICC when the period effect was common to all clusters and have not assessed the effect of ICC when period effects vary between clusters. In scenarios where these effects varied between clusters, the baseline ICC was 0.20 which in many

contexts would be considered large. Additionally, there was large variability in the period effect; the effect of a less variable period effect needs further exploration. It is not known how common it is for the period and intervention effects to vary between clusters in practice; however, we have based this simulation on real trial data. Large clusters were used in the simulation study to reflect the motivating deworming trial; however similar results were seen with a smaller mean cluster size of 250 (data not shown). We used a large number of clusters in each group to avoid small sample issues.

Whilst further research is needed to explore the potential for bias in a wider range of designs and settings, we have demonstrated that there is a potential for the standard model to give biased intervention effect estimates and under-coverage of confidence intervals. These simulations provide clear evidence that the standard model for analysis of SWTs can be both highly sensitive to the data meeting the model assumptions and highly dependent on non-randomised horizontal comparisons. We urge those conducting SWTs to ensure an appropriate analysis is used.

References

1. Miguel E, Kremer M. Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 2004; **72**: 159-217, DOI DOI 10.1111/j.1468-0262.2004.00481.x.
2. Davey C, Aiken AM, Hayes RJ, Hargreaves JR. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasi-randomized stepped-wedge trial. *International Journal of Epidemiology* 2015; **44**: 1581-1592, DOI 10.1093/ije/dyv128.
3. Aiken AM, Davey C, Hargreaves JR, Hayes RJ. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication. *International Journal of Epidemiology* 2015; **44**: 1572-1580, DOI 10.1093/ije/dyv127.
4. Davey C, Hargreaves J, Thompson JA, Copas AJ, Beard E, Lewis JJ, Fielding KL. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. *Trials* 2015; **16**: 358, DOI 10.1186/s13063-015-0838-3.
5. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007; **28**: 182-191, DOI 10.1016/j.cct.2006.05.007.
6. Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* 2001; **88**: 973-985, DOI 10.1093/biomet/88.4.973.
7. Ukoumunne OC, Thompson SG. Analysis of cluster randomized trials with repeated cross-sectional binary measurements. *Stat Med* 2001; **20**: 417-433, DOI 10.1002/1097-0258(20010215)20:3<417::AID-SIM802>3.0.CO;2-G.
8. Koepsell TD, Martin DC, Diehr PH, Psaty BM, Wagner EH, Perrin EB, Cheadle A. Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs: a mixed-model analysis of variance approach. *J Clin Epidemiol* 1991; **44**: 701-713, DOI http://dx.doi.org/10.1016/0895-4356(91)90030-D.
9. Turner RM, White IR, Croudace T, Group PIPS. Analysis of cluster randomized cross-over trial data: a comparison of methods. *Stat Med* 2007; **26**: 274-289, DOI 10.1002/sim.2537.
10. Morgan KE, Forbes AB, Keogh RH, Jairath V, Kahan BC. Choosing appropriate analysis methods for cluster randomised cross-over trials with a binary outcome. *Stat Med* 2017; **36**: 318-333, DOI 10.1002/sim.7137.
11. Hooper R, Teerenstra S, de Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Stat Med* 2016, DOI 10.1002/sim.7028.

12. Girling AJ, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Stat Med* 2016; **35**: 2149, DOI 10.1002/sim.6850.
13. Bottomley C, Kirby MJ, Lindsay SW, Alexander N. Can the buck always be passed to the highest level of clustering? *BMC Med Res Methodol* 2016; **16**: 29, DOI 10.1186/s12874-016-0127-1.
14. Miguel E, Kremer M. Replication data for: Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities.Harvard Dataverse. 2014, Version: V2.1 DOI 10.7910/DVN/28038.
15. Campbell MK, Piaggio G, Elbourne DR, Altman DG, Group C. Consort 2010 statement: extension to cluster randomised trials. *bmj* 2012; **345**: e5661, DOI 10.1136/bmj.e5661.
16. Woertman W, de Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL, Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J Clin Epidemiol* 2013; **66**: 752-758, DOI 10.1016/j.jclinepi.2013.01.009.
17. Hemming K, Taljaard M. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. *J Clin Epidemiol* 2016; **69**: 137-146, DOI 10.1016/j.jclinepi.2015.08.015.
18. Hemming K, Girling A. A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster-randomized trials. *Stata Journal* 2014; **14**: 363-380.
19. Heagerty PJ, Zeger SL. Marginalized multilevel models and likelihood inference. *Statistical Science* 2000; **15**: 1-19, DOI 10.1214/ss/1009212670.
20. Ji X, Fink G, Robyn PJ, Small SS. Randomization inference for stepped-wedge cluster-randomised trials: An application to community-based health insurance. *Annals of Applied Statistics* 2016, DOI 10.1214/16-AOAS969.
21. Scott JM, deCamp A, Juraska M, Fay MP, Gilbert PB. Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. *Stat Methods Med Res* 2014, DOI 10.1177/0962280214552092.
22. Liang KY, Zeger SL. Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika* 1986; **73**: 13-22, DOI 10.1093/biomet/73.1.13.
23. Granston T. Addressing Lagged Effects and Interval Censoring in the Stepped Wedge Design of Cluster Randomized Clinical Trials. *University of Washington* 2014; **Doctor of Philosophy**.

Table 1: Summary of simulation study data scenarios

	Description	Similar to motivating example?
Common to all simulations		
Number of groups	3	Yes
Number of time-periods	2. In period 1 group 1 receive the intervention. In period 2 groups 1 and 2 receive the intervention	Yes
Number of clusters	75	Yes
Cluster size	Log-normal(6.9, 0.74) in each year. Geometric mean=1027	Yes
Correlation of measurements within clusters	Independent, assumes each measurement is from a different individual	No
Mean outcome in year one	Odds = 6.61	Yes
Mean change in outcome from year one to year two	Odds ratio = 0.32	Yes
Different scenarios		
Period effect	1. Common period effect, High variability	No
	2. Common period effect, Low variability	No
	3. Varying period effect, Decreasing Variability	Yes
	4. Varying period effect, Stable variability	No
Intervention effect	A. Log(OR)=0.41 common to all clusters	No
	B. Log(OR)=0.41, varying between clusters	No
Intervention effect in group 2		
Simulation study 1:	Intervention effect in group 2 the same as group 1 log(OR)=0.41	No
Simulation study 2:	Intervention effect in group 2 is log(OR)=1.5 and group 1 is log(OR)=0.41	No

Table 2: Intervention effect estimates from motivating example with different analysis models

Model	Odds Ratio (95% CI)	Standard error	P value	P value of random period or intervention effect
Separate year analysis (Vertical comparisons)				
Year one	1.67 (0.90,3.10)	0.32	0.11	
Year two	1.19 (0.95, 1.50)	0.12	0.13	
Combined analysis				
Standard model	1.74 (1.67, 1.81)	0.02	<0.001	
Random Period model	1.26 (1.02, 1.57)	0.11	0.03	<0.001
Random Intervention model	1.25 (0.96, 1.62)	0.13	0.09	<0.001