Frequentist inference for cluster randomised trials with multiple primary outcomes

Samuel I. Watson, University of Birmingham
Joshua Akinyemi, University of Ibadan
Karla Hemming, University of Birmingham

July 22, 2021

Abstract

The use of a single primary outcome is generally either recommended or required by many influential randomised trial guidelines to avoid the problem of “multiple testing”. Without correction, the probability of rejecting at least one of a set of null hypotheses (the family-wise error rate) is often much greater than the nominal rate of any single test so that statistics like p-values and confidence intervals have no reliable interpretation. Cluster randomised trials though may require multiple outcomes to adequately describe the effects of often complex and multi-faceted interventions. We propose a method for inference for cluster randomised trials with multiple outcomes that ensures a nominal family-wise error rate and produces simultaneous confidence intervals with nominal “family-wise” coverage. We adapt the resampling-based step-down procedure of Romano and Wolf (2005) using a randomisation-test approach within a generalised linear model framework. We then adapt the Robbins-Monro search procedure for confidence interval limits proposed by Garthwaite and Buckland (1996) to this stepdown process to produce a set of confidence intervals. We show that this procedure has nominal error rates and coverage in a simulation-based study of parallel and stepped-wedge cluster randomised studies and compare results from the analysis of a real-world stepped-wedge trial under both the proposed and more standard analyses.

Keywords: cluster randomised trial, inference, coverage, multiple testing
1 Introduction

The requirement to state a single primary outcome for a randomised controlled trial has become accepted, even required, practice. For example, the influential CONSORT statement on clinical trials requires the pre-specification of a primary outcome, which they describe as “the pre-specified outcome considered to be of greatest importance to relevant stakeholders”, and recommends against multiple primary outcomes [The Lancet, 2010]. The reason for this is to ensure appropriate control of the ‘false discovery rate’ when using null hypothesis significance testing [Wason et al., 2014]. If there are multiple outcomes each with their own associated treatment effect being tested separately, then we are implicitly testing a family of null hypotheses against an alternative that at least one of them is false. Without correction, the type I error rate for this family of null hypotheses will be much greater than the nominal rate of any single test [Romano and Wolf, 2005a]. Indeed, the CONSORT statement notes that that multiple primary outcomes are not recommended as it “incurs the problem of multiplicity of analyses” [The Lancet, 2010].

Cluster randomised trials are a widely used method to evaluate interventions applied to groups of people, such as clinics, schools, or villages. Often these interventions target ‘higher level’ processes and can be complex in nature [Murray, 1998; Hemming et al., 2015; Eldridge and Kerry, 2012]. Recent examples from our own work include an incentive scheme to improve implementation of a broad package of education and activities designed to improve employee health in the workplace [Thrive at Work Wellbeing Programme Collaboration, 2019a], or a community health worker programme targeting multiple health conditions [Dunbar et al., 2018]. The effects of such complex interventions cannot be adequately summarised by a single outcome. ‘Fudging’ this process by creating a composite outcome is undesirable since it requires applications of arbitrary weights across outcomes.
and discards information by collapsing a multivariate outcome to a univariate one. The requirement for a single primary outcome therefore clashes with the needs of many cluster randomised trials. The solution is to ensure appropriate methods are used where there are multiple outcomes of interest rather than restricting the outcomes from which we can make inferences. However, the question of appropriate analysis for randomised trials, and particularly cluster randomised trials, with multiple outcomes can be contentious and complex.

The Food and Drug Administration (FDA), the main regulatory body for medicines in the United States, declares that “If the purpose of the trial is to demonstrate effects on all of the designated primary variables, then there is no need for adjustment of the type I error” [US Food and Drug Administration, 2017]. They also identify a “gatekeeping” approach where “statistical significance” on a primary outcome is required before a second one can be analysed and state this does not need correction for multiple testing. However, in both cases, the type I error for the two (or more) outcomes jointly is unlikely to be nominal with this procedure and their validity is debatable. Other authors differentiate aiming to declare “statistical significance” on at least one of a group of null hypotheses to declaring it for all, and propose different solutions for both [Lafaye de Micheaux et al., 2014, Rubin, 2021].

Where a correction for multiple testing is deemed necessary, we can divide solutions into: (i) multivariate methods that model the joint distribution of the outcomes, which is particularly favoured by Bayesian practitioners [Gelman et al., 2012]; and (ii) univariate solutions that aim to ensure inferential statistics for a set of estimands collectively have the appropriate Frequentist properties [Farcomeni, 2008]. Despite the different approaches and guidance, [Wason et al. 2014] estimated that only around half of all randomised trials with multiple outcomes or arms corrected for multiple testing and outcomes. No evidence is available on the use of corrections for multiple testing in cluster randomised trials specif-
ically, but there are few, if any, comprehensive discussions of methods in this area currently available. No guidance is available on methods for correcting confidence intervals or if it is needed; the FDA note that it is complex and beyond the scope of their advice.

This article adapts and evaluates a univariate method for the use of multiple primary outcomes in cluster randomised trials. We do not aim to resolve the debate about the necessity of these methods, rather to provide a novel method that does ensure a nominal type I error rate and confidence set coverage when multiple outcomes are used in cluster trials, should one be required. We do not consider the problem with the goal of declaring “statistical significance”, although we do use terms such as “error rate” for convenience. Our contention is that even in the absence of declarations of “statistical significance”, statistics such as $p$-values and confidence intervals must possess the appropriate Frequentist properties in order to be interpretable. 95% confidence intervals are presented in applied research on the basis that they have a coverage of 95%, if instead they had unknown coverage that might range, say, between 50% and 99% then they would have little ability to support principled inference from statistical analysis of randomised trials and other studies. If a $p$-value of 0.03 could mean the null model(s) would produce datasets with more extreme test statistics in anywhere from 2% to 20% of repetitions, then it is similarly unusable. Our aim therefore is to derive $p$-values and confidence intervals with a nominal type I error rate for a family of null hypotheses and coverage of confidence sets for the parameters of interest.
2 Family wise error rates

2.1 The multiple testing problem

We first suppose that data $X$ are generated from some probability distribution $P$, which belongs to some family of probability distributions $\Omega$. The family $\Omega$ could be a parametric, semi-parametric, or non-parametric model. The multiple testing problem arises when we have a set of hypotheses $H_j$ versus $H_j'$ for $j = 1, \ldots, J$, following the notation of Romano and Wolf [2005a]. Each of these hypotheses is a subset $\omega_j \subset \Omega$ and is equivalent to testing $P \in \omega_j$ against $P \not\in \omega_j$. So for any subset $K \subset 1, \ldots, J$, $H_K = \cap_{j \in K} H_j$ is the hypothesis that $P \in \cap_{j \in K} \omega_j$. We assume each null hypothesis $H_j$ is based on a test statistic $T_j$; we denote the $\alpha$-quantile of the distribution of $T_j$ as $c_j(\alpha, P)$. In a traditional null hypothesis testing framework we “reject” $H_j$ in favour of $H_j'$ at the $\alpha$ level, if $T_j \geq c_j(1 - \alpha, P)$, which clearly has probability $\alpha$. Conversely, the $p$-value $p_j$ of the test is where $T_j = c_j(1 - p_j, P)$, so that the probability of observing $Pr(T_j > c_j(1 - p_j, P)|H_j) = p_j$. The family-wise error rate (FWER) of this set of hypotheses is the probability of “rejecting” at least one true null hypothesis. That is, if $I = I(P) \subset 1, \ldots, J$ are the indices of the true null hypotheses, so $j \in I$ if and only if $P \in \omega_j$, then the FWER is the probability under $P$ of rejecting any $H_{j \in I}$, i.e. $Pr(\cup_{j \in I} T_j > c_j(1 - \alpha, P))$, which should be $\alpha$.

2.2 Solutions

Solutions to the multiple testing problem aim to ensure that $FWER \leq \alpha$. Control over the FWER is said to be strong if it holds for any combination of true and false null hypotheses, and weak if it only holds when all null hypotheses are true [Dudoit et al., 2003]. Several approaches exist to control the FWER. The Bonferroni method is probably
most well known, which sets the critical value for the test of the null hypothesis to be 
$c_j(1 - \alpha/J, P)$. However, while this method exerts strong control over the FWER. It
is very conservative and assumes no correlation between the test statistics. Holm [1979]
proposed a less conservative ‘stepdown’ approach to multiple testing. One orders the test
statistics from largest to smallest and then compares the largest statistic to the critical
value $c_j(1 - \alpha/J, P)$. If the test statistic is larger than this value, then the null hypothesis
is rejected, otherwise we do not reject any null hypothesis and stop. If we rejected, then the
next largest test statistic is compared to $c_j(1 - \alpha/(J-1), P)$, and again it is either rejected,
or we do not reject all remaining null hypotheses and stop, and so forth. The stepdown
method is less conservative than the Bonferroni method, but it can still be conservative as
it does not take into account the dependence structure in the data and of the test statistics.

Romano and Wolf [2005a,b] developed a resampling based version of the stepdown
method. Being based on randomisation or permutation tests provides an advantage as it
allows for arbitrary dependence structures in the data while maintaining strong control
of the FWER. Their process is optimal in a maximin sense. These properties, and the
use of randomisation tests, are useful for cluster randomised trials with multiple outcomes
as the construction of appropriate test statistics can be difficult owing to the difficulty of
estimating reliable standard errors with the potentially complex covariance structure and
often small number of clusters [Watson et al., 2021; Leyrat et al., 2018]. Furthermore, as
we discuss in Section 4, this procedure provides a means to construct confidence intervals.
Other methods for controlling the FWER have been proposed, which are surveyed elsewhere
Farcomeni, 2008.

We describe the general stepdown procedure of Romano and Wolf firstly in terms of
accepting or rejecting each null hypothesis at an $\alpha$-level. We let $c_K(\alpha, P)$ denote an $\alpha$-
quantile of the distribution of the statistic:

\[ T_K = \max_{j \in K} T_j \]  

for any subset of null hypotheses \( K \). We also denote \( T_{[r]} \) as the \( r \)th largest test statistic so that

\[ T_{[1]} \geq T_{[2]} \geq \ldots \geq T_{[J]} \]  

corresponding to hypotheses \( H_{[1]} \), \( H_{[2]} \), \ldots, \( H_{[J]} \). Then the idealised algorithm is:

1. Let \( K_1 = 1, \ldots, J \). If \( T_{[1]} \leq c_{K_1}(1 - \alpha, P) \) then accept all hypotheses and stop; otherwise, reject \( H_{[1]} \) and continue;

2. Let \( K_2 \) be the indices of all the hypotheses not previously rejected. If \( T_{[2]} \leq c_{K_2}(1 - \alpha, P) \), then accept all remaining hypotheses and stop; otherwise, reject \( H_{[2]} \) and continue;

\vdots

\( J \). If \( T_{[J]} \leq c_{K_J}(1 - \alpha, P) \) then do not reject \( H_{[J]} \), otherwise reject.

In this procedure it is assumed the critical values are known. However, as Romano and Wolf show, they can be replaced by estimates from randomisation tests. In particular, the null hypothesis implies that the distribution of \( X \) is invariant under a set of transformations in \( A \), which has \( L \) elements. So, \( aX \) and \( X \) have the same distribution for all \( a \in A \) whenever \( X \) has distribution \( P \in \omega \).

Our observed test statistics with our sample data are \( T_j(X) \). The test statistic generated by the \( l \)th permutation is \( T_j(a_lX) \) for \( a_l \in A \) and \( l = 1, \ldots, L \). For each permutation we can determine the test statistic as in Equation (1) as \( T_{K,l} = \max_{j \in K} T_j(a_lX) \). As before we
denote $T_{K|r}$ as the $r$th largest of all the permutational test statistics $\{T_{K;l}; l = 1, \ldots, L\}$ of the test statistics from the permutations. Then our estimator for the critical value is:

$$\hat{c}_K(1 - \alpha, P) = T_{K,|L(1-\alpha)|}$$

where $T_{K,|L(1-\alpha)|}$ is the $L(1-\alpha)$th (or nearest integer) largest value from the permutations. We can see how this procedure produces $p$-values for a two-sided hypothesis that also maintains the FWER for a given $\alpha$ [Romano and Wolf, 2016], in particular:

$$p_K = \frac{1}{L} \sum_{l=1}^{L} 1[|T_K(a_lX)| \geq |T_K(X)|]$$

(4)

For a one-sided test we would not use the absolute values of the test statistics.

Often the size of $A$ can be very large, and increases exponentially with the number of clusters. A Monte Carlo approach can be used that instead generates a random subset of $A$ of fixed sized in order to generate realisations of the test statistics. If we conduct $M$ such permutations then the estimator of the $p$-value for a given null hypothesis versus some alternative is

$$\hat{p}_K = \frac{1}{M+1} \sum_{m=1}^{M} (1 + 1[T(a_mX) \leq T(x)])$$

(5)

Obtaining $p$-values in this way is described in detail by [Romano and Wolf, 2016]. In the next section we discuss permutation tests in the context of cluster randomised trials.

3 Permutation tests and $p$-values for cluster trials

We next introduce a generalised linear mixed model commonly used in the analysis of cluster randomised trials (e.g. Hooper et al., 2016). We denote $Y_{ict}$ as the outcome of the $i$th individual, $i = 1, \ldots, N$, in cluster $c = 1, \ldots, C$ at time $t = 1, \ldots, T$. We do not restrict
the outcome, it could be continuous or discrete. We also specify the linear predictor:

\[ \eta_{ict} = \mu + \delta D_{ct} + X_{ict}' \beta + \theta_{ct} \]  

(6)

where \( D_{ct} \) is an indicator for whether cluster \( c \) has received the intervention at time \( t \) and so \( \delta \) is the parameter of interest, our “treatment effect”. We also have a vector of individual and/or cluster-level covariates, \( X_{ict} \), which may also contain temporal fixed effects. The parameter \( \theta_{ct} \) represents a general ‘random-effect’ term that captures the within cluster and cluster-time correlation, although we do not provide a specific structure here. The overall model is then

\[ Y_{ict} \sim P(h(\eta_{ict})) \]  

(7)

where \( h(.) \) is a link function. For example, \( P \) could be a Binomial distribution and \( h(.) \) the logistic link function.

Gail et al. [1996] provided the first extensive examination of permutation tests for cluster-based study designs. Their work principally used unweighted differences of cluster means as the basis of permutation tests. In particular, if \( \bar{Y}_{ct} = \sum_i Y_{ict} \) then one can base inference on the mean difference:

\[ U = \frac{1}{J} \frac{1}{T} \sum (\bar{Y}_{ct} D_{ct} - \bar{Y}_{ct} (1 - D_{ct})) \]  

(8)

Under the null hypothesis that \( H_0 : U = 0 \) the clusters are exchangeable, so we can re-assign them to a new allocation \( a \in A \). In this way critical values of the distribution of \( U \) under this null, or associated \( p \)-values, can be generated. This statistic is the basis of a statistical package for inference from stepped-wedge cluster randomised trials using randomisation tests (\texttt{cptest}; Gallis et al. [2018]). The authors of that package also implement a “within-period” version of the analysis which takes estimates “within period” treatment effects as:

\[ U_t = \frac{1}{J} \sum (\bar{Y}_{ct} D_{ct} - \bar{Y}_{ct} (1 - D_{ct})) \]  

(9)
and then takes a weighted average of them based on the number of clusters in treatment and control conditions in each time period. One can extend these ideas to allow for covariate adjustment [Gail et al., 1996]. Braun and Feng [2001] examine optimal permutation tests for cluster randomised trials. The test statistics they consider are weighted sums of residuals. In particular, they consider a score-based statistic derived from the marginal means of the observations modelled separately from the correlation structure of the data. The mean of each observation is

\[ h^{-1}(\mu_{ict}) = \mu + \delta D_{ct} + X'_{ict}\beta \]  

(10)

which is devoid of the cluster effect \( \theta_{ct} \). The "quasi-score" statistic is then:

\[ \sum_c \{D^*_cG_cV_c^{-1}[Y_c - \mu_c]\}_{|\delta=\delta^*} \]  

(11)

where \( D^*_c[D^*_{c1}, D^*_{c1}, ..., D^*_{c_T}, D^*_{c_T}] \) is a \((1 \times n_c)\) vector of modified intervention indicators equal to 1 if the intervention was present in cluster \( c \) at time \( t \) and -1 otherwise, and where \( n_c = \sum_t n_{ct} \) and \( n_{ct} \) is the number of individuals in cluster \( c \) at time \( t \). \( G_c \) is a \((1 \times n_c)\) vector with elements \((\partial h^{-1}_{ict}/\partial \eta_{ict})^{-1}\), and \( V_c \) is an \((n_c \times n_c)\) covariance matrix for cluster \( c \) with non-zero elements off its diagonal. As an example, if we assume the data are normally distributed with mean given by Equation (6), identity link function, variance \( \sigma^2 \), and \( \theta_{ct} \sim N(0, \tau^2) \), then the diagonal elements of \( V_c \) are \( \sigma^2 + \tau^2 \) and the off-diagonal elements are \( \tau^2 \). More complex structures might include temporal decay in correlation, for example. We use \( \Theta \) to represent the parameters of the variance-covariance matrix. Finally \([Y_c - \mu_c]\) represents generalised residuals: \( Y_c = [Y_{1c1}, Y_{2c}, ..., Y_{n_{c1}c1}, Y_{1c2}, ..., Y_{n_{cT}cT}] \) is a \((1 \times n_c)\) vector of outcomes and \( \mu_c \) is a \((1 \times n_c)\) vector of means.

For the permutation test to be valid the 'nuisance' parameters \((\mu, \beta, \Theta)\) must be invariant to permutation [Braun and Feng 2001]. This means we cannot re-estimate them for
each new permutation. In practice the maximum likelihood estimates of these parameters are used to construct the test statistic, so that we use the estimates:

$$\hat{\mu}_{ict} = h(\hat{\mu} + \delta^* D_{ct} + X_{ict}' \hat{\beta})$$  \hspace{1cm} (12)

for the linear predictor under the null $H_0 : \delta = \delta^*$. Estimating $\Theta$ is more difficult, however, particularly when the number of clusters is small [McNeish and Stapleton, 2016, Watson et al., 2021]. As an alternative to (11) we can replace $G_c V_c^{-1}$ with a $(1 \times n_c)$ vector of ones $1_c$:

$$\sum_c \sum_t \sum_i \{D_{ict}^* 1_{ict} \{Y_{ict} - \mu_{ict}\}\} = \sum_c \sum_t \sum_i \{D_{ict}^* 1_{ict} \{Y_{ict} - \mu_{ict}\}\}$$  \hspace{1cm} \hspace{1cm} (13)

so that the sum of residuals is weighted only by the size of each cluster or cluster-time period. [Braun and Feng, 2001] provide proof of the optimality of the quasi-score (11) in the context of a single hypothesis test; the second test in (13) is likely to be suboptimal if there is a (strong) dependence structure in the data. Nevertheless, in simulation-based experiments, they show both have nominal or near-nominal type I error rates for various levels of intra-cluster correlation.

For the purposes of multiple testing we use a studentized version of (13):

$$T(X)|_{\delta = \delta_0} = \frac{\sum_c \sum_t \sum_i \{D_{ict}^* 1_{ict} \{Y_{ict} - \mu_{ict}\}\}}{\sqrt{\sum_c \sum_t \sum_i \{D_{ict}^* 1_{ict} \{Y_{ict} - \mu_{ict}\}\}^2}}$$  \hspace{1cm} \hspace{0.5cm} (14)

where the terms on the right-hand side have been evaluated at $\delta = \delta^*$. In the absence of studentization, the variances of the test statistics are not scale-free and depend on, among other things, the null hypothesis being tested so that different tests will have different power under stepdown testing [Romano and Wolf, 2005a]. The lack of balance is particularly consequential for the construction of confidence sets discussed in the next section. While confidence sets constructed on the basis of stepdown methods will have joint coverage of
1 − \alpha, without balance the individual coverage probabilities of each interval will differ, perhaps substantially [Romano and Wolf, 2005b].

The test statistic described in this section can be used with the stepdown procedure of Section 2.2 to provide a basis for generating \( p \)-values for multiple hypothesis tests across multiple outcomes with a nominal FWER. However, a point estimate and \( p \)-value is often not considered sufficient for providing inferences from trials. The \( p \)-value can provide relatively little indication of the magnitude of sampling variation associated with a particular parameter, and often confidence intervals are expected or required.

4 Confidence sets and multiple testing

The multiple testing problem extends to the construction of simultaneous confidence intervals or a “confidence set”. Let the parameters of interest be \( \delta_j \) with associated confidence intervals \([L_j, U_j]\), so that \([L_1, U_1] \times [L_2, U_2] \times \ldots \times [L_J, U_J]\) forms a confidence set. Similar to the FWER, we want appropriate control of the coverage of the 100(1 − \alpha)\% confidence set such that the process produces confidence sets with the property:

\[
Pr(L_j \leq \delta_j \leq U_j \forall j) = 1 - \alpha
\]  

we refer to this as ‘family-wise coverage’. If we construct 100(1 − \alpha)\% confidence intervals independently then the probability that at least one interval in the set excludes the true value can significantly exceed \( \alpha \). There have been some attempts to construct exact confidence sets for parameters based on the stepdown procedure [Romano and Wolf, 2005b]. For example, [Guilbaud, 2008], extending the proposal of Hayter and Hsu [1994], uses the acceptance/rejection of null hypotheses by the stepdown procedure as a basis of determining upper or lower limits of confidence intervals if we conclude they are strictly negative.
or positive, respectively. However, these procedures can only provide information on the upper or lower bound respectively - the other end of the interval is infinity - so they provide little extra information on the extent of sampling variation beyond the \( p \)-value.

As an alternative, consider for a moment, a single parameter \( \delta_1 \). Its 100(1 − \( \alpha \))% confidence interval is \([L_1, U_1]\); for any value \( \delta_1^* \) inside this interval the null hypothesis \( H_1 : \delta_1 = \delta_1^* \) will not be rejected in favour of the two-sided alternative \( H_1' : \delta_1 \neq \delta_1^* \) at the \( \alpha \) level. The question is then how to find the values of \( L_1 \) and \( U_1 \) efficiently. As suggested by [Braun and Feng 2001], one could iteratively perform a series of permutation tests to identify the limits as \( U_1 = \sup \{ \delta_1^* : \text{do not reject } \delta_1 = \delta_1^* \} \) and \( L_1 = \inf \{ \delta_1^* : \text{do not reject } \delta_1 = \delta_1^* \} \).

However, this procedure is inefficient, particularly when testing multiple parameters: if there are \( M \) permutations per test and \( J \) outcomes, then for each increment in \( U \) we must calculate \( JM \) permutation test statistics and perform the stepdown procedure. Moreover, since the test statistic and its randomisation distribution depends on the values of the other null hypotheses being tested, a very large number of combinations of values of the parameters must be tested to ensure we have identified with reasonable certainty the limits of the confidence set.

[Garthwaite and Buckland 1992] developed a method for searching for confidence interval endpoints efficiently, which [Garthwaite 1996] later adapted for use with randomisation tests. Their method is based on the search process devised by [Robbins and Monro 1951], who developed a stochastic approximation procedure to find the \( \alpha \)-quantile of a particular distribution. Multivariate Robbins-Monro processes follow the same procedures as their univariate equivalents [Ruppert 1985]. For our multiple testing scenario the upper limits to the confidence set correspond to where all hypotheses \( H_j : \theta_j = U_j \) for \( j = 1, ..., J \) are all rejected in favour of the two-sided alternative with a FWER of \( \alpha \) but for any smaller values of \( U_j \) not all hypotheses are rejected, and equivalently for the lower limits.
Returning to Equation 4 and following the logic of the stepdown procedure, if the FWER rate for this family of hypotheses is preserved then \( Pr(|T_{K_j}(aX)| \geq |T_{[j]}(X)|) = \alpha \) (and equivalently \( Pr(|T_{K_j}(aX)| < |T_{[j]}(X)|) = 1 - \alpha \)) where \( T_{K_j}(aX) \) is the largest of the permutational test statistic corresponding to the hypotheses \( H_{[j]} \) to \( H_{[J]} \). This provides a probabilistic basis for the search procedure based on single iterations of a randomisation test.

At the \( q \)th step of \( Q \) steps total, we have estimates of the upper confidence interval limits of our \( J \) parameters \( u_q = [u_{1q}, u_{2q}, ..., u_{jq}] \). We generate the set of test statistics \( T_j(X)|_{\delta = u_{jq}} \), which correspond to the null hypotheses \( H_j : \theta_j = u_{jq} \). We then generate a single permutation of a randomisation test for the same hypotheses \( T_j(a_qX)|_{\delta = u_{jq}} \). The probability that \( |T_j(a_qX)|_{\delta = u_{jq}} \) is less than \( |T_j(X)|_{\delta = u_{jq}} \) is \( \alpha_q \). The estimates of the upper limits are then updated as:

\[
 u_{j,q+1} = \begin{cases} 
 u_{jq} - s_j \alpha / q & \text{if } |T_{K_j}(a_qX)|_{\delta = u_{jq}} < |T_{[j]}(X)|_{\delta = u_{jq}} \\
 u_{jq} + s_j (1 - \alpha) / q & \text{if } |T_{K_j}(a_qX)|_{\delta = u_{jq}} \geq |T_{[j]}(X)|_{\delta = u_{jq}}
\end{cases}
\]  

(16)

where \( s_j \) is the “step length constant”. The mean increment of each estimate at each step is \( \alpha_q s_j (1 - \alpha) / q - (1 - \alpha) s_j \alpha / q = s_j (\alpha_q - \alpha) / q \) so that it is clear that each step reduces the expected distance from \( U_j \). Similarly for the lower limits, the updating rule is:

\[
 l_{j,q+1} = \begin{cases} 
 l_{jq} + s_j \alpha / q & \text{if } |T_{K_j}(a_qX)|_{\delta = l_{jq}} < |T_{[j]}(X)|_{\delta = l_{jq}} \\
 l_{jq} - s_j (1 - \alpha) / i & \text{if } |T_{K_j}(a_qX)|_{\delta = l_{jq}} \geq |T_{[j]}(X)|_{\delta = l_{jq}}
\end{cases}
\]  

(17)

The step length constants are \( s_j = k(u_{jq} - \hat{\theta}_j) \) and \( s_j = k(\hat{\theta}_j - l_{jq}) \) for the upper and lower limits, respectively, where \( \hat{\theta}_j \) is a point estimate of the parameter and:

\[
k = \frac{2}{z_{\alpha/2}(2\pi)^{-1/2} \exp(-z_{\alpha/2}^2/2)}
\]  

(18)
where $z_\alpha$ is the $\alpha$-quantile of the standard normal distribution. Under relatively weak regularity conditions this process converges in probability to the correct confidence limits as $l \to \infty$ [Garthwaite and Buckland, 1992; Ruppert, 1985].

4.1 Computation

An R package to execute the analyses described in this paper is available from CRAN as crctStepdown.

5 Simulation studies

To examine the FWER and simultaneous coverage of the procedures outlined in this article we conduct a series of simulation studies. We start with the simplest relevant design and move to more complex constructions. A brief description of each simulation is described here, a full description can be found in the Supplementary Material.

Each simulation is based on a hypothetical or real two-arm cluster randomised trial of different designs. All outcomes are simulated and modelled using exponential-family models. In all simulations we set the number of individuals per cluster to 20. We consider both when all treatment effects are zero and when only a subset are and also vary the number of clusters per arm. An important parameter for cluster trial design and analysis is the intraclass correlation coefficient (ICC), which is the ratio of cluster-level to total variance. For non-linear models the ICC is not constant as it depends on the particular realisations of covariates and parameter values. We set cluster-level random effect terms to have variance to give ICCs in the range of 0.01 to 0.10. The treatment effect parameters for each simulation are a vector, $\delta$, with length equal to the number of outcomes.

In each simuation, we compare a small and a large number of clusters per arm, with the
exact numbers given below. Standard uncorrected test statistics, like an $F$-test, generally do not provide reliable inference for trials with fewer than around ten clusters per arm [Leyrat et al., 2018; Watson et al., 2021]. While (small-sample) corrected test statistics are available with variable performance, permutation tests should provide exact inference and we should expect to see nominal FWERs and coverage with both small and large numbers of clusters [Li et al., 2017]. Uncorrected test statistics should only provide nominal FWER with a large number of clusters and a single ‘correct’ null hypothesis.

The four simulation models we compare are:

- (1) Two-arm, parallel cRCT with two outcomes and no covariates. There are two outcomes that are Poisson distributed and only measured in a single post-intervention period.

- (2) Two-arm, parallel cRCT, with three outcomes, and covariates. The three outcomes are Poisson, Bernoulli, and Normally-distributed. We include one normally distributed covariate. As before, all outcomes are only measured in a single post-intervention period.

- (3) Stepped-wedge cRCT with four outcome measures and covariates. Of the four outcomes, three were Poisson distributed and one normally distributed. We include one normally distributed covariate. For all simulations of this type we include 20 clusters and ten ‘steps’ (two cluster per sequence). This simulation is based on an ongoing trial of the effect of a remote consulting intervention on consultation rates at primary care clinics.

- (4) Two-arm, parallel cRCT with three outcomes and no covariates. The purpose of this final simulation was to examine the effect of both very large numbers of clusters
and a correlation between trial outcomes. The three outcomes were multivariate normally distributed with standard deviation terms all equal to one and the correlation between any two outcomes equal to $\rho$.

5.1 Simulation methods and evaluation

Each set of simulations is run 10,000 times. We then estimate the FWER for $p < 0.05$, which would have a nominal rate of 5%, and also estimate coverage of 95% confidence sets. We also rerun each set of simulations but use instead uncorrected $t$-statistics to obtain $p$-values and confidence sets for the parameters of interest to compare to the corrected procedure. The $t$-statistics we use are the default ones returned by `lme4` using the standard errors estimated from the correctly specified model (i.e. with cluster and cluster-time random effects, where relevant). 95% confidence intervals for each parameter were constructed in the default way of $\hat{\delta} + / - 1.96 \times SE$. To conduct the simulations we used the R package `crctStepdown` described above.

5.2 Results

Table 1 reports the results from simulation experiments (1) to (3). Under all tested conditions the FWER (for $p < 0.05$) and coverage (95%) were nominal under the proposed corrected procedure. Running time for a single iteration of the procedure ranged from 20 seconds (for the two outcome, single-period simulation) to two minutes (for the four outcome, stepped-wedge design). Figure 1 shows an example of the confidence interval search method output. Table 2 reports the results of simulation experiment (4). As before, FWER and coverage were nominal in all cases.

We compare our results to a “standard analysis” that uses uncorrected $t$-tests as the
Figure 1: Example of the search procedure for the upper limits of four treatment effects simulated from the stepped-wedge design

test statistics. FWERs were all non-nominal except in the case where there was only one true null and there were a large number of clusters. For example, Table 2 shows FWER of 0.05 for 50 clusters per arm and one true null (and equivalently FWER of around 0.15 for three true nulls). In the simulations where there were a small number of clusters rates were not nominal even when there was only one true null, reflecting previous results [Leyrat et al., 2018, Watson et al., 2021]. None of the confidence sets had appropriate coverage under the uncorrected method.

6 Applied example

To provide a real-world example of the the use of the methods proposed in this article, we re-analyse a cluster randomised trial of a financial incentive to improve workplace health and wellbeing in small and medium sized enterprises (SME) in the United King-
| Clust./arm | Parameters (δ) | Corrected FWER | CI coverage | Uncorrected FWER | CI coverage |
|-----------|---------------|----------------|-------------|------------------|-------------|
| (1)       | 7 (0,0)       | 0.055          | 0.962       | 0.168            | 0.834       |
|           | 20            | 0.072          | 0.940       | 0.124            | 0.878       |
|           | 7 (0,0.5)     | 0.046          | 0.969       | 0.104            | 0.820       |
|           | 20            | 0.048          | 0.978       | 0.068            | 0.866       |
| (2)       | 7 (0,0,0)     | 0.056          | 0.950       | 0.216            | 0.784       |
|           | 20            | 0.038          | 0.962       | 0.176            | 0.824       |
|           | 7 (0,0,0.5)   | 0.052          | 0.942       | 0.152            | 0.780       |
|           | 20            | 0.043          | 0.949       | 0.102            | 0.846       |
|           | 7 (0,0.5,0.5) | 0.048          | 0.939       | 0.108            | 0.774       |
|           | 20            | 0.052          | 0.947       | 0.066            | 0.838       |
| (3)       | 7 (0,0,0,0)   | 0.055          | 0.924       | 0.216            | 0.788       |
|           | 20            | 0.070          | 0.942       | 0.498            | 0.508       |
|           | 7 (0,0.5,0,1) | 0.044          | 0.920       | 0.110            | 0.774       |
|           | 20            | 0.052          | 0.944       | 0.294            | 0.530       |

Table 1: Results of simulation experiments (1) - (3) with 10,000 iterations each. Each iteration used 1,000 permutations for the permutation test and 2,000 iterations in each of the lower and upper confidence interval search processes. (1) Two outcomes simulation; (2) Three outcomes simulation; (3) Stepped-wedge simulation.
| $\rho$ | Clust./arm | Parameters ($\delta$) | Corrected FWER | CI coverage | Uncorrected FWER | CI coverage |
|--------|------------|-----------------------|----------------|-------------|------------------|-------------|
| 0.1    | 10         | (0,0,0)               | 0.052          | 0.950       | 0.186            | 0.817       |
| 0.1    | 50         | (0,0,0)               | 0.050          | 0.959       | 0.149            | 0.856       |
| 0.5    | 10         | (0,0,0)               | 0.051          | 0.954       | 0.172            | 0.831       |
| 0.5    | 50         | (0,0,0)               | 0.050          | 0.960       | 0.146            | 0.858       |
| 0.1    | 10         | (0,0.5,1)             | 0.053          | 0.940       | 0.065            | 0.823       |
| 0.1    | 50         | (0,0.5,1)             | 0.051          | 0.938       | 0.056            | 0.849       |
| 0.5    | 10         | (0,0.5,1)             | 0.045          | 0.940       | 0.064            | 0.822       |
| 0.5    | 50         | (0,0.5,1)             | 0.053          | 0.941       | 0.054            | 0.856       |

Table 2: Results of the simulation experiments (4) with 10,000 iterations. Each iteration used 1,000 permutations for the permutation test and 2,000 iterations in each of the lower and upper confidence interval search processes.
6.1 Outcomes

A single primary outcome was specified in the protocol, which was the question “Does your employer take positive action on health and wellbeing?” However, given the potential lack of insight it might provide into the functioning of the intervention, several secondary outcomes were specified to capture the “causal chain” between intervention and employee health and wellbeing. For each of three separate health categories (mental, musculoskeletal, and lifestyle health) employees were asked:

1. whether the employer provided information in this area;

2. whether the employer had provided activities and services in this area;

3. whether the employee had made a conscious effort to improve in this area;

4. whether the employee had attended any groups or activities in this area at work;

5. whether the employee had attended any groups or activities in this area outside of work.

for a total of 15 outcomes.
6.2 Re-analysis

The original analysis of the trial took a Bayesian approach. The Frequentist re-analysis we conduct here is principally for illustrative purposes, and so we only take a subset of the data and simplify some of the outcomes. In particular, we take only the main control arm and the high incentive intervention arm to estimate the effect of the high incentive. We focus on the set of secondary outcomes listed above, which we collapse into five separate outcomes; whether the employer provided information across all three health areas, and then whether there was a positive response for any of the health areas for the remaining outcomes, for a total of five outcomes. All outcomes are modelled using a Bernoulli-logistic regression model, following the notation above, with \( t = 0 \) for baseline and \( t = 1 \) for post-intervention:

\[
Y_{k,ict} \sim \text{Bernoulli}(\logit(\mu_k + \delta_k D_{ct} + \theta_{k,c} + \theta_{k,ct}))
\]  

(19)

6.3 Results

Table E shows the results of an ‘uncorrected’ univariate analysis using the standard output from R’s \texttt{lme4} package alongside ‘corrected’ results using the method proposed in this article. We make several observations. The uncorrected analysis would suggest there is likely good evidence that the intervention improved employer provision of information and activities and services, and increased employee taking part at work. However, this conclusion might contradict our understanding of the causal processes since it would seem contradictory for employees to make more effort but not report making more effort. The results corrected for multiple testing appear to be more consistent in that employers appeared to make more effort but the employees did not take up the new services. The effect of the intervention is also more uncertain than suggested by the uncorrected confidence intervals. In particular, the confidence intervals under the corrected method, which are
| Outcome                                    | Estimated parameter | No correction | Correction |
|--------------------------------------------|---------------------|---------------|------------|
| Employer provided information              | 2.91                | <0.01         | 0.01       |
|                                            |                     | [1.98, 3.97]  | [0.29, 2.96]|
| Employer provided activities and services   | 2.11                | <0.01         | 0.03       |
|                                            |                     | [1.31, 2.99]  | [0.43, 2.93]|
| Employee made a conscious effort            | 0.22                | 0.44          | 0.33       |
|                                            |                     | [-0.33, 0.77] | [-0.48, 0.88]|
| Employee took part at work                  | 1.13                | <0.01         | 0.23       |
|                                            |                     | [0.50, 1.75]  | [-0.30, 1.65]|
| Employee took part outside work             | 0.27                | 0.11          | 0.15       |
|                                            |                     | [-0.06, 0.61] | [-0.05, 0.84]|

Table 3: Results from re-analysis of the workplace wellbeing trial. Results are log odds-ratios.
based on exact randomisation tests, are not symmetric for several outcomes, unlike under the uncorrected approach. So, smaller effect sizes, particularly for the first two outcomes, are more plausible than the uncorrected method would suggest.

7 Discussion

This article presents the first method to provide Frequentist statistics for cluster randomised trials with multiple outcomes that controls for the FWER and coverage of simultaneous confidence intervals. We build on a range of previous work including: randomisation tests for cluster trials [Gail et al., 1996, Gallis et al., 2018], univariate methods for corrections for multiple testing that use randomisation tests [Romano and Wolf, 2005a,b, 2016], and procedures for estimating confidence interval limits based on randomisation tests [Garthwaite and Buckland, 1992, Garthwaite, 1996]. Altogether the proposed method can deal with several issues that are common to cluster randomised trials as it allows for multiple outcomes, it can incorporate other features such as restricted randomisation methods, which are often used in trials with a small number of clusters [Watson et al., 2021, Li et al., 2017], and it provides exact inference when there are a small number of clusters, which can lead to non-nominal error rates of standard test procedures and hence confidence intervals with non-nominal coverage. Several small-sample corrections exist that can provide nominal error rates with a small number of clusters [Watson et al., 2021], however there is no obvious way these would be incorporated efficiently into a multiple testing procedure.

We have used optimal methods for: (i) correction for multiple testing; (ii) choice of test statistic; and (iii) confidence interval limit search. However, while each might be optimal for its chosen problem, in combination they may still be inefficient. Since there are no current alternatives for the problem we consider here though, we cannot compare power
under different approaches and we did not consider power. These methods are useful for
the analysis of cluster trials with multiple outcomes and the treatment effect parameters
from the linear predictors of multiple univariate models, however, it is not clear how or
if they could be applied to cluster trials with multiple arms. In multi-arm trials there
may be one or more outcomes, but clusters may receive different ‘doses’ or variants of the
treatment. There are a variety of treatment effects and null hypotheses of interest including
pairwise comparisons between arms and a global joint null, which can be estimated from
a single univariate model with indicators for each arm [Watson et al., 2021]. Pairwise null
hypotheses in these models do not make statements about the value of the treatment effects
in arms outside the pair under comparison as it is left unspecified, so it is not obvious then
how a randomisation test could be conducted for the pairwise comparison that is invariant
to randomised allocation. The multiple treatment effects of interest in a multi-arm study
clearly fall in the realm of multiple testing. Nevertheless, we believe the methods proposed
in this article will be a useful tool for the analysis of cluster randomised trials in many
cases.

References

The Lancet. Consort 2010. *The Lancet*, 375(9721):1136, 2010. ISSN 01406736. doi:
10.1016/S0140-6736(10)60456-4.

James M S Wason, Lynne Stecher, and Adrian P Mander. Correcting for multiple-testing
in multi-arm trials: is it necessary and is it done? *Trials*, 15(1):364, dec 2014. ISSN 1745-
6215. doi: 10.1186/1745-6215-15-364. URL https://trialsjournal.biomedcentral.com/articles/10.1186/1745-6215-15-364.
Joseph P. Romano and Michael Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108, 2005a. ISSN 01621459. doi: 10.1198/016214504000000539.

David M. Murray. *Design and Analysis of Group Randomised Trials*. Oxford University Press Inc., New York, NY, 1998.

Karla Hemming, Richard Lilford, and Alan J. Girling. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *Statistics in Medicine*, 34(2):181–196, jan 2015. ISSN 02776715. doi: 10.1002/sim.6325. URL [http://doi.wiley.com/10.1002/sim.6325](http://doi.wiley.com/10.1002/sim.6325).

Sandra Eldridge and Sally Kerry. *A Practical Guide to Cluster Randomised Trials in Health Services Research*. John Wiley & Sons, Ltd, Chichester, UK, jan 2012. ISBN 9781119966241. doi: 10.1002/9781119966241. URL [http://doi.wiley.com/10.1002/9781119966241](http://doi.wiley.com/10.1002/9781119966241).

Thrive at Work Wellbeing Programme Collaboration. Evaluation of a policy intervention to promote the health and wellbeing of workers in small and medium sized enterprises - a cluster randomised controlled trial. *BMC public health*, 19(1):493, may 2019a. ISSN 1471-2458. doi: 10.1186/s12889-019-6582-y. URL [https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-6582-y](https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-6582-y).

Elizabeth L Dunbar, Emily B Wroe, Basimyene Nhlema, Chiyoumbekezo Kachimanga, Ravi Gupta, Celia Taylor, Annie Michaelis, Katie Cundale, Luckson Dullie, Arnold Jumbe,
Lawrence Nazimera, Ryan McBain, Richard J Lilford, and Samuel Ian Watson. Evaluating the impact of a community health worker programme on non-communicable disease, malnutrition, tuberculosis, family planning and antenatal care in Neno, Malawi: protocol for a stepped-wedge, cluster randomised controlled trial. *BMJ Open*, 8(7): e019473, jul 2018. ISSN 2044-6055. doi: 10.1136/bmjopen-2017-019473. URL [http://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2017-019473](http://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2017-019473).

US Food and Drug Administration. Multiple Endpoints in Clinical Trials. Guidance for Industry. Technical report, Silver Spring, MD, USA, 2017.

Pierre Lafaye de Micheaux, Benoit Liquet, Sébastien Marque, and Jérémie Riou. Power and Sample Size Determination in Clinical Trials with Multiple Primary Continuous Correlated Endpoints. *Journal of Biopharmaceutical Statistics*, 24(2):378–397, mar 2014. ISSN 1054-3406. doi: 10.1080/10543406.2013.860156. URL [https://www.tandfonline.com/doi/full/10.1080/10543406.2013.860156](https://www.tandfonline.com/doi/full/10.1080/10543406.2013.860156).

Mark Rubin. When to adjust alpha during multiple testing: a consideration of disjunction, conjunction, and individual testing. *Synthese*, jul 2021. ISSN 0039-7857. doi: 10.1007/s11229-021-03276-4. URL [https://link.springer.com/10.1007/s11229-021-03276-4](https://link.springer.com/10.1007/s11229-021-03276-4).

Andrew Gelman, Jennifer Hill, and Masanao Yajima. Why We (Usually) Don’t Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211, 2012. ISSN 19345747. doi: 10.1080/19345747.2011.618213.

Alessio Farcomeni. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, 27.
Sandrine Dudoit, Juliet Popper Shaffer, and Jennifer C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003. ISSN 08834237. doi: 10.1214/ss/1056397487.

Sture Holm. Board of the Foundation of the Scandinavian Journal of Statistics A Simple Sequentially Rejective Multiple Test Procedure Author ( s ): Sture Holm Published by : Wiley on behalf of Board of the Foundation of the Scandinavian Journal of Statistics Stable U. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Joseph P. Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005b. ISSN 00129682. doi: 10.1111/j.1468-0262.2005.00615.x.

Samuel I. Watson, Alan J. Girling, and K. Hemming. Design and analysis of three-arm parallel group randomised trials with small numbers of clusters. *Statistics in Medicine*, 40(5):1133–1146, 2021.

Clémence Leyrat, Katy E Morgan, Baptiste Leurent, and Brennan C Kahan. Cluster randomized trials with a small number of clusters: Which analyses should be used? *International Journal of Epidemiology*, 47(1):321–331, 2018. ISSN 14643685. doi: 10.1093/ije/dyx169.

Joseph P. Romano and Michael Wolf. Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics and Probability Letters*, 113:38–40, 2016. ISSN 01677152. doi: 10.1016/j.spl.2016.02.012. URL http://dx.doi.org/10.1016/j.spl.2016.02.012
Richard Hooper, Steven Teerenstra, Esther de Hoop, and Sandra Eldridge. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Statistics in Medicine*, 35(26):4718–4728, nov 2016. ISSN 02776715. doi: 10.1002/sim.7028. URL [http://doi.wiley.com/10.1002/sim.7028](http://doi.wiley.com/10.1002/sim.7028).

Mitchell H. Gail, Steven D. Mark, Raymond J. Carroll, Sylvan B. Green, and David Pee. On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, 15(11):1069–1092, 1996. ISSN 02776715. doi: 10.1002/(SICI)1097-0258(19960615)15:11⟨1069::AID-SIM220⟩3.0.CO;2-Q.

John A. Gallis, Fan Li, Hengshi Yu, and Elizabeth L. Turner. Cvcrand and Cptest: Commands for Efficient Design and Analysis of Cluster Randomized Trials Using Constrained Randomization and Permutation Tests. *The Stata Journal: Promoting communications on statistics and Stata*, 2018. ISSN 1536-867X. doi: 10.1177/1536867x1801800204.

Thomas M. Braun and Ziding Feng. Optimal permutation tests for the analysis of group randomized trials. *Journal of the American Statistical Association*, 96(456):1424–1432, 2001. ISSN 1537274X. doi: 10.1198/016214501753382336.

Daniel McNeish and Laura M Stapleton. Modeling Clustered Data with Very Few Clusters. *Multivariate Behavioral Research*, 51(4):495–518, 2016. ISSN 00273171. doi: 10.1080/00273171.2016.1167008. URL [http://dx.doi.org/10.1080/00273171.2016.1167008](http://dx.doi.org/10.1080/00273171.2016.1167008).

Olivier Guilbaud. Simultaneous confidence regions corresponding to Holm’s step-down procedure and other closed-testing procedures. *Biometrical Journal*, 50(5):678–692, 2008. ISSN 03233847. doi: 10.1002/bimj.200710449.

Anthony J. Hayter and Jason C. Hsu. On the relationship between stepwise decision
procedures and confidence sets. *Journal of the American Statistical Association*, 89(425):128–136, 1994. ISSN 1537274X. doi: 10.1080/01621459.1994.10476453.

Paul Garthwaite and Stephen Buckland. Generating Monte Carlo Confidence Intervals by the Robbins-Monro Process Author(s): Paul H. Garthwaite and Stephen T. Buckland Source: *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 41, No. 1 Published by: Wiley. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 41(1):159–171, 1992.

Paul H Garthwaite. Confidence Intervals from Randomization Tests. *Biometrics*, 52(4):1387, 1996. ISSN 0006341X. doi: 10.2307/2532852.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, sep 1951. ISSN 0003-4851. doi: 10.1214/aoms/1177729586. URL [http://projecteuclid.org/euclid.aoms/1177729586](http://projecteuclid.org/euclid.aoms/1177729586).

David Ruppert. A Newton-Raphson Version of the Multivariate Robbins-Monro Procedure. *Annals of Statistics*, 13(1):234–245, 1985.

Fan Li, Elizabeth L Turner, Patrick J Heagerty, William M Vollmer, Elizabeth R De-long, and David M Murray. An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes. *Statistics in Medicine*, (October 2016):3791–3806, 2017. doi: 10.1002/sim.7410.

Thrive at Work Wellbeing Programme Collaboration. Evaluation of a policy intervention to promote the health and wellbeing of workers in small and medium sized enterprises— a cluster randomised controlled trial. *BMC Public Health*, 19(1):493, dec 2019b. ISSN 1471-2458. doi: 10.1186/s12889-019-6582-y. URL [https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-6582-y](https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-6582-y).
A Simulation details

A.1 (1) Two-arm, parallel cRCT with two outcomes and no covariates

For the first simulation we consider a parallel cluster trial with two outcomes measured once in the post-intervention period. Both outcomes are specified as Poisson distributed variables:

$$Y_{j,ic} \sim \text{Poisson}(\exp(\mu_j + \delta_j D_c + \theta_{j,c}))$$  \hspace{1cm} (20)

for $j = 1, 2$, where $\theta_{j,c} \sim N(0, \sigma^2_\theta)$. We set $\mu_j = 1$ and consider both $\delta = (0, 0)$ and $\delta = (0, 0.5)$ to compare the FWER under different combinations of true null hypotheses.

Analyses of cluster randomised trials usually consider the intraclass correlation coefficient (ICC) $\rho_j = \text{Var}(\theta_{j,c})/\text{Var}(Y_{j,ic})$, however the ICC for non-linear models depends on the realised values of the covariates and the parameter values and so will differ between simulations. We choose $\sigma^2_\theta$ to give a reasonable range of ICCs; for this simulation $\sigma^2_\theta = 0.2^2$ gives ICCs in the range of approximately 0.01 to 0.10.
A.2 (2) Two-arm, parallel cRCT, three outcomes, and covariates

We next step up the complexity of the design by introducing covariates and including three outcomes of different distributions:

\[ Y_{1,ic} \sim \text{Poisson}(\exp(\mu_1 + \delta_1 D_c + \beta_1 X_{ic} + \theta_{1,c})) \] (21)

\[ Y_{2,ic} \sim \text{Bernoulli}(\logit(\mu_2 + \delta_2 D_c + \beta_2 X_{ic} + \theta_{2,c})) \] (22)

\[ Y_{3,ic} \sim N(\mu_3 + \delta_3 D_c + \beta_3 X_{ic} + \theta_{3,c}, \sigma^2) \] (23)

We maintain the same values of \( \sigma^2 \theta \) and the numbers of clusters and individuals per cluster as the previous simulation. Similarly \( \mu_j = 1 \) and \( \beta_j = 1 \) for all \( j = 1, 2, 3 \). We vary the choice of \( \delta \) as either \((0, 0, 0), (0, 0, 0.5), \) or \((0, 0.5, 0.5)\); as with the following set of simulations we do not consider a completely exhaustive set of permutations of simulation parameters.

A.3 (3) Stepped-wedge cRCT with four outcome measures and covariates

For the final set of simulations we reproduce a simulation used in the design of an ongoing stepped-wedge cluster randomised trial. The trial evaluates an intervention designed to enable and facilitate the use of remote (via telephone or smartphone) primary care consulting in a number of low and middle income countries. The intervention was rolled out to 20 primary care clinics in a stepped-wedge design over ten monthly steps comprising two clusters per step. Patients with a set of pre-specified chronic conditions were included in the study as they regularly used the clinic.

The researchers were interested in the effect of the intervention on: (i) the rate of remote consulting (number of remote consultations per patient per month); (ii) the rate...
of face-to-face consulting (number of face-to-face consultations per patient per month) – it
was unknown if remote consultation functions as a substitute, complement, or neither for
face-to-face consultation; (iii) the prescribing rate of medication (number of prescriptions
issued per patient per month), which was seen as the primary safety endpoint; and (iv)
the trust the patient had in the consultation process (evaluated using a ‘trust score’ from
a specialised survey instrument). All four outcomes were therefore specified as ‘primary
outcomes’. We present analysis of the real data from this trial in Section 6.

We simulate these outcomes as:

\[
Y_{1,ict} \sim \text{Poisson}(\exp(\mu_1 + \delta_1 D_{ct} + \beta_1 X_{ct} + \theta_{1,c} + \theta_{1,ct}))
\]  

\[
Y_{2,ict} \sim \text{Poisson}(\exp(\mu_2 + \delta_2 D_{ct} + \beta_2 X_{ct} + \theta_{2,c} + \theta_{2,ct}))
\]  

\[
Y_{3,ict} \sim \text{Poisson}(\exp(\mu_3 + \delta_3 D_{ct} + \beta_3 X_{ct} + \theta_{3,c} + \theta_{3,ct}))
\]  

\[
Y_{4,ict} \sim N(\mu_4 + \delta_4 D_{c} + \beta_4 X_{ct} + \theta_{4,c} + \theta_{4,ct}, \sigma^2)
\]

We set the parameters as \(\mu = (\log(1), \log(0.5), \log(10), 0), \beta = (1, 1, 1, 1), \theta_{j,c} \sim N(0, 0.2^2), \theta_{j,ct} \sim N(0, 0.09^2)\) (implying a cluster autocorrelation coefficient of 0.8), and \(\sigma^2 = 1\). We
examine two sets of treatment effect parameters: \(\delta = (0, 0, 0, 0)\) and \(\delta = (0, 0.5, 0, 1)\).

A.4 (4) Two-arm, parallel cRCT with three outcomes and no
covariates

For the final set of simulation experiments, generated data from the following model

\[
Y_{i,ct} \sim MVN(\eta, \Sigma)
\]  

where \(Y_{i,ct} = [Y_{1,ict}, Y_{2,ict}, Y_{3,ict}]\), \(\eta = [\eta_{1,ict}, \eta_{2,ict}, \eta_{3,ict}]\) and \(\eta_{j,ict} = \delta_{j} D_{c} + \theta_{j,c}\) and \(\Sigma\) is a
3 \times 3 covariance matrix with diagonal elements of 1 and off-diagonal elements of \(\rho\). We set
the parameters as \(\theta_{ct} \sim N(0, 0.05)\) and compared \(\delta = (0, 0, 0)\) and \(\delta = (0, 0.5, 1)\).