Permutation-based multiple testing corrections for p-values and confidence intervals for cluster randomised trials

Watson, Samuel I.; Akinyemi, Joshua O.; Hemming, Karla

DOI:
10.1002/sim.9831

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Watson, SI, Akinyemi, JO & Hemming, K 2023, 'Permutation-based multiple testing corrections for p-values and confidence intervals for cluster randomised trials', *Statistics in Medicine*. https://doi.org/10.1002/sim.9831

Received: 22 August 2022 | Revised: 17 May 2023 | Accepted: 29 May 2023

DOI: 10.1002/sim.9831

RESEARCH ARTICLE

Statistics in Medicine WILEY

Permutation-based multiple testing corrections for *P*-values and confidence intervals for cluster randomized trials

Samuel I. Watson[1] | **Joshua O. Akinyemi**[2] | **Karla Hemming**[1]

[1]Institute of Applied Health Research, University of Birmingham, Birmingham, UK

[2]College of Medicine, University of Ibadan, Ibadan, Oyo State, Nigeria

Correspondence
Samuel I. Watson, Institute of Applied Health Research, University of Birmingham, Birmingham, UK.
Email: s.i.watson@bham.ac.uk

Funding information
Medical Research Council, Grant/Award Number: MR/V038591/1; National Institute for Health and Care Research, Grant/Award Numbers: EP/V028936/1, SRF-2017-10-002

In this article, we derive and compare methods to derive *P*-values and sets of confidence intervals with strong control of the family-wise error rates and coverage for estimates of treatment effects in cluster randomized trials with multiple outcomes. There are few methods for *P*-value corrections and deriving confidence intervals, limiting their application in this setting. We discuss the methods of Bonferroni, Holm, and Romano and Wolf and adapt them to cluster randomized trial inference using permutation-based methods with different test statistics. We develop a novel search procedure for confidence set limits using permutation tests to produce a set of confidence intervals under each method of correction. We conduct a simulation-based study to compare family-wise error rates, coverage of confidence sets, and the efficiency of each procedure in comparison to no correction using both model-based standard errors and permutation tests. We show that the Romano-Wolf type procedure has nominal error rates and coverage under non-independent correlation structures and is more efficient than the other methods in a simulation-based study. We also compare results from the analysis of a real-world trial.

KEYWORDS
cluster randomized trial, coverage, inference, multiple testing

1 | INTRODUCTION

For a randomized controlled trial, the requirement to state a single primary outcome has become accepted, even required, practice. For example, the influential CONSORT statement on clinical trials requires the pre-specification of a single primary outcome, which they describe as the "outcome considered to be of greatest importance to relevant stakeholders," and recommends against multiple primary outcomes.[1] The reason for this is to ensure appropriate control of the "false discovery rate" when using null hypothesis significance testing.[2] 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.[3] Indeed, the CONSORT statement notes that that multiple primary outcomes are not recommended as it "incurs the problem of multiplicity of analyses."[4]

Cluster randomized 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.[5-7] 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,[8] or a community health worker program targeting multiple health conditions.[9] The effects of such complex interventions cannot be adequately summarized by a single outcome. 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 randomized 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 randomized trials, and particularly cluster randomized 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."[10] 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. Other authors differentiate aiming to declare "statistical significance" on at least one of a group of null hypotheses to requiring statistical significance for all tests in order to reject any individual test, and propose different solutions for both.[11,12]

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 favored by Bayesian practitioners;[13] and (ii) univariate solutions that aim to ensure inferential statistics for a set of estimands collectively have the appropriate frequentist properties.[14] In this article, we focus on the latter approaches in a frequentist setting. Despite the different approaches and guidance, Wason et al[2] estimated that only around half of all randomized trials with multiple outcomes or arms corrected for multiple testing. No evidence is available on the use of corrections for multiple testing in cluster randomized trials specifically, but there are few, if any, comprehensive discussions of methods in this area currently available. Furthermore, almost all discussion of multiple testing adjustment relates to corrections for *P*-values, with few, if any, solutions for confidence intervals. The FDA note that correcting confidence intervals is complex and beyond the scope of their advice. However, the duality between hypothesis testing and confidence intervals means that we should be able to identify the bounds of a "confidence set" adjusted for multiple testing.[3,15] The primary limiting factor to using corrected confidence intervals is that there are no proposed methods for determining these bounds efficiently.

In this article, we develop several methods for adjusting *P*-values for multiple testing for a cluster randomized trial setting using permutation-test based methods, by adapting existing methods of correction, and propose a novel method to derive corrected confidence sets. We then compare these methods in a simulation-based study to evaluate type I error rates and efficiency of the different procedures. Our analysis is based on generalized linear mixed models, which are frequently used in the analysis of cluster trials. We also focus on permutation-based methods, since these methods provide exact inference at all sample sizes. A small number of clusters, which is common to many cluster trials, can result in small sample biases in the standard error estimator and inflated type 1 errors,[16-18] which results in complication when it comes to considering additional corrections for multiple testing. Section 2 provides a review and discussion of methods for correcting for multiple testing and their adaptation to a cluster randomized trials setting, Section 3 presents a simulation-based comparison, Section 4 provides an applied example, and Section 5 concludes.

2 | MULTIPLE TESTING IN CLUSTER RANDOMIZED TRIALS

2.1 | The multiple testing problem

We first suppose that data $\mathbf{X}$ are generated from some probability distribution P, which belongs to some family of probability distributions Ω. The family Ω could be a parametric, semi-parametric, or non-parametric model. The multiple testing problem arises when we have a set of hypotheses H_j vs $H_{j'}$ for $j = 1, \ldots, J$, following the notation of Romano and Wolf.[3] These hypotheses in our context are typically estimates of the treatment effect of an intervention on multiple outcomes. Each of the hypotheses is a subset $\omega_j \subset \Omega$ and is equivalent to testing $P \in \omega_j$ against $P \notin \omega_j$. 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 α-quantile of the distribution of T_j as $c_j(\alpha, P)$. In a traditional null hypothesis testing framework we "reject" H_j in favor of $H_{j'}$ at the α level, if $T_j \geq c_j(1-\alpha, P)$, which clearly has probability α. 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 $H_{j\in I}$, that is, $Pr(\cup_{j\in I} T_j > c_j(1-\alpha, P))$, which should be α.

2.2 | Methods for correcting for multiple testing

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.[19] 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 $c_j(1-\alpha/J, P)$. Equivalently, P-values that maintain the FWER for the family of null hypotheses ensure that $Pr(\cup_{j\in I} T_j > c_j(1-p, P)) = p$, so a crude "corrected" P-value for the null hypothesis H_j using the Bonferroni method would be $\min(Jp_j, 1)$. However, while this method exerts strong control over the FWER, it is highly conservative.

Holm[20] 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. A crude corrected P-value could therefore be obtained by multiplying the smallest to the largest P-values by J, $J-1$ and so forth, respectively. The Holm method is less conservative than the Bonferroni method,[20] but it may still be inefficient as, like the Bonferroni method, it does not explicitly take into account the dependence structure in the data. Romano and Wolf[3,15] developed an efficient resampling based version of Holm's stepdown method, which can use permutation-based tests in the context of a cluster randomized trial.

2.3 | Permutation-based corrections for multiple testing

An issue that complicates analyses of cluster randomized trials is that test statistics can fail to have the expected sampling distribution in a range of circumstances, but particularly when the number of clusters is small.[16-18,21] This issue means determining the critical value of a hypothesis test, even in the absence of any multiple testing issue, can be difficult. While there exist several small sample corrections in the literature their performance often depends on the correlation structure, which is not known.[16,18]

An alternative approach is to use a permutation testing method based on the randomization scheme for the trial. In particular, the null hypothesis implies that the distribution of the data $\mathbf{X}$ is invariant under a set of transformations in $\mathbf{A}$, which has L elements. So, $a\mathbf{X}$ and $\mathbf{X}$ have the same distribution for all $a \in \mathbf{A}$ whenever $\mathbf{X}$ has distribution $P \in \omega$. $\mathbf{A}$ in the context of cluster randomized trials is the set of all transformations that could be generated by the randomization mechanisms, for example, all ways of dividing the clusters into two groups for a parallel design. Our observed test statistics with our sample data are $T_j(\mathbf{X})$. The test statistic generated by the lth permutation is $T_j(a_l\mathbf{X})$ for $a_l \in \mathbf{A}$ and $l = 1, \ldots, L$. We can use this approach to estimate the critical values for the Bonferroni or Holm corrections. For example, for Bonferroni:

$$\hat{c}_j(1-\alpha/J, P) = T_{j,|L(1-\alpha/J)|}, \tag{1}$$

where $T_{j,|L(1-\alpha/J)|}$ is the $L(1-\alpha/J)$th (or nearest integer) largest value from the permutations. And a crude, corrected two-sided P-value is:

$$p_j = \min\left(\frac{J}{L}\sum_{l=1}^{L}\mathbf{1}\left[\text{abs}(T_j(a_l\mathbf{X})) \geq \text{abs}(T_j(\mathbf{X}))\right], 1\right), \tag{2}$$

where $\mathbf{1}$ is the indicator function and abs is the absolute value. The same approach can be used for the Holm method.

Romano and Wolf[3,15] developed a modified stepdown approach to take advantage of resampling methods. Their process is optimal in a maximin sense. We describe the general stepdown procedure of Romano and Wolf first in terms of accepting or rejecting each null hypothesis at an α-level. We let $c_{\mathbf{K}}(\alpha, P)$ denote an α-quantile of the distribution of the statistic:

$$T_K = \max_{j\in K} T_j, \tag{3}$$

for any subset of null hypotheses K. We also denote $T_{|r|}$ as the rth largest test statistic so that

$$T_{|1|} \geq T_{|2|} \geq \cdots \geq T_{|J|}, \tag{4}$$

corresponding to hypotheses $H_{|1|}, H_{|2|}, \ldots, H_{|J|}$. Then the idealized 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;

⋮

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. One can see that this algorithm replicates Holm's procedure, but allows us to use permutation-based methods to estimate the critical values where they are not known.

For each permutation we can determine the test statistic as in Equation (3) as $T_{K,l} = \max_{j \in K} T_j(a_l \mathbf{X})$. As before we denote $T_{K,|r|}$ as the rth largest of all the permutational test statistics $\{T_{K,l}; l = 1, \ldots, L\}$. Then our estimator for the critical value is:

$$\hat{c}_K(1-\alpha, P) = T_{K,|L(1-\alpha)|}. \tag{5}$$

We can see how this procedure produces P-values for a two-sided hypothesis that also maintains the FWER for a given α,[22] in particular:

$$p_K = \frac{1}{L}\sum_{l=1}^{L} \mathbf{1}\left[\text{abs}(T_K(a_l\mathbf{X})) \geq \text{abs}(T_K(\mathbf{X}))\right]. \tag{6}$$

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

Often the size of $\mathbf{A}$ 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 $\mathbf{A}$ of fixed sized in order to generate realizations of the test statistics. If we conduct M such permutations then the estimator of the P-value for a given null hypothesis vs some alternative is

$$\hat{p}_K = \frac{1}{M+1}\sum_{m=1}^{M} \left(1 + \mathbf{1}[\text{abs}(T(a_m\mathbf{X})) \leq \text{abs}(T(\mathbf{x}))]\right). \tag{7}$$

Obtaining P-values in this way is described in detail by Romano.[22] Values of $M = 1000$ or greater are often used as this results in relatively small Monte Carlo error, although much larger values (eg, 10 000 or 100 000) may be preferred for formal or final analyses.

In subsequent sections, we develop and compare Bonferroni, Holm, and Romano-Wolf methods, however, we note there are several other multiple testing corrections in the literature, including Hochberg's "step-up" procedure,[23] Hommel's "stagewise" procedure,[24] and Šidák's procedures[25] (see also Reference 14 for a discussion). More exhaustive comparisons of these methods in other settings, such as References 26-29, show that they all maintain a FWER $\leq \alpha$, but that Holm's, Hommel's, and Hochberg's procedures generally are the most efficient and perform very similarly. However, these comparisons do not include the Romano-Wolf method, which purports to be at least as efficient as Holm's procedure.[3] We note that Westfall and Young[30] propose an early version of a resampling based multiple testing correction similar to Romano-Wolf, which is included in the comparison by Alberton et al[29] in the context of modeling brain imaging data. We adapt only a subset of all methods, but believe the application of other methods in the context we describe below, including any developed after the publication of this article, should be clear from the discussion of these four key approaches.

2.4 | Permutation test statistics for cluster trials

We next introduce a generalized linear mixed model commonly used in the analysis of cluster randomized trials.[31] We denote Y_{ict} as the outcome of the ith individual, $i = 1, \ldots, N$, in cluster $c = 1, \ldots, C$ at time $t = 1, \ldots, T$. We include a temporal dimension in this discussion for generality, however, it can be ignored as required. Our simulation-based

comparisons include both examples with and without a temporal dimension. We do not restrict the outcome, it could be continuous or discrete. We specify the linear predictor:

$$\eta_{ict} = \mu_0 + \delta D_{ct} + X'_{ict}\beta + \theta_{ct}, \tag{8}$$

where D_{ct} is an indicator for whether cluster c has received the intervention at time t and so δ 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 θ_{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

$$Y_{ict} \sim P(h(\eta_{ict})), \tag{9}$$

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

Gail et al[32] 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.[33] Several other authors have also developed and evaluated permutation-tests and test statistics in the context of cluster trials.[34-39] Here, we build on the statistic proposed by Braun and Feng.[40]

Braun and Feng[40] examine optimal permutation tests for cluster randomized trials specifically. They derive a "quasi-score" statistic using the marginal likelihood of the data modeled separately from the correlation structure of the data. The marginal mean of each observation, ignoring the cluster-effects θ_{ct}, is

$$h^{-1}(\mu_{ict}) = \mu_0 + \delta D_{ct} + X'_{ict}\beta. \tag{10}$$

The "quasi-score" statistic, which is weighted sum of generalized residuals, is then:

$$\sum_c \{D^*_c \mathbf{G}_c \mathbf{V}_c^{-1}[\mathbf{Y}_c - \boldsymbol{\mu}_c]\}|_{\delta=\delta^*}, \tag{11}$$

where $D^*_c[D^*_{c1}, D^*_{c1}, D^*_{c1}, \ldots, D^*_{cT}, D^*_{cT}]'$ 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. $\mathbf{G}_c$ is a $(1 \times n_c)$ vector with elements $(\partial h^{-1}_{ict}/\partial \eta_{ict})^{-1}$, and $\mathbf{V}_c$ 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 μ_{ict}, identity link function, variance σ^2, and $\theta_{ct} \sim N(0, \tau^2)$, then the diagonal elements of $\mathbf{V}_c$ are $\sigma^2 + \tau^2$ and the off-diagonal elements are τ^2. More complex structures might include temporal decay in correlation, for example. We use Θ to represent the parameters of the variance-covariance matrix. Finally $[\mathbf{Y}_c - \boldsymbol{\mu}_c]$ are generalized residuals: $\mathbf{Y}_c = [Y_{1c1}, Y_{2c}, \ldots, Y_{n_{c1}c1}, Y_{1c2}, \ldots, Y_{n_{cT}cT}]$ is a $(1 \times n_c)$ vector of outcomes and $\boldsymbol{\mu}_c$ is a $(1 \times n_c)$ vector of means.

For the permutation test to be valid the "nuisance" parameters (μ, β, Θ), that is, those other than δ, must be invariant to permutation.[40] 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:

$$\hat{\mu}_{ict} = h(\hat{\mu}_0 + \delta^* D_{ct} + X'_{ict}\hat{\beta}), \tag{12}$$

for the linear predictor under the null $H_0 : \delta = \delta^*$. Estimating Θ is more difficult, however, particularly when the number of clusters is small.[16,21] As an alternative to (11) we can replace $\mathbf{G}_c\mathbf{V}_c^{-1}$ with a $(1 \times n_c)$ vector of ones:

$$\sum_c \sum_t \sum_i \{D^*_{ict}[Y_{ict} - \mu_{ict}]\}|_{\delta=\delta^*} \tag{13}$$

so that the sum of residuals is "weighted" only by the size of each cluster or cluster-time period. One can see that under homoscedasticity the two test statistics will be approximately proportional. The weighted statistic weights the residuals in proportion to their variance, so in non-linear models with differing variances (eg, different linear predictors over time) we may expect to see an improvement in efficiency.

The quasi-score statistics are the motivation behind quasi-likelihood approaches, including GEE methods.[40,41] Thus, the tests and corrections described here can be implemented within a GEE framework. However, our simulations in

Section 3, model estimation, and the software we provide to implement the methods uses a more explicitly GLMM formulation. The quasi-score statistic is equivalent for full and marginal likelihoods using linear Gaussian models, or when using the "unweighted" variant described below. For non-linear alternatives though, the quasi-score statistic is an approximation to the full likelihood. In terms of our implementation of the computation of (11), we use a GLMM formulation (Equation 8) and use the original estimates of the covariance parameters to generate an estimated inverse covariance matrix $\hat{\mathbf{V}}^{-1}$, which is then re-used for each iteration.

For the purposes of correcting for multiple testing we use studentized versions of the two test statistics:

$$T_w = T_w(\mathbf{X})|_{\delta=\delta_0} = \frac{\sum_c \{D_c^* \mathbf{G}_c \mathbf{V}_c^{-1}[\mathbf{Y}_c - \boldsymbol{\mu}_c]\}}{\sqrt{\sum_c \{D_c^* \mathbf{G}_c \mathbf{V}_c^{-1}[\mathbf{Y}_c - \boldsymbol{\mu}_c]\}^2}}, \tag{14}$$

$$T_u = T_u(\mathbf{X})|_{\delta=\delta_0} = \frac{\sum_c \sum_t \sum_i \{D_{ict}^*[Y_{ict} - \mu_{ict}]\}}{\sqrt{\sum_c \sum_t \sum_i \{D_{ict}^*[Y_{ict} - \mu_{ict}]\}^2}}, \tag{15}$$

where the terms on the right-hand side have been evaluated at $\delta = \delta^*$. We describe T_w as the "weighted test statistic" and T_u as "unweighted." 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.[3] 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 permutational methods will have joint coverage of $1 - \alpha$, without balance the individual coverage probabilities of each interval will differ, perhaps substantially.[15]

2.5 | 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 δ_j with associated confidence intervals $[L_j, U_j]$, so that $[L_1, U_1] \times [L_2, U_2] \cdots \times [L_J, U_J]$, $\mathbf{U} = [U_1, \ldots, U_J]$ and $\mathbf{L} = [L_1, \ldots, L_J]$, 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:

$$Pr(\cup_j \delta_j \in [L_j, U_j]) = 1 - \alpha, \tag{16}$$

we refer to this as "family-wise coverage," which we use analogously to "simultaneous coverage" used in other contexts. 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 α. For Bonferroni, an obvious modification is to instead estimate $100(1 - \alpha/J)\%$ confidence intervals to achieve a family-wise coverage of $100(1 - \alpha)\%$. There have been some attempts to construct exact confidence sets for parameters analytically based on the stepdown procedure.[15] For example, Guilbaud,[42] extending the proposal of Hayter and Hsu,[43] 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 δ_1. Its $100(1 - \alpha)\%$ confidence interval is $[L_1, U_1]$: for any value δ_1^* inside this interval the null hypothesis $H_1 : \delta_1 = \delta_1^*$ will not be rejected in favor of the two-sided alternative $H_{1'} : \delta_1 \neq \delta_1^*$ at the α level. The question is then how to find the values of L_1 and U_1 efficiently. 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 $\mathbf{U}$ we must calculate JM permutation test statistics and perform the desired correction. Moreover, since the test statistic and its permutational 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[44] developed a method for searching for confidence interval endpoints efficiently, which Garthwaite[45] later adapted for use with permutation tests. Their method is based on the search process devised by Robbins and Munro,[46] who developed a stochastic approximation procedure to find the α-quantile of a particular distribution.

Multivariate Robbins-Munro processes follow the same procedures as their univariate equivalents.[47] For our multiple testing scenario the upper limits to the confidence set correspond to where all hypotheses $H_j : \theta_j = U_j$ for $j = 1, \ldots, J$ are all rejected in favor of the two-sided alternative with a FWER of α but for any smaller values of U_j not all hypotheses are rejected, and equivalently for the lower limits. Rabideau et al[48,49] have also independently proposed this method for confidence interval estimation for cluster randomized trials, although not in the context of multiple testing.

For each method, at the qth step of Q steps total, we have estimates of the upper confidence interval limits of our J parameters $\mathbf{u}_q = [u_{1q}, u_{2q}, \ldots, u_{Jq}]$. We generate the set of test statistics $T_j(\mathbf{X})|_{\delta=u_{jq}}$, which correspond to the null hypotheses $H_j : \delta_j = u_{jq}$. We then generate a single permutation of a permutation test for the same hypotheses $\text{abs}(T_j(a_q\mathbf{X}))_{\delta=u_{jq}}$. Each method then defines a procedure for determining whether to reject these hypotheses or not, which are described in the preceding sections. For example, with the Romano-Wolf stepdown procedure: reject hypothesis $H_{|1|}$ if $\text{abs}(T_{K_1}(a_q\mathbf{X})) < \text{abs}(T_{|1|}(\mathbf{X}))$ otherwise do not reject any hypothesis and stop; if $H_{|1|}$ was rejected then reject hypothesis $H_{|2|}$ if $\text{abs}(T_{K_2}(a_q\mathbf{X})) < \text{abs}(T_{|2|}(\mathbf{X}))$ otherwise do not reject any further hypotheses and stop, and so forth.

The estimates of the upper limits are updated based on the single permutation draw as (we drop the subscript $\delta = u_{jq}$ for ease of notation, but the test statistics are evaluated at this value for each iteration):

$$u_{j,q+1} = \begin{cases} u_{jq} - s_j\alpha^*/q & \text{if } H_j \text{ rejected} \\ u_{jq} + s_j(1-\alpha^*)/q & \text{otherwise,} \end{cases} \tag{17}$$

where s_j is the "step length constant." With no correction and with Romano-Wolf $\alpha^* = \alpha$, for Bonferroni $\alpha^* = \alpha/J$, and for Holm $\alpha^* = \alpha/J$ for $H_{|1|}$, $\alpha^* = \alpha/(J-1)$ for $H_{|2|}$, and so forth. Similarly for the lower limits, the updating rule is:

$$l_{j,q+1} = \begin{cases} l_{jq} + s_j\alpha^*/q & \text{if } H_j \text{ rejected} \\ l_{jq} - s_j(1-\alpha^*)/q & \text{otherwise.} \end{cases} \tag{18}$$

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:

$$k = \frac{2}{z_{1-\alpha}(2\pi)^{-1/2}\exp(-z_{1-\alpha}^2/2)}, \tag{19}$$

where z_α is the α-quantile of the standard normal distribution. The algorithm proceeds for a pre-selected number of iterations; in the simulations in the subsequent section we have used 2000 iterations. A sensible starting value for this algorithm is the approximate uncorrected confidence interval limits, for example, for the upper limit $u_j, 0 = \hat{\beta}_j + 2SE_j$ where SE_j is the standard error of β_j from the univariate model.

2.6 | Computation

An R package developed by the authors to execute the analyses described in this article is available from CRAN as `crct-Stepdown` (version 0.2.1 at the time of writing) including implementations of the Romano-Wolf, Holm, and Bonferroni methods for correcting P-values and confidence sets using permutation-based tests.

3 | SIMULATION STUDY

3.1 | Methods

We conduct a simulation-based study to examine the FWER, family-wise coverage, and efficiency of the procedures outlined in the previous sections for cluster randomized trials. We compare the following procedures:

1. A "naive" no correction approach using the reported standard errors and test statistics from the output of the `lme4` package for R. 95% confidence intervals for each parameter were constructed as $\hat{\delta} + / - 1.96 \times SE$.
2. No correction with P-values and confidence sets derived from permutation based tests.
3. The Bonferroni method using permutation based tests.
4. The Holm method using permutation based tests.
5. The Romano-Wolf method using permutation based tests.

For methods 2-5 we use both the weighted and unweighted test statistic resulting in nine methods. For the Bonferroni and Holm methods we only use permutation-based inference rather than the perhaps more standard approach of adjusting *P*-values reported by mixed model fitting software. Model-based inference can fail to have nominal FWERs for reasons other than multiple testing, such as biases arising from small numbers of clusters, which would further complicate interpretation of the results. We include a comparison with methods 1 and 2 to illustrate this issue in our context.

3.1.1 | Data generating processes

We use three different data generating processes of cluster randomized trials, described below. We opt for specific scenarios of rising complexity to examine the performance of the nine different methods (including both unweighted and weighed versions of the permutation-based methods). All outcomes are simulated and modeled using exponential-family models. In all simulations we set the number of individuals per cluster to 20 and simulate either seven or 14 clusters per arm. The choice of number of clusters is informed by two key considerations. First, the simulations take a very long time to run given the number of GLM models required to be estimated for the permutation tests and search procedures (for three outcomes and 10 000 iterations we require 90 million models), and so we aimed to choose the smallest number that would provide the desired inference. Second, we wanted to include scenarios where there was likely small sample bias in "standard" non-permutation based estimators of standard errors due to the low number of clusters, and one where such biases were likely minimal. Previous literature on cluster trials suggests small sample biases are likely minimal at 14 clusters or more per arm, but present with seven clusters per arm,[17] although permutation-based methods provide exact inference at any sample size. We provide estimates of FWER without correction and with non-permutation based estimators to examine whether there are likely small sample biases. However, we recognize that 14 clusters per arm may still be considered "small." The treatment effect parameters for each simulation are a vector, δ, with length equal to the number of outcomes and with different combinations of either 0 or 1, allowing for when all treatment effects are zero and when only a subset are.

(1) Two-arm, parallel cRCT, two outcomes

The first simulation data generating process ("model (1)") represents a two arm parallel cluster trial with two outcomes measured once in the post-intervention period. Both outcomes Y_j are continuous, Gaussian variables for $j = 1, 2$. This model is intended to examine the effect of correlation, which we model at the individual and cluster levels. For individual i in cluster c:

$$\begin{pmatrix} Y_{1,ic} \\ Y_{2,ic} \end{pmatrix} \sim \left(\begin{pmatrix} \mu_1 + \delta_1 D_c + \theta_{1,c} \\ \mu_2 + \delta_2 D_c + \theta_{2,c} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right), \tag{20}$$

where μ_j are intercept parameters, D_c is an indicator for whether the cluster is treated or not, and $\theta_{j,c}$ are cluster level random effect modeled as:

$$\begin{pmatrix} \theta_{1,c} \\ \theta_{2,c} \end{pmatrix} \sim \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \pi\tau_1\tau_2 \\ \pi\tau_1\tau_2 & \tau_2^2 \end{pmatrix} \right). \tag{21}$$

The parameters ρ and π are correlation parameters at the individual and cluster levels, respectively with σ_j and τ_j the standard deviation of the individual-level outcomes and cluster-random effect terms, respectively. Clusters are assigned in a 1:1 ratio with seven or 14 clusters per arm and 20 or 10 individuals per cluster. 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. We set $\sigma_j^2 = 1$ and $\tau_j^2 = 0.05$, which gives a marginal intraclass correlation coefficient (ICC) ($\text{ICC}_j = Var(\theta_{j,c})/Var(Y_{j,ic})$) of 0.05. We also set $\rho = \pi$ and examine a range of values. We do not report outcomes using the weighted test statistic with this example as it is proportional the unweighted test statistic as both models are Gaussian with identity link, so there will be no difference in performance.

(2) Two-arm, parallel cRCT, two differently distributed outcomes

For the next set of simulations ("model (2)") we consider a parallel cluster trial with two outcomes measured once in the post-intervention period. Simulation parameters are as the previous example, unless stated below. The first outcome is specified as Poisson distributed:

$$Y_{1,ic} \sim \text{Poisson}(\exp(\mu_1 + \delta_1 D_c + \theta_{1,c})),$$

and the second outcome as Gaussian distributed:

$$Y_{2,ic} \sim N(\mu_2 + \delta_2 D_c + \theta_{2,c}, 1),$$

where the random effects are specified as in Equation (21) with $\pi = 0$. We again set $\mu_j = 1$ and consider both $\delta = (0, 0)$ and $\delta = (0, 0.5)$. The ICC for non-linear models depends on the realized values of the covariates and the parameter values and so will differ between simulations. We again choose $\sigma_j^2 = 0.05$, which gives a range of ICCs between approximately 0.01 and 0.2 for the Poisson model and 0.05 for the Gaussian model.

(3) Two-arm parallel cRCT with baseline measures, three outcomes

We finally extend the parallel cluster trial model ("model (3)") to include baseline measures, which incorporates a temporal dimension and hence more complex covariance structure. The trial includes seven clusters in each arm, with half receiving the intervention in the second time period. We simulate three outcomes, with index t representing time period:

$$\begin{aligned} Y_{1,ict} &\sim \text{Poisson}(\exp(\mu_1 + \delta_1 D_{ct} + T_1 + \theta_{1,ct})) \\ Y_{2,ict} &\sim N(\mu_2 + \delta_2 D_{ct} + T_2 + \theta_{2,ct}, 1) \\ Y_{3,ict} &\sim \text{Bernoulli}(\text{logit}(\mu_3 + \delta_3 D_{ct} + T_3 + \theta_{3,ct})), \end{aligned} \tag{22}$$

where, now, D_{ct} equals one if the cluster has the intervention in time period t and zero otherwise and T is a fixed effect for the second time period. We use an auto-regressive specification for $\theta_{j,ct}$ to facilitate incorporation of correlation between outcomes. In particular,

$$Cov(\theta_{j,ct}, \theta_{j,ct'}) = \lambda^{|t-t'|}\sigma_j^2 \tag{23}$$

$$Cov(\theta_{j,ct}, \theta_{j',ct'}) = \lambda^{|t-t'|}\sigma_j\sigma_{j'}\rho \tag{24}$$

for $j \neq j'$. The random effects have a multivariate normal specification as before zero correlation. We maintain the same number of individuals per cluster. We set $\mu_j = -1$ and $\tau_j = 1$ for all $j = 1, 2, 3$. We vary the choice of δ as either $(0, 0, 0)$ or $(0, 0.5, 0)$; as with the previous set of simulations we do not consider a completely exhaustive set of permutations of simulation parameters. We set $\lambda = 0.7$.

3.1.2 | Simulation methods

Each set of simulations is run 10 000 times. We note the Monte Carlo error will be moderately higher than expected due to variation arising from the permutation tests, confidence set search procedure, and simulations. We use 1000 iterations for the permutation test P-values and 2000 steps for the search procedure as these produced stable values for these simulations (although we note that for more outcomes longer runs were often required for the confidence interval search procedure for it to reach a stable equilibrium). Point estimates of parameters were obtained from univariate generalized linear mixed models estimated with the R package `lme4` for models (1) and (2), we similarly obtained estimates of variance parameters from these models for the weighted test statistics. For example (3) we obtained parameter estimates from a generalized linear model with no random effects given the lack of widely available software for estimating autoregressive random effects models; weighted test statistics were generated using a covariance matrix created with the values of λ, σ_l, and ρ used in the data generating process.

3.1.3 | Evaluation

We estimate the FWER for $p \leq 0.05$, which has a nominal rate of 5%, and also estimate coverage of 95% confidence sets. We also estimate the mean 95% confidence interval width for each parameter δ to compare the efficiency of the procedures.

3.2 | Results

Figure 1 shows the family wise error rates and coverage from model (1) with the permutation-based methods for different levels of the correlation coefficient. We exclude the "naive" approach from these plots as it has non-nominal marginal

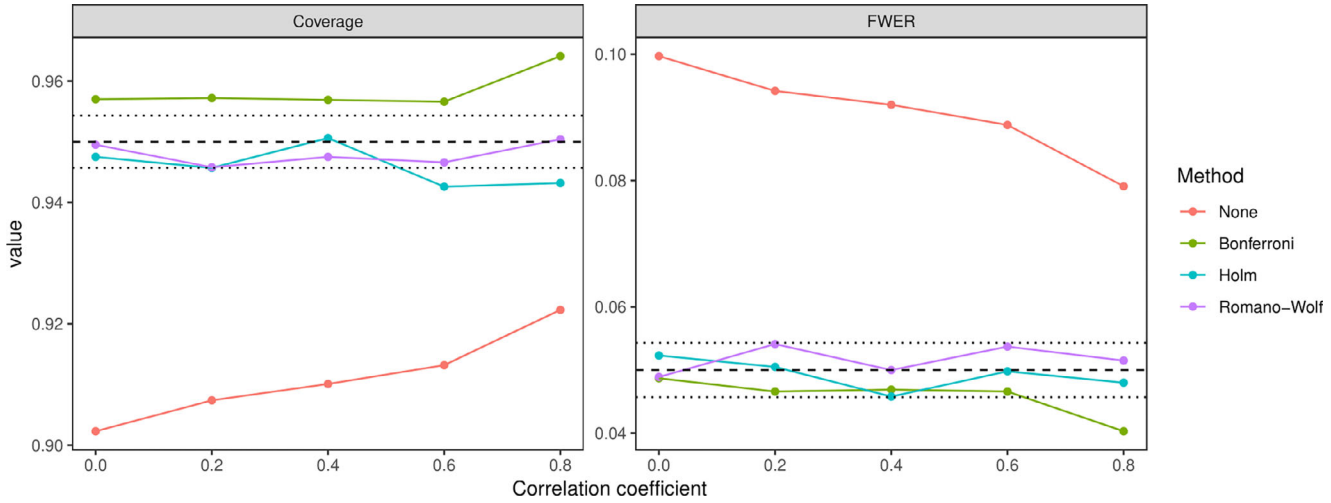

FIGURE 1 Family wise error rate and coverage under model (1) for four methods with different levels of the correlation coefficient ρ. The dashed line shows the nominal rates and the dotted lines approximate Monte Carlo confidence intervals. "None" refers to no correction.

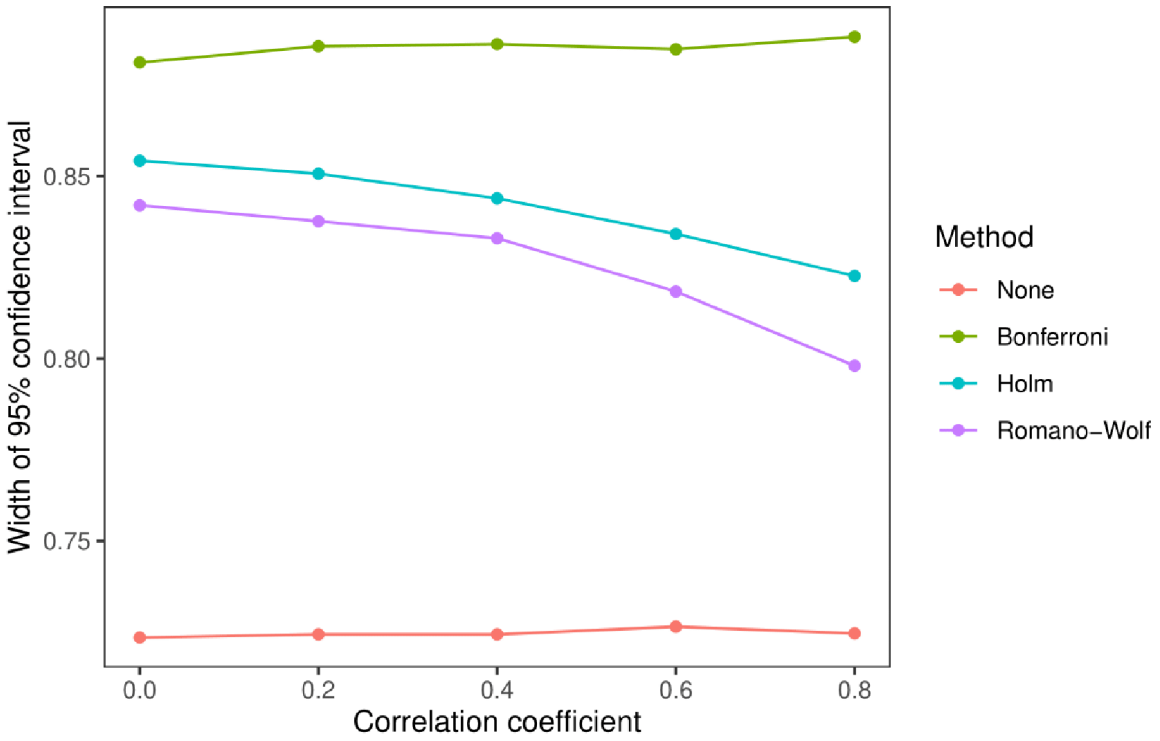

FIGURE 2 95% confidence interval width model (1) for four methods with different levels of the correlation coefficient ρ. "None" refers to no correction.

type I error and coverage without correction (see below). All three corrections ensured nominal error rates at lower levels of correlation ($\rho \leq 0.6$), however at higher levels of correlation Bonferroni was conservative. Without correction, the FWER declined as the correlation increased but was still approximately 0.08 at $\rho = 0.8$. Only the Romano-Wolf and Holm methods ensured nominal family wise coverage at any level of correlation. Figure 2 shows the 95% confidence interval width for the four methods for the same model. For the two methods with nominal or near nominal error rates (Romano-Wolf and Holm), Romano-Wolf was moderately more efficient with narrower confidence intervals. The other methods displayed approximately constant confidence interval widths, with their respective widths reflecting the coverage results.

TABLE 1 Results of simulation experiments with two outcomes, seven clusters per arm, and with 10 000 iterations each.

					CI width	
Method	**Test statistic**	δ	**FWER**	**Coverage**	δ_1	δ_2
None (naive)	-	(0,0)	0.158	0.844	0.653	0.497
None (permutation)	Unweighted		0.099	0.902	0.725	0.605
	Weighted		0.102	0.903	0.726	0.599
Bonferroni	Unweighted		**0.048**	0.958	0.881	0.746
	Weighted		**0.051**	**0.954**	0.900	0.761
Holm	Unweighted		**0.051**	**0.947**	0.855	0.716
	Weighted		0.056	0.942	0.869	0.710
Romano-Wolf	Unweighted		**0.053**	**0.948**	0.841	0.708
	Weighted		**0.049**	**0.947**	0.824	0.720
None (naive)	-	(0,0.5)	0.067	0.845	0.654	0.484
None (permutation)	Unweighted		**0.048**	0.914	0.722	0.637
	Weighted		**0.049**	0.916	0.727	0.650
Bonferroni	Unweighted		0.026	0.960	0.885	0.789
	Weighted		0.022	0.962	0.902	0.831
Holm	Unweighted		**0.049**	**0.954**	0.851	0.754
	Weighted		0.045	**0.953**	0.870	0.773
Romano-Wolf	Unweighted		**0.048**	0.957	0.839	0.740
	Weighted		**0.051**	**0.947**	0.819	0.796

Note: Each iteration used 1000 permutations for the permutation test and 2000 iterations in each of the lower and upper confidence interval search processes. Bold results for FWER and coverage show those within approximated 95% Monte Carlo confidence interval of the nominal value.

Table 1 reports the results from model (2). Under all tested conditions the FWER was approximately nominal in all scenarios for all multiple testing corrections when both parameters were zero. However, when only one parameter was zero, Bonferroni was conservative as expected with a FWER ≈ 0.025 at $\alpha = 0.05$ which was also reflected in coverage being greater than the nominal rate. Romano-Wolf and Holm had nominal rates in all scenarios. Confidence interval width followed the same pattern as model (1) with Romano-Wolf generally being more efficient. Use of the weighted test statistic did not make much difference qualitatively with some confidence intervals larger and some smaller. Without correction, using a permutation test approach resulted in a FWER of ≈ 0.10 when there were two true null hypotheses, as expected. Using the naive output of `lme4` resulted in even worse performance due to the small sample bias in the test statistics, also as expected,[16,17] with FWER around 30%-50% higher. Table 2 reports the results from the two outcome trial simulations with a larger 20 clusters per arm. The same pattern is observed as the smaller two-arm experiments, but the small sample bias using the naive method is reduced. To illustrate the computational efficiency of the procedure, a single run of the function to derive *P*-values and confidence sets took between 1 and 10 s depending on the number of outcomes and size and number of the clusters.

Table 3 shows the results from the three outcome simulations with baseline measures. Despite the more complex covariance structure and imbalance in the number of observations between control and treatment conditions, Holm and Romano-Wolf maintained nominal FWER and coverage. Again, Romano-Wolf was the most efficient correction. Its confidence intervals were between 10% and 50% larger than the uncorrected results. We also note that the uncorrected approach maintain marginally nominal rates for each univariate outcome in all scenarios.

4 | APPLIED EXAMPLE

To provide a real-world example of the the use of the methods proposed in this article, we re-analyse a cluster randomized trial of a financial incentive to improve workplace health and wellbeing in small and medium sized enterprises (SME) in

TABLE 2 Results of simulation experiments with two outcomes, 14 clusters per arm, 10 individuals per cluster, and with 10 000 iterations each.

					CI width	
Method	**Test statistic**	δ	**FWER**	**Coverage**	δ_1	δ_2
None (naive)	-	(0,0)	0.125	0.878	0.573	0.416
None (permutation)	Unweighted		0.095	0.908	0.597	0.457
	Weighted		0.099	0.901	0.600	0.466
Bonferroni	Unweighted		**0.053**	**0.953**	0.706	0.543
	Weighted		0.046	0.959	0.714	0.566
Holm	Unweighted		**0.053**	**0.948**	0.693	0.529
	Weighted		**0.048**	**0.950**	0.700	0.544
Romano-Wolf	Unweighted		**0.052**	**0.951**	0.686	0.527
	Weighted		**0.047**	**0.952**	0.687	0.539
None (naive)	-	(0,0.5)	0.057	0.878	0.871	0.399
None (permutation)	Unweighted		**0.049**	0.915	0.597	0.469
	Weighted		**0.052**	0.927	0.600	0.561
Bonferroni	Unweighted		0.023	0.961	0.707	0.558
	Weighted		0.026	0.969	0.715	0.690
Holm	Unweighted		**0.053**	0.935	0.693	0.546
	Weighted		**0.049**	0.963	0.698	0.660
Romano-Wolf	Unweighted		**0.049**	0.956	0.685	0.542
	Weighted		**0.048**	**0.960**	0.679	0.655

Note: Each iteration used 1000 permutations for the permutation test and 2000 iterations in each of the lower and upper confidence interval search processes. Bold results for FWER and coverage show those within approximated 95% Monte Carlo confidence interval of the nominal value.

TABLE 3 Results of simulation experiments for the parallel cluster trial with baseline measures (3).

					CI width		
Method	**Test statistic**	δ	**FWER**	**Coverage**	δ_1	δ_2	δ_3
None (permutation)	Unweighted	(0,0,0)	0.129	0.876	0.872	1.199	0.700
	Weighted		0.148	0.848	1.033	1.201	0.767
Bonferroni	Unweighted		**0.046**	**0.954**	1.332	1.714	1.050
	Weighted		**0.047**	0.962	1.834	1.709	1.141
Holm	Unweighted		**0.048**	**0.946**	1.202	1.652	1.092
	Weighted		**0.048**	0.942	1.554	1.612	1.101
Romano-Wolf	Unweighted		**0.052**	0.956	1.014	1.633	0.986
	Weighted		**0.049**	**0.954**	1.611	1.498	0.997
None (permutation)	Unweighted	(0,0.5,0)	0.093	0.869	0.783	1.271	0.705
	Weighted		0.093	0.852	0.996	1.282	0.724
Bonferroni	Unweighted		0.038	0.940	1.303	1.895	1.043
	Weighted		0.032	0.942	1.824	1.834	1.134
Holm	Unweighted		0.055	0.939	1.234	1.730	1.090
	Weighted		0.045	0.949	1.593	1.646	1.105
Romano-Wolf	Unweighted		**0.048**	**0.953**	1.000	1.820	1.008
	Weighted		**0.049**	**0.954**	1.678	1.548	1.031

Note: Each iteration used 1000 permutations for the permutation test and 2000 iterations in each of the lower and upper confidence interval search processes. Bold results for FWER and coverage show those within approximated 95% Monte Carlo confidence interval of the nominal value.

the United Kingdom. The original trial was relatively complex and included four trial arms with pre- and post-intervention observations comprising a standard control condition (no incentive), two treatment conditions (high and low incentive), and a second control arm with no baseline measures also with no incentive. The trial enrolled 152 clusters (SMEs), which were randomly allocated in an equal ratio to each of the trial arms; 100 SMEs completed the trial. Up to 15 employees were sampled and interviewed from each cluster. The full protocol is published elsewhere[8] (at the time of writing the results from the trial are under review).

4.1 | 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.

4.2 | 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 modeled using a Bernoulli-logistic regression model, following the notation above, with $t = 0$ for baseline and $t = 1$ for post-intervention:

$$Y_{k,ict} \sim \text{Bernoulli}(\text{logit}(\mu_{0,k} + \delta_k D_{ct} + \theta_{k,c} + \theta_{k,ct})). \tag{25}$$

We used 4000 permutation test iterations and 10 000 steps in the confidence interval search procedure. For illustration, this re-analysis took 8 min on a desktop PC with Intel Core i7-9700K with 16GB RAM and Windows 10.

4.3 | Results

Table 4 shows the results of an analysis using the naive method (a model-based analysis using `lme4` with no multiple testing correction), alongside "corrected" results using the Bonferroni, Holm, and Romano-Wolf methods. We first note that the convergence of the confidence interval search procedure was highly sensitive to the starting values and we could not achieve reliable convergence for the weighted Bonferroni and Holm approaches. The algorithm could take a long time to find the right part of the parameter space, particularly since the search distance decays with the number of iterations. Convergence can be assessed graphically. Figure 3 shows a search procedure that has not converged after 10 000 iterations. Several of the upper bounds make a large jump and then require a large number of iterations to eventually reach to the bound. Figure 4 shows a convergent search procedure.

We make several observations about the results. 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

TABLE 4 Results from re-analysis of the workplace wellbeing trial.

Outcome	Statistic	None (Naive)	Bonferroni	Holm	R-W
Employer provided information	Estimate	2.91			
	95% CI (Unweighted)	[1.98, 3.97]	[0.14, 2.96]	[0.13, 2.94]	[0.34, 2.96]
	P-value (Unweighted)	0.03	0.03	0.01	0.01
	95% CI (Weighted)		NR*	NR*	[0.28, 2.96]
	P-value (Weighted)		0.02	0.02	<0.01
Employer provided	Estimate	2.11			
	95% CI (Unweighted)	[1.31, 2.99]	[−0.29, 3.04]	[−0.22, 2.71]	[−0.11, 3.22]
	P-value (Unweighted)	<0.01	0.21	0.15	0.05
	95% CI (Weighted)		NR*	NR*	[−0.16, 2.95]
	P-value (Weighted)		0.21	0.13	0.04
Employee made a conscious effort	Estimate	0.22			
	95% CI (Unweighted)	[−0.33, 0.77]	[−0.89, 0.98]	[−0.72, 1.32]	[−0.77, 1.45]
	P-value (Unweighted)	0.44	1.00	0.38	0.37
	95% CI (Weighted)		NR*	NR*	[−0.84, 1.45]
	P-value (Weighted)		1.00	0.38	0.36
Employee took part at	Estimate	1.13			
	95% CI (Unweighted)	[0.50, 1.75]	[−0.37, 1.72]	[−0.55, 1.74]	[−0.39, 1.85]
	P-value (Unweighted)	<0.01	1.00	0.88	0.27
	95% CI (Weighted)		NR*	NR*	[−0.43, 1.90]
	P-value (Weighted)		1.00	0.84	0.29
Employee took part outside work	Estimate	0.27			
	95% CI (Unweighted)	[−0.06, 0.61]	[−0.09, 0.95]	[−0.06, 0.97]	[−0.69, 0.83]
	P-value (Unweighted)	0.11	0.34	0.17	0.18
	95% CI (Weighted)		NR*	NR*	[−0.70, 0.83]
	P-value (Weighted)		0.34	0.16	0.17

Note: Results are log odds-ratios, 95% confidence intervals, and *P*-values. Permutation test *P*-values used 4000 iterations, and the confidence interval search procedure used 10 000 steps for Bonferroni, Holm, and Romano-Wolf (RW) methods. The "None (Naive)" method refers to a model-based analysis using `lme4` with no multiple testing correction. * NR (not reported) because a reliable convergence could not be achieved.

since it would seem contradictory for employees to make more effort but not report making more effort. The results corrected for multiple testing using Romano-Wolf appear to be more consistent in that employers appeared to make more effort but the employees did not take up the new services with small and negative effects now shown to be compatible with the data for the latter three outcomes. The effect of the intervention is also more uncertain than suggested by the uncorrected confidence intervals. In particular, the confidence intervals under the corrected methods, which are based on exact permutation 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.

5 | DISCUSSION

We have proposed how one can estimate frequentist statistics for cluster randomized trials with multiple outcomes that control for the FWER and coverage of simultaneous confidence intervals. These methods also apply generally in any scenario where multiple tests from GLMMs are used. Where a correction for multiple testing is desired in a cluster trial

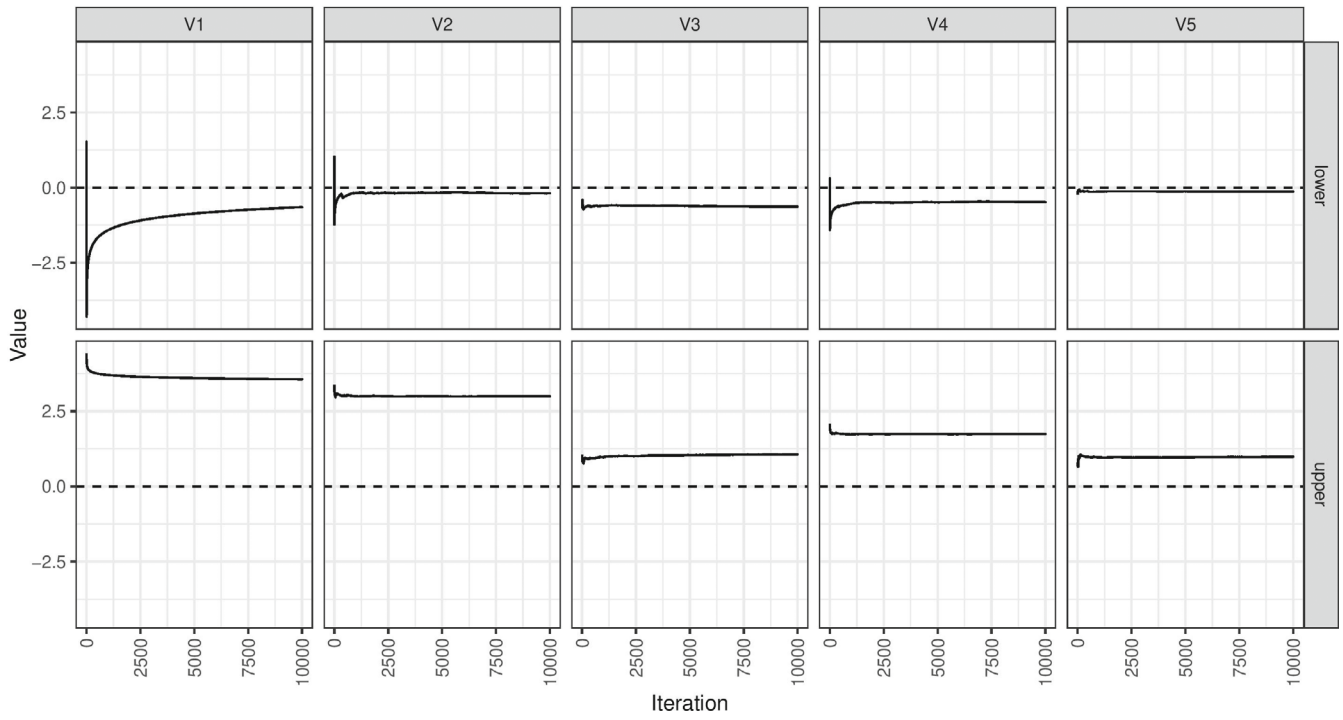

FIGURE 3 Example of the confidence interval search for the upper and lower confidence interval limits for the five outcomes in the applied example using the Holm correction for the cluster trial example demonstrating lack of convergence for some bounds.

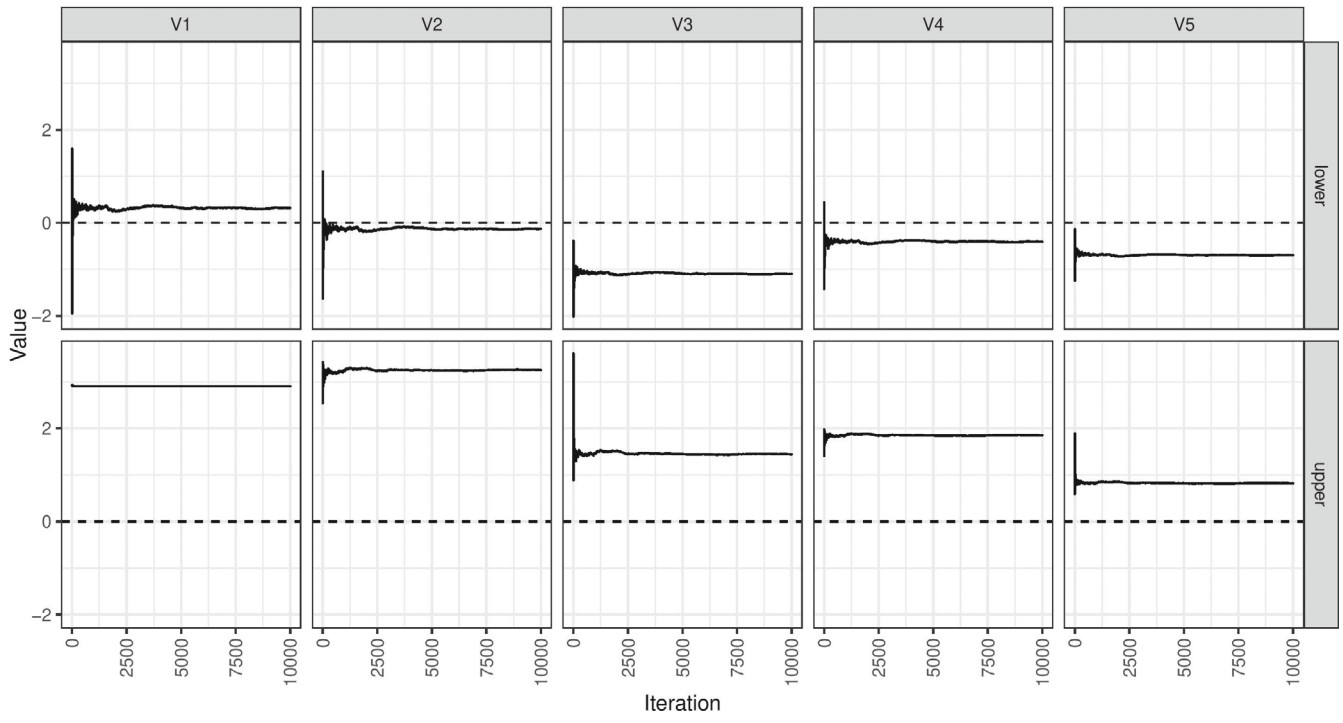

FIGURE 4 Example of the confidence interval search for the upper and lower confidence interval limits for the five outcomes in the applied example using the Romano-Wolf correction for the cluster trial example demonstrating convergence.

setting, the Romano-Wolf approach would be recommended as it maintains nominal rates in a variety of scenarios including with differing levels of between-outcome correlation, cluster and individual sample sizes, and covariance structures, it is also more efficient than the alternatives and in the example we considered, exhibits better convergence of confidence interval search procedures. Where a multiple testing correction is not desired, permutation-based methods are likely to provide marginally nominal error rates and so are also recommended when other methods may exhibit biases. We also compared a weighted test statistic based on the score statistic proposed by Romano and Wolf,[3] but did not find this provided any obvious benefit over an unweighted sum of generalized residuals. We do note, however, that while these methods do provide the desired properties, many regulatory agencies, including the FDA, do not (yet) accept statistics derived from re-sampling based methods, which may limit their application. Researchers may also consider other methods if multiple testing corrections are required such as "intersection-union" testing.[50]

There have been no previous comparisons of multiple testing corrections in the context of cluster randomized trials as far as we are aware, but our results generally reflect those from other settings. For example, Ozenne et al[28] compared several multiple testing corrections for linear latent variable models, including a resampling-based procedure, although not Romano-Wolf. They showed this method maintained strong control of the FWER and was more efficient than Bonferroni. Vickerstaff et al[27] considered the question for individual level randomized trials with a linear model, and suggested that Hommel's[24] and Hochberg's[23] methods were marginally more efficient than Bonferroni or Holm, but they did not include a permutation-based procedure, not non-linear models. Alberton et al[29] also shows permutation-based methods to outperform other corrections in the context of analysing brain imaging data.

We have examined methods from a range of previous work including: permutation tests for cluster trials,[32,51] univariate methods for corrections for multiple testing that use permutation tests,[3,15,22] and procedures for estimating confidence interval limits based on permutation tests.[44,45,48,49] Altogether the proposed methods can deal with several issues that are common to cluster randomized trials as they allow for multiple outcomes, they can incorporate other features such as restricted randomization methods, which are often used in trials with a small number of clusters. Watson et al,[16] Li et al,[18,37] and Zhou et al[35] discuss permutation tests with restricted randomization methods. Permutation-based methods provide 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,[16,17] however there is no obvious way these would be incorporated efficiently into a multiple testing procedure. After conducting the analyses presented in this article, an updated and more efficient version of the confidence interval search procedure was brought to our attention.[52] This method improves the efficiency of the search procedure, and requires fewer steps by making larger steps on average, although would not affect the results presented here. We aim to incorporate the algorithm in our R package implementing these methods (`crctStepdown`).

The tools developed for this article can be incorporated at the design stage of a cluster trial to determine power using simulation-based approaches. 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.[16,35] 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 permutation test could be conducted for the pairwise comparison that is invariant to randomized 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 randomized trials in many cases.

DATA AVAILABILITY STATEMENT

The data used in the applied example are available from the authors on request.

ORCID

Samuel I. Watson https://orcid.org/0000-0002-8972-769X
Karla Hemming https://orcid.org/0000-0002-2226-6550

REFERENCES

1. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC Med*. 2010;340:c332. doi:10.1136/bmj.c332
2. Wason JMS, Stecher L, Mander AP. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials*. 2014;15(1):364. doi:10.1186/1745-6215-15-364
3. Romano JP, Wolf M. Exact and approximate stepdown methods for multiple hypothesis testing. *J Am Stat Assoc*. 2005;100(469):94-108. doi:10.1198/016214504000000539
4. Lancet T. Consort 2010. *Lancet*. 2010;375(9721):1136. doi:10.1016/S0140-6736(10)60456-4
5. Murray DM. *Design and Analysis of Group Randomised Trials*. New York, NY: Oxford University Press Inc.; 1998.
6. Hemming K, Lilford R, Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *Stat Med*. 2015;34(2):181-196. doi:10.1002/sim.6325
7. Eldridge S, Kerry S. *A Practical Guide to Cluster Randomised Trials in Health Services Research*. Chichester, UK: John Wiley & Sons, Ltd; 2012.
8. 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*. 2019;19(1):493. doi:10.1186/s12889-019-6582-y
9. Dunbar EL, Wroe EB, Nhlema B, et al. 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*. 2018;8(7):e019473. doi:10.1136/bmjopen-2017-019473
10. US Food and Drug Administration. Multiple endpoints in clinical trials. Guidance for industry. Technical Report. Silver Spring, MD: US Food and Drug Administration; 2017.
11. Lafaye de Micheaux P, Liquet B, Marque S, Riou J. Power and sample size determination in clinical trials with multiple primary continuous correlated endpoints. *J Biopharm Stat*. 2014;24(2):378-397. doi:10.1080/10543406.2013.860156
12. Rubin M. When to adjust alpha during multiple testing: a consideration of disjunction, conjunction, and individual testing. *Synthese*. 2021;199:10969-11000. doi:10.1007/s11229-021-03276-4
13. Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. *J Res Educ Eff*. 2012;5(2):189-211. doi:10.1080/19345747.2011.618213
14. Farcomeni A. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat Methods Med Res*. 2008;17(4):347-388. doi:10.1177/0962280206079046
15. Romano JP, Wolf M. Stepwise multiple testing as formalized data snooping. *Econometrica*. 2005;73(4):1237-1282. doi:10.1111/j.1468-0262.2005.00615.x
16. Watson SI, Girling AJ, Hemming K. Design and analysis of three-arm parallel group randomised trials with small numbers of clusters. *Stat Med*. 2021;40(5):1133-1146.
17. Leyrat C, Morgan KE, Leurent B, Kahan BC. Cluster randomized trials with a small number of clusters: which analyses should be used? *Int J Epidemiol*. 2018;47(1):321-331. doi:10.1093/ije/dyx169
18. Li F, Turner EL, Heagerty PJ, Vollmer WM, Delong ER, Murray DM. An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes. *Stat Med*. 2017;36(24):3791-3806. doi:10.1002/sim.7410
19. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Stat Sci*. 2003;18(1):71-103. doi:10.1214/ss/1056397487
20. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 1979;6(2):65-70.
21. McNeish D, Stapleton LM. Modeling clustered data with very few clusters. *Multivariate Behav Res*. 2016;51(4):495-518. doi:10.1080/00273171.2016.1167008
22. Romano JP, Wolf M. Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Stat Probab Lett*. 2016;113:38-40. doi:10.1016/j.spl.2016.02.012
23. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988;75:800-802. doi:10.1093/biomet/75.4.800
24. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*. 1988;75:383-386. doi:10.1093/biomet/75.2.383
25. Holland B, DiPonzio CM. An improved sequentially rejective Bonferroni test procedure. *Biometrics*. 1987;43:417. doi:10.2307/2531823
26. Stevens JR, Masud AA, Suyundikov A. A comparison of multiple testing adjustment methods with block-correlation positively-dependent tests. *PLoS ONE*. 2017;12:e0176124. doi:10.1371/journal.pone.0176124
27. Vickerstaff V, Omar RZ, Ambler G. Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes. *BMC Med Res Methodol*. 2019;19:129. doi:10.1186/s12874-019-0754-4
28. Ozenne B, Budtz-Jørgensen E, Ebert SE. Controlling the familywise error rate when performing multiple comparisons in a linear latent variable model. *Comput Stat*. 2022;38:1-23. doi:10.1007/s00180-022-01214-7
29. Alberton BA, Nichols TE, Gamba HR, Winkler AM. Multiple testing correction over contrasts for brain imaging. *NeuroImage*. 2020;216:116760. doi:10.1016/j.neuroimage.2020.116760
30. Westfall PH, Young SS. *Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment*. 1st ed. Hoboken, NJ: Wiley; 1993.
31. Hooper R, Teerenstra S, de Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Stat Med*. 2016;35(26):4718-4728. doi:10.1002/sim.7028

32. Gail MH, Mark SD, Carroll RJ, Green SB, Pee D. On design considerations and randomization-based inference for community intervention trials. *Stat Med*. 1996;15(11):1069-1092. doi:10.1002/(SICI)1097-0258(19960615)15:11<1069::AID-SIM220>3.0.CO;2-Q
33. Thompson J, Davey C, Hayes R, Hargreaves J, Fielding K. Permutation tests for stepped-wedge cluster-randomized trials. *Stata J Promot Commun Stat Stata*. 2019;19:803-819. doi:10.1177/1536867X19893624
34. Wang R, De Gruttola V. The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials. *Stat Med*. 2017;36:2831-2843. doi:10.1002/sim.7329
35. Zhou Y, Turner EL, Simmons RA, Li F. Constrained randomization and statistical inference for multi-arm parallel cluster randomized controlled trials. *Stat Med*. 2022;41:1862-1883. doi:10.1002/sim.9333
36. Murray DM, Hannan PJ, Pals SP, McCowen RG, Baker WL, Blitstein JL. A comparison of permutation and mixed-model regression methods for the analysis of simulated data in the context of a group-randomized trial. *Stat Med*. 2006;25:375-388. doi:10.1002/sim.2233
37. Li F, Lokhnygina Y, Murray DM, Heagerty PJ, DeLong ER. An evaluation of constrained randomization for the design and analysis of group-randomized trials. *Stat Med*. 2016;35:1565-1579. doi:10.1002/sim.6813
38. Li F, Lu W, Wang Y, et al. A comparison of analytical strategies for cluster randomized trials with survival outcomes in the presence of competing risks. *Stat Methods Med Res*. 2022;31:1224-1241. doi:10.1177/09622802221085080
39. Blaha O, Esserman D, Li F. Design and analysis of cluster randomized trials with time-to-event outcomes under the additive hazards mixed model. *Stat Med*. 2022;41:4860-4885. doi:10.1002/sim.9541
40. Braun TM, Feng Z. Optimal permutation tests for the analysis of group randomized trials. *J Am Stat Assoc*. 2001;96(456):1424-1432. doi:10.1198/016214501753382336
41. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73:13-22. doi:10.1093/biomet/73.1.13
42. Guilbaud O. Simultaneous confidence regions corresponding to Holm's step-down procedure and other closed-testing procedures. *Biom J*. 2008;50(5):678-692. doi:10.1002/bimj.200710449
43. Hayter AJ, Hsu JC. On the relationship between stepwise decision procedures and confidence sets. *J Am Stat Assoc*. 1994;89(425):128-136. doi:10.1080/01621459.1994.10476453
44. Garthwaite P, Buckland S. Generating Monte Carlo confidence intervals by the Robbins-Monro process. *J R Stat Soc Ser C Appl Stat*. 1992;41(1):159-171.
45. Garthwaite PH. Confidence intervals from randomization tests. *Biometrics*. 1996;52(4):1387. doi:10.2307/2532852
46. Robbins H, Monro S. A stochastic approximation method. *Ann Math Stat*. 1951;22(3):400-407. doi:10.1214/aoms/1177729586
47. Ruppert D. A Newton-Raphson version of the multivariate Robbins-Monro procedure. *Ann Stat*. 1985;13(1):234-245.
48. Rabideau DJ, Wang R. Randomization-based confidence intervals for cluster randomized trials. *Biostatistics*. 2021;22:913-927. doi:10.1093/biostatistics/kxaa007
49. Rabideau DJ, Wang R. Randomization-based inference for a marginal treatment effect in stepped wedge cluster randomized trials. *Stat Med*. 2021;40:4442-4456. doi:10.1002/sim.9040
50. Yang S, Moerbeek M, Taljaard M, Li F. Power analysis for cluster randomized trials with continuous coprimary endpoints. *Biometrics*. 2022. doi:10.1111/biom.13692
51. Gallis JA, Li F, Yu H, Turner EL. cvcrand and cptest: commands for efficient design and analysis of cluster randomized trials using constrained randomization and permutation tests. *Stata J*. 2018;18:357-378. doi:10.1177/1536867x1801800204
52. Garthwaite PH, Jones MC. A stochastic approximation method and its application to confidence intervals. *J Comput Graph Stat*. 2009;18:184-200. doi:10.1198/jcgs.2009.0011

How to cite this article: Watson SI, Akinyemi JO, Hemming K. Permutation-based multiple testing corrections for *P*-values and confidence intervals for cluster randomized trials. *Statistics in Medicine*. 2023;1-18. doi: 10.1002/sim.9831