Title

A hybrid approach to sample size re-estimation in cluster randomized trials with continuous outcomes

Journal

Statistics in Medicine, 43(24)

ISSN

0277-6715

Authors

Sarkodie, Samuel K
Wason, James MS
Grayling, Michael J

Publication Date

2024-10-30

DOI

10.1002/sim.10205

Peer reviewed

Received: 3 October 2023 | Revised: 1 August 2024 | Accepted: 7 August 2024

DOI: 10.1002/sim.10205

RESEARCH ARTICLE

Statistics in Medicine WILEY

A hybrid approach to sample size re-estimation in cluster randomized trials with continuous outcomes

Samuel K Sarkodie[1] | **James MS Wason**[1] | **Michael J Grayling**[2]

[1]Population Health Sciences Institute, Newcastle University, Newcastle Upon Tyne, United Kingdom

[2]Statistics and Decision Sciences, Janssen R&D, High Wycombe, United Kingdom

Correspondence
Samuel K Sarkodie, Population Health Sciences Institute, Newcastle University, Newcastle Upon Tyne, United Kingdom.
Email: s.k.sarkodie2@newcastle.ac.uk

Funding information
National Institute for Health and Care Research (NIHR), Grant/Award Number: NIHR301614

This study presents a hybrid (Bayesian-frequentist) approach to sample size re-estimation (SSRE) for cluster randomised trials with continuous outcome data, allowing for uncertainty in the intra-cluster correlation (ICC). In the hybrid framework, pre-trial knowledge about the ICC is captured by placing a Truncated Normal prior on it, which is then updated at an interim analysis using the study data, and used in expected power control. On average, both the hybrid and frequentist approaches mitigate against the implications of misspecifying the ICC at the trial's design stage. In addition, both frameworks lead to SSRE designs with approximate control of the type I error-rate at the desired level. It is clearly demonstrated how the hybrid approach is able to reduce the high variability in the re-estimated sample size observed within the frequentist framework, based on the informativeness of the prior. However, misspecification of a highly informative prior can cause significant power loss. In conclusion, a hybrid approach could offer advantages to cluster randomised trials using SSRE. Specifically, when there is available data or expert opinion to help guide the choice of prior for the ICC, the hybrid approach can reduce the variance of the re-estimated required sample size compared to a frequentist approach. As SSRE is unlikely to be employed when there is substantial amounts of such data available (ie, when a constructed prior is highly informative), the greatest utility of a hybrid approach to SSRE likely lies when there is low-quality evidence available to guide the choice of prior.

KEYWORDS
adaptive design, Bayesian-frequentist, expected power, hybrid design, interim analysis, internal pilot, intra-class correlation

1 | INTRODUCTION

Estimation of the required sample size is one of a trial's most important, but challenging, aspects. This is because sample size estimation typically depends on foreknowledge of nuisance parameters that are difficult to stipulate at the design stage.[1] In cluster randomised trials (CRTs), where the unit of randomisation is a cluster (or group, eg, hospital, school), one such nuisance parameter is the intra-cluster correlation (ICC), which measures the correlation between outcomes from different individuals within a cluster. Several authors have discussed the difficulty in obtaining reliable estimates of the ICC,[2,3] and the implications of its misspecification on the statistical power of a trial.[4,5]

To mitigate against misspecification of the ICC, previous studies have established the usefulness of allowing for uncertainty in the ICC within the sample size calculation.[6,7] Specifically, some authors have encouraged the application of confidence interval techniques,[2,8] Bayesian methods,[9,10] and hybrid[11] (Bayesian-frequentist) approaches to account for the uncertainty in the ICC. In the hybrid framework, a prior is placed on the ICC with quantities such as statistical assurance,[12,13] probability of success (PoS), or expected power (EP)[14] then controlled in the sample size calculation in place of the conventional frequentist power. A limitation of approaches of this kind is that their utility can be highly dependent on the choice of prior, with the possibility always present that it may poorly reflect the data from the intended trial.

In theory, this limitation could be addressed by sample size re-estimation (SSRE), an adaptive design that uses study data gathered during the trial to re-estimate nuisance parameter(s) at an interim analysis.[15] Since a portion of the trial data is used to compute these parameter estimates at the interim analysis, they may be expected to provide a better estimate of their actual values than any pre-trial 'guess', even guesses that account for uncertainty such as those represented through a prior in a hybrid approach.[16]

Previous works have investigated the potential of SSRE within the context of CRTs in the parallel-group[16,17] (PG) and stepped-wedge domains.[15] Methods proposed in these publications re-estimated the total outcome variance, the ICC, the target effect powered for, or some subset of these parameters. The investigations were frequentist in nature, meaning the re-estimation used interim point estimates of the nuisance parameters to update the required sample size. Though the methods performed well on average, the variability in their re-estimated required sample size could undermine their utility in practice.[18] This is a consequence of challenges with the precision of estimation in CRTs, particularly in relation to estimating the ICC, which can be difficult to estimate even on the completion of a large trial. That is, existing frequentist methods for SSRE in CRTs neglect any uncertainty in the interim point estimates of the nuisance parameters. Sarkodie et al[11] recently demonstrated that the hybrid approach may provide utility at the design stage of a CRT. However, this proposed approach provides no flexibility to evaluate the selected prior at the design stage based on data from an interim analysis for potential adjustment in the final sample size. Thus, a natural question of interest is whether a hybrid approach could also be useful for SSRE.

Consequently, in this paper, we develop a hybrid approach to SSRE for PG-CRTs. This is achieved by assuming a prior for the ICC at the design stage of the trial. This prior is then updated at an interim analysis, to a posterior, based on available data. The posterior is then used in determining the re-estimated required sample size to control the EP to a desired level, following the approach described for pre-trial EP control described by Sarkodie et al.[11] Following accrual of the re-estimated sample size, the final analysis uses all available data in a conventional frequentist analysis to determine whether the null hypothesis can be rejected. Such a hybrid approach to SSRE seems intuitively appealing, as it may directly account for uncertainty in the ICC through both a pre-trial prior and also by considering the variability of the interim ICC estimate. To ascertain whether this approach is useful in practice, we explore both blinded and unblinded methods of SSRE, and perform a comparison between the existing frequentist and our proposed hybrid approach.

2 | METHODS

First, we describe how SSRE can be performed in both the frequentist and hybrid frameworks. For brevity, the methodology focuses on interim updating of the required number of clusters throughout, as increasing the number of clusters typically has a bigger impact on power than increasing the cluster size.[7] However, we provide an example of updating the required cluster size in the Supplementary Material, demonstrating that the underlying methodology for updating the cluster size is similar to that which follows for re-estimating the required number of clusters.

2.1 | Setting and notation

We consider the case of a PG-CRT where clusters are randomised to receive an experimental or a control treatment. We assume the primary outcome is continuous and normally distributed with variance σ^2. Accordingly, let Y_{ij} be the outcome from patient $i = 1, \dots, n_j$ in cluster $j = 1, \dots, C$. While we acknowledge that sample sizes can vary between clusters, we restrict our attention to having the same number of participants per cluster, and thus assume $n_j = n$ for all j. Furthermore, for simplicity, we assume equal allocation of clusters to the experimental and control treatments. Given the (likely) non-independence in the data, one possible approach is to fit a linear mixed model to the data at both interim and final analyses. At the final analysis, and at the interim analysis in unblinded SSRE procedures, the model is assumed to be:

$$Y_{ij} = \theta + X_j\mu + c_j + e_{ij}. \tag{1}$$

Here, θ is an intercept term (the mean in the control arm), $X_j = 1$ if cluster j is allocated to the experimental arm and $X_j = 0$ otherwise, $c_j \sim N(0, \sigma_c^2)$ is a random effect for cluster j, and $e_{ij} \sim N(0, \sigma_e^2)$ is the individual-level error. Note that $\sigma^2 = \sigma_c^2 + \sigma_e^2$, and that the ICC $\rho = \sigma_c^2/\sigma^2$. Then, μ is the treatment effect of interest and we specify our one-sided null hypothesis as $H_0 : \mu \leq 0$. The test statistic for H_0 is:

$$t = \frac{\hat{\mu}}{\sqrt{var(\hat{\mu})}},$$

which can be computed using, for example, REML estimation. The degrees of freedom for the test statistic will always be assumed to be that in a corresponding balanced ANOVA analysis. That is, $df = nC - C - 1$. Therefore, for a target type I error of α, H_0 will be rejected if t is greater than $t_{1-\alpha,df}$, the $(1-\alpha)$-quantile of a central t distribution on df degrees of freedom.

At the interim analysis in blinded SSRE procedures, where cluster assignment is unknown, the model fitted is instead:

$$Y_{ij} = \theta + c_j + e_{ij}.$$

A blinded procedure in this context implies that the treatment status of an observation is undisclosed, but there is awareness of which observations belong to the same cluster.[15]

In either case, the interim analysis results in estimates $\hat{\sigma}^2_{c,int}$ and $\hat{\sigma}^2_{e,int}$ which can be combined into an interim estimate of the ICC $\hat{\rho}_{int} = \hat{\sigma}^2_{c,int}/\left(\hat{\sigma}^2_{c,int} + \hat{\sigma}^2_{e,int}\right)$.

2.2 | Sample size re-estimation procedure

A high-level summary of how SSRE functions in both frequentist and hybrid frameworks is as follows.

First, a sample size is chosen for when the interim analysis will occur. As we assume n is fixed, this corresponds to selecting a certain number of clusters, C_{int}, from which data will have been collected at the interim analysis. This could be achieved by utilising some proportion of an initially calculated sample size based on assumed values for required parameters. Alternatively, a pragmatic sample size could be selected, for example, based on the number of clusters required to make a sufficiently precise estimate of the ICC.

The trial is then conducted until the interim required sample size is achieved, and the ICC estimated (ie, $\hat{\rho}_{int}$ is computed). Given the value of $\hat{\rho}_{int}$ (and using other selected design parameters, eg, the target type I error rate), the required sample size is re-estimated. That is, a value for the final target number of clusters, C_{reest} is computed. It is the interim estimation of ρ (blinded or unblinded) and the method of utilising $\hat{\rho}_{int}$ to compute C_{reest} (frequentist or hybrid) that will differ between the compared methods.

Next, if $C_{reest} \leq C_{int}$, the study terminates and the final analysis is conducted. Otherwise, the trial continues until data from C_{reest} clusters has been accrued, with the final analysis then conducted using data from both stages. This final analysis is conducted using the approach outlined above (ie, without adjustment for the inclusion of the interim analysis); thus consideration of the potential for type I error inflation will be important.

2.3 | Sample size re-estimation in the frequentist framework

The classical method of sample size estimation for a PG-CRT in the frequentist framework is to first calculate the sample size required for a corresponding individually randomised trial (IRT), and then multiply it by a ‘design effect’ (or ‘variance inflation factor’) to account for clustering. The sample size for the IRT (N_{IRT}), assuming power of $1-\beta$ is desired when

$\mu = \delta > 0$, is obtained as:

$$N_{IRT} = \frac{4(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{\delta^2}.$$

Then, the design effect for the considered type of PG-CRT is given by:

$$DE(\rho) = 1 + (n-1)\rho.$$

Hence, if there are n measurements per cluster, the required number of clusters is, for a particular value of ρ:

$$C(\rho) = N_{IRT} DE(\rho)/n. \tag{2}$$

In SSRE within the frequentist framework, the interim estimated ICC ($\hat{\rho}_{int}$) is simply inserted into Equation (2). That is, the method sets $C_{reest} = C(\hat{\rho}_{int})$.

2.4 | Sample size re-estimation in the hybrid framework

Sample size calculation in the hybrid framework amounts to averaging the frequentist power over any uncertainty in nuisance parameters by placing priors on these parameters. In this framework, two quantities are commonly used for sample size determination: the EP and the PoS. In this work, where a prior is placed only on the ICC, the standard definitions of the PoS and EP become equivalent[11] and can be expressed as:

$$EP(\psi, \mu, C) = \int_0^1 P(\mu, n, C, \alpha, \sigma, \rho)\psi(\rho|\theta)\, \mathrm{d}\rho,$$

where $P(\mu, n, C, \alpha, \sigma, \rho)$ is the probability of rejecting H_0 under a PG-CRT design, equal to

$$\Phi\left[\mu\sqrt{\frac{Cn}{4\{1+(n-1)\rho\}\sigma^2}} - \Phi^{-1}(1-\alpha)\right],$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution and $\psi(\rho|\theta)$ is the prior density for an ICC of ρ, which is dependent on parameters θ. Assuming EP of $1-\gamma$ is then desired when $\mu = \delta > 0$, sample size calculation is then performed by numerically searching for the minimal C such that $EP(\psi, \delta, C) \geq 1-\gamma$. In practice, γ is often set to be equal to the value of β from the frequentist framework.

Since the ICC (typically) ranges between 0 and 1, we select a prior distribution with support on [0,1]. Note that for simplicity and efficiency, a conjugate prior distribution may be desirable. However, conjugacy cannot be achieved in this study since the likelihood for the ICC is complex: this necessitates approximation using MCMC to make sampling from the non-conjugate posterior distribution possible. We achieve this through the *rjags* package in R. We select a Truncated Normal distribution, truncated on [0,1], as the prior for this study. We acknowledge that other distributions, such as the Beta distribution, could similarly be used. Nonetheless, an evaluation of alternative priors by Sarkodie et al[11] revealed no significant sensitivity to the exact choice of prior, given they held (approximately) the same mean and variance. Therefore, for simplicity, we consider only a Truncated Normal prior and denote the choice by $TN(0, 1, m, s)$, where m and s are the mean and standard deviation (SD) parameters of the Normal distribution before truncation.

For Normally distributed outcome data, Ukoumunne[8] proposed an approximation for the variance of an ICC estimate using Fisher's method.[19] Using this, it is assumed that:

$$\hat{\rho} \sim N\left[\rho, \frac{2(1-\rho)^2[(1+(n-1)\rho]^2}{n(n-1)C}\right].$$

Adopting this as the likelihood, the posterior for ρ can be determined at the interim analysis. Consequently, on calculating $\hat{\rho}_{int}$ at the interim analysis, the prior $TN(0, 1, m, s)$ can then be updated to a posterior that is a function of $\hat{\rho}_{int}, n, C_{int}, m$, and s, which we denote for brevity as $\hat{\psi}(\rho|m, s, \hat{\rho}_{int})$.

The posterior is then substituted into the EP to update the sample size. That is, the method sets C_{reest} as the minimal value such that $EP(\hat{\psi}, \delta, C_{reest}) \geq 1 - \gamma$.

2.5 Simulation study

The parameters for our simulation study are motivated by the study of Hankonen et al[20] which sought to reduce adolescent sedentary behaviour by improving physical activity. The interim estimate of the ICC for this study was $\hat{\rho}_{int} = 0.059$. An internal pilot of 25 clusters with an average cluster size of 17 was used to estimate the ICC at this interim analysis. Accordingly, we set $C_{int} = 26$ (to allow equal allocation of clusters to the control and experimental treatments) and $n = 17$. The study desired 80% power ($\beta = 0.2$) to detect a difference of $\delta = 0.3$ for an SD of $\sigma = 1.3$ and $\alpha = 0.025$.

To evaluate the performance of the SSRE techniques, we conducted a thorough simulation study. Specifically, we wanted to evaluate how varied values of the prior parameters m and s impacted the operating characteristics. We consider $m = 0.01, 0.059, 0.10$ and $s = 0.01, 0.1, 1.00$, to give a range of possible concordances of the prior densities in relation to the value of $\hat{\rho}_{int}$. These priors are shown in Figure 1.

We consider the performance of the hybrid approach for these m and s, alongside the performance of the frequentist method, for both blinded and unblinded SSRE, for a range of possible values of ρ. We also complete this work under the null ($\mu = 0$) and the alternative ($\mu = \delta$), in order to empirically estimate the type I error rates and power of the various SSRE procedures. For simplicity, we follow previous works in setting $\gamma = \beta$.

In all simulations, data is generated using Equation (1); the value of ρ, combined with the fixed assumption $\sigma = 1.3$, sets the assumptions required regarding σ_c and σ_e. Further, μ was set to either 0 or δ according to whether the interest was the type I error rate or power. Finally, θ was set to 0 without loss of generality and the X_j were arbitrarily set to imply balance in allocation between the arms.

For each combination of assumed parameters and particular SSRE approach, we then empirically compute several measures to assess performance, based on the results of 10,000 simulation runs. Firstly, the probability of rejecting H_0 is estimated. If the computed test statistic at the final analysis in simulation replicate i is t_i, and the final number of clusters is $C_{(reest,i)}$, this probability is:

$$\mathbb{P}(\text{Reject } H_0) = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} \mathbb{I}(t_i > t_{1-\alpha, df_i}),$$
$$df_i = nC_{(reest,i)} - C_{(reest,i)} - 1.$$

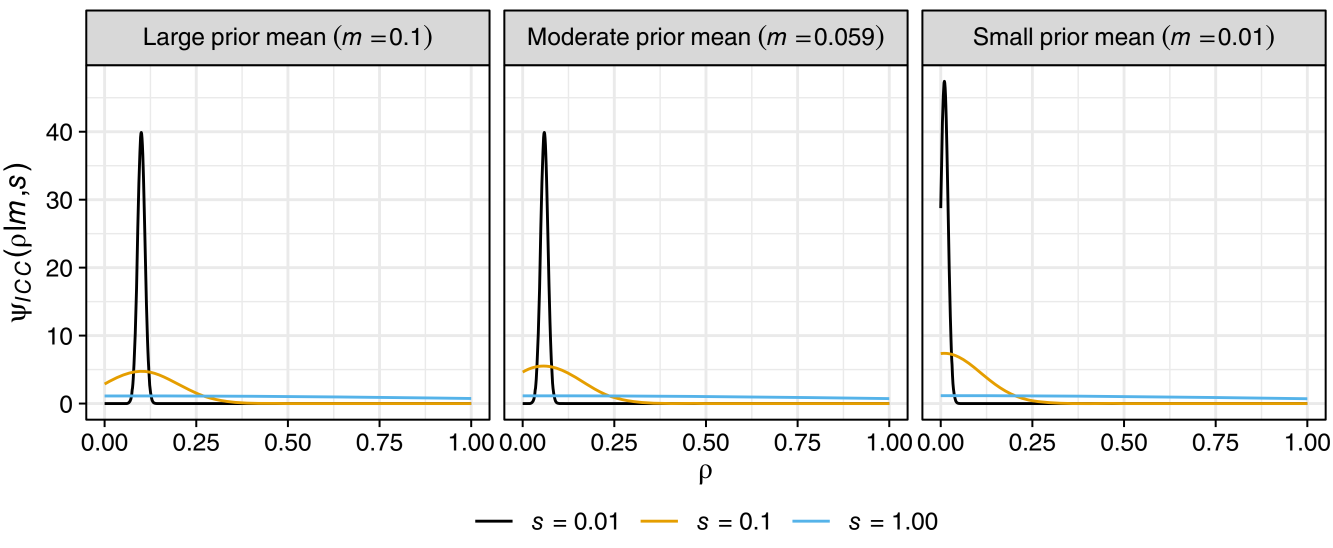

FIGURE 1 Plot of utilised truncated normal prior distributions. Plots are faceted by the use of $m = 0.01, 0.059, 0.10$ and all combinations of $s = 0.01, 0.10, 1.00$ are considered.

Our other two metrics relate to the ability of the SSRE procedures to specify the 'correct' required number of clusters reliably. That is, we think of the goal of each SSRE procedure as being to 'estimate' the sample size that would have been used if the true values of the design parameters were known. Thus we consider the re-estimated sample size as an estimator, with the target estimand being the 'oracle' sample size that would have been chosen in the frequentist framework if all parameters were known. Natural measures of the performance of these estimators are then its bias and mean square error (MSE); these are also computed. A SSRE that performs well will have a bias close to 0 and a low MSE. The bias and MSE are given by:

$$Bias = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} C_{(reest,i)} - C(\rho),$$

$$MSE = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} [C_{(reest,i)} - C(\rho)]^2.$$

Here, C(ρ) is the 'oracle' required number of clusters for the particular value of ρ assumed in the simulations that generated $C_{(reest,1)}, \dots , C_{(reest,10{,}000)}$.

Another important performance measure that is considered in this paper is the proportion of trials where the SSRE procedure underestimates (underpowered), overestimates (overpowered), or correctly estimates the required sample size. As noted above, if C(ρ) is the 'oracle' required number of clusters for a particular value of ρ, then we define a single trial as correctly powered if its sample size falls within $\pm 10\%$ of the oracle sample size, that is, if $0.9C(\rho) \leq C_{(reest,i)} \leq 1.1C(\rho)$. By contrast trial is considered underpowered if $C_{(reest,i)} < 0.9C(\rho)$ and overpowered if $C_{(reest,i)} > 1.1C(\rho)$. To estimate the proportion of cases where the trial is correctly estimated, underestimated, or overestimated, we use:

$$\mathbb{P}(\text{Correctly powered}) = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} \mathbb{I}\{0.9C(\rho) \leq C_{(reest,i)} \leq 1.1C(\rho)\},$$

$$\mathbb{P}(\text{Underpowered}) = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} \mathbb{I}\{C_{(reest,i)} < 0.9C(\rho)\},$$

$$\mathbb{P}(\text{Overpowered}) = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} \mathbb{I}\{C_{(reest,i)} > 1.1C(\rho)\}.$$

The code to reproduce the results in this paper is available from https://github.com/sks2023/article_codes.

3 | RESULTS

3.1 | Practical updating of the required number of clusters based on interim ICC estimate

To provide some intuition on how the choice of prior can influence the re-estimated required number of clusters in the hybrid framework, we present a plot of posterior modes in Figure 2, that is, modal values of $\psi(\rho|m, s, \hat{\rho}_{int})$, over $\rho \in [0, 1]$, are given as a function of m, s, and $\hat{\rho}_{int}$.

Figure 2 shows the interplay between $\hat{\rho}_{int}$, the posterior mode, and the re-estimated required number of clusters, given the prior mean and SD. Generally, as the interim estimate of the ICC increases, a monotonic relationship between the variables is observed where larger prior mean and SD values result in a larger posterior mode, consequently resulting in a larger number of required clusters. A caveat to this is that when $\hat{\rho}_{int}$ takes smaller values (approximately $\hat{\rho}_{int} < 0.13$), a highly informative prior ($s = 0.01$) can result in a larger required number of clusters.

Note that the impact of m on the posterior mode and the number of required clusters diminishes as s becomes large. That is, the final sample size is not heavily dependent on the prior mean when the prior is non-informative and vice versa. Put differently, an inaccurate prior mean will have less impact on the final sample size if the prior is non-informative. Note also that although the three curves for the 'weakly' informative prior ($s = 0.1$) are evidently distinct, they are not as widely separated as for the informative priors ($s = 0.01$).

According to the study by Hankonen et al, the estimated ICC at the interim analysis was 0.059. If we were to use this information in frequentist re-estimation we would need 68 clusters. However, within the hybrid framework, it is observed

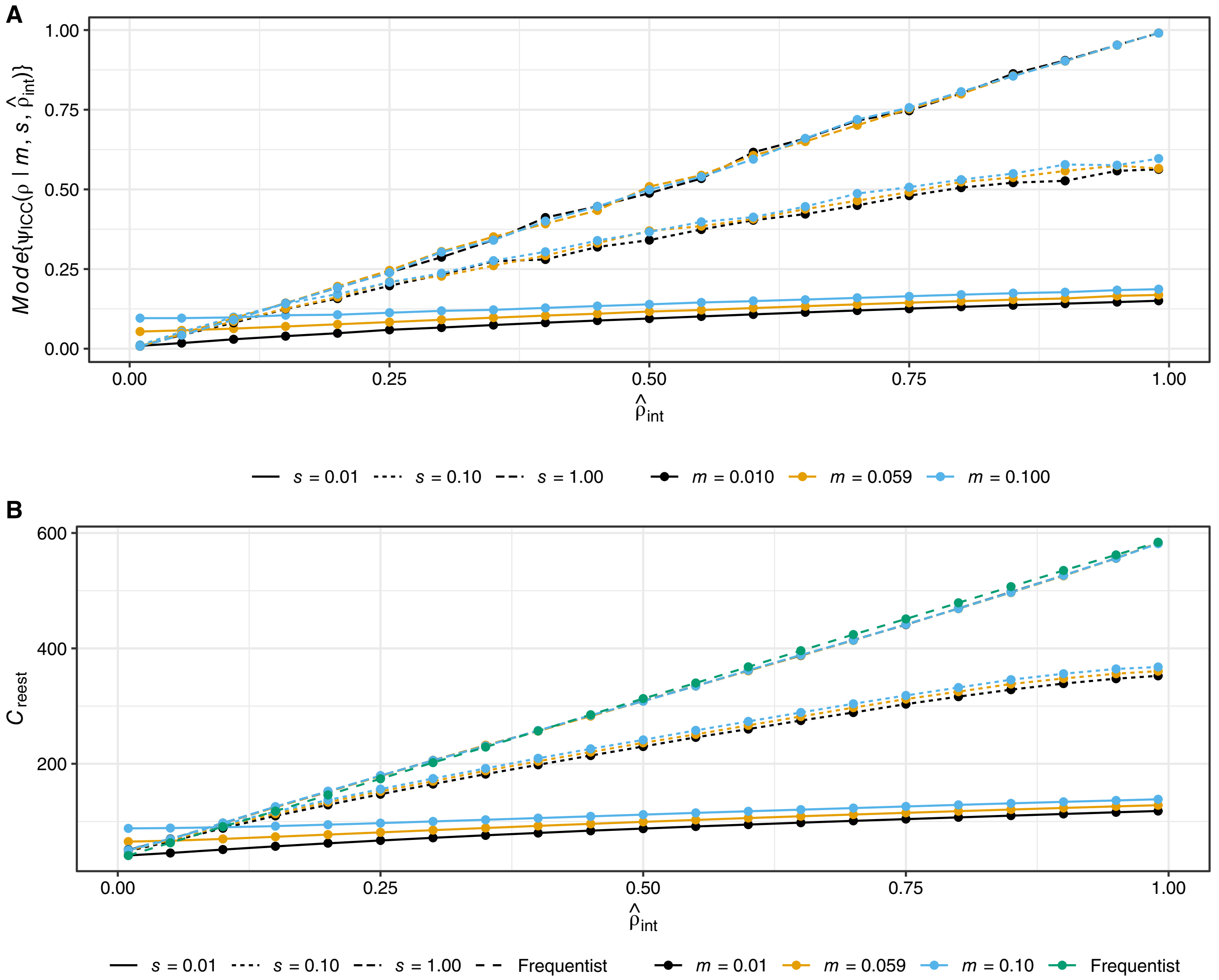

FIGURE 2 Plots of $\hat{\rho}_{int}$ against the posterior mode (A) and the re-estimated required number of clusters (B), given the prior mean (m) and SD (s) are shown, for all combinations of $m = 0.01, 0.059, 0.10$ and $s = 0.01, 0.10, 1.00$. The re-estimated required number of clusters for the frequentist design is also shown in B.

from Figure 2B that the required number of clusters ranges from 46 to 88, depending on the prior mean and standard deviation. We refer the reader to the Supplementary Material for a more focused plot in which this result is clearer.

Importantly, it is also observed from Figure 2B that, as would be expected, leveraging a highly informative prior results in a more constant re-estimated required number of clusters across $\hat{\rho}_{int}$. By contrast, the frequentist framework, or the use of a weakly or non-informative prior results in a clear relationship where the re-estimated required number of clusters can increase rapidly in $\hat{\rho}_{int}$.

Of course, the superiority of one method's re-estimated required number of clusters over another depends on the (unknown) true value of the ICC. For this reason, we now present the results of our simulation study evaluating the average performance of the various SSRE procedures.

3.2 | Re-estimated sample size, power, and type I error rates for correctly specified priors

In what follows, we evaluate the distribution of the re-estimated number of clusters, the power, and the type I error rate, for selected priors in the hybrid framework. Specifically, we assume that $\rho = 0.059$, and the priors are 'correctly specified' (ie, $m = \rho$). Then, we explore how on average, a highly informative prior, a weakly informative prior, and a non-informative prior impact the performance of the SSRE procedure in the hybrid framework. Our results are stratified by the use of blinded or unblinded SSRE, and are also compared to the performance of the frequentist approach. A summary of the performance measures for the SSRE procedures is given in Table 1, while the distribution of the re-estimated required number of clusters is presented in more detail in Figure 3. Recall that for the parameters from the motivating example, when $\rho = 0.059$, 68 clusters are required for a frequentist power of $\sim 80\%$.

On average, both the hybrid and frequentist approaches can mitigate against the implications of misspecifying the ICC at the trial's design stage, as indicated by their mean values of C_{reest}. Note that the interim estimate of the ICC is dependent only on the interim model (ie, blinded or unblinded) and the value of μ. In the blinded model, the average value of $\hat{\rho}_{int}$ is 0.0711 when $\mu = \delta$ and 0.0583 when $\mu = \delta$. In the unblinded model, $\hat{\rho}_{int}$ is always 0.0583 on average regardless of the value of μ. As the SSRE technique leverages the interim estimate of the ICC to make a final determination on the sample size, the likelihood of obtaining an accurate sample size is dependent on the closeness of the interim estimate to the truth. However, unlike the frequentist approach whose final sample size is dependent only on $\hat{\rho}_{int}$, the final sample size in the hybrid framework is a function of $\hat{\rho}_{int}$ and other parameters which include the prior SD. Thus, although the frequentist and hybrid approaches may compute the same interim ICC estimate, their final average re-estimated required number of cluster sample sizes may differ.

As a result of the patterns observed in Figure 2, the average re-estimated required number of clusters in the hybrid framework increases as s increases. For $s = 1.00$, as seen in Table 1, the difference in sample sizes between the hybrid and frequentist frameworks is relatively small, with this phenomenon expected based on previous studies.[11] Explicitly, a maximum increase of 10% in average sample size is observed between the frequentist approach and the hybrid approach with a non-informative prior. In this setting, this small increase in average sample size may be considered beneficial if it translates into power more reliably above the desired level. Of importance in the SSRE procedures is the control of the type I error rate, which appears similar in both frameworks with the small observed differences attributed to simulation error. We note that the results for the type I error rate were similar across all additional simulation scenarios and thus are not presented below, that is, in general, some small inflation was observed.

In both frameworks, the interim ICC estimates from the blinded model are biased when there is a non-zero treatment effect (ie, for $\mu = \delta$). Thus, the model overestimates the interim ICC on average and re-estimates a larger required sample size. Although regulatory agencies prefer blinded models,[1,21] this may be less necessary in CRTs as cluster allocations are not always blinded. We comment further on this in the Discussion.

Having established from Table 1 that there is no considerable difference in the average value of C_{reest} across the frameworks, we next examine the variability in this quantity. Particularly, our interest lies in whether the hybrid framework has less variability than the frequentist framework, which is known to have high variability in terms of the re-estimated required sample size. For the hybrid framework, we again evaluate how the selected prior SD impacts the performance. The findings are shown in Figure 3.

In comparison to the hybrid approach, the frequentist approach results in higher variability in the re-estimated required number of clusters. This is evidenced by the lower variance and lower interquartile ranges recorded in the hybrid framework compared to the frequentist. Furthermore, the variability is significantly lower for a highly informative prior,

TABLE 1 A summary of the performance of the cluster size re-estimation procedures is shown for the case where $m = \rho = 0.059$.

			Mean of C_{reest}			
Interim model	**Framework**	s	$\mu = 0$	$\mu = \delta$	**Power**	**Type I error rate**
Blinded	Frequentist	N/A	68	75	0.80	0.029
Blinded	Hybrid	0.01	68	69	0.80	0.026
Blinded	Hybrid	0.10	73	79	0.83	0.026
Blinded	Hybrid	1.00	75	82	0.84	0.026
Unblinded	Frequentist	N/A	68	68	0.80	0.033
Unblinded	Hybrid	0.01	68	68	0.80	0.027
Unblinded	Hybrid	0.10	73	73	0.83	0.031
Unblinded	Hybrid	1.00	75	75	0.84	0.030

Note: For all hybrid designs, the assume prior ψ is of the form $TN(0, 1, 0.059, s)$ for given s.

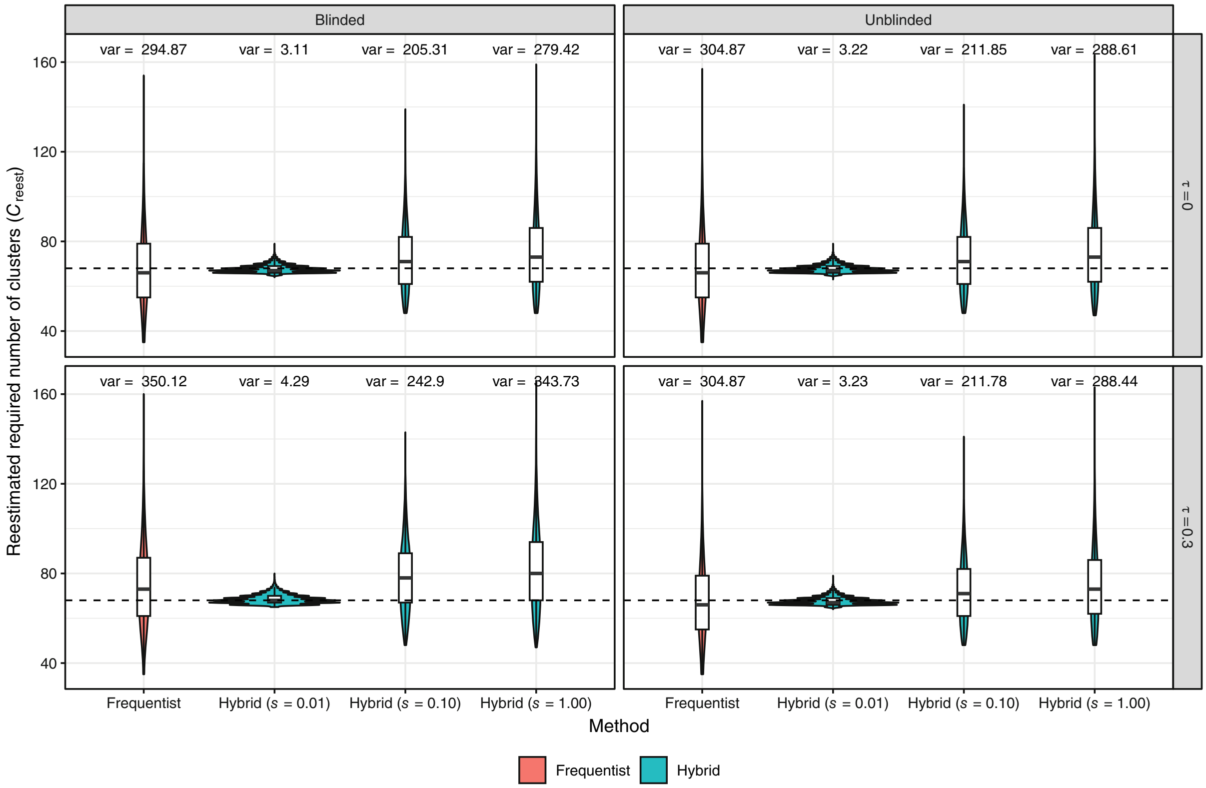

FIGURE 3 Violin and boxplots showing the variability in the re-estimated sample sizes (C_{reest}) for the frequentist and hybrid methods ($s = 0.01, 0.1, 1.00$), with the respective variances ($Var(C_{reest})$) and a horizontal dashed line indicating the oracle sample size is also displayed. Results are faceted by blinded versus unblinded sample size re-estimation and the value of the treatment effect. In all cases, $m = \rho = 0.059$ is assumed.

TABLE 2 Proportion of cases where the SSRE procedure underestimates (underpowered), overestimates (overpowered), or correctly estimates the required sample size.

Interim model	Framework	μ	Correct (%)	Underpowered (%)	Overpowered (%)
Blinded	Frequentist	0	28.3	39.0	32.7
Blinded	Frequentist	δ	27.9	39.4	32.7
Blinded	Hybrid ($s = 0.01$)	0	100.0	0.0	0.0
Blinded	Hybrid ($s = 0.10$)	0	37.3	23.1	39.6
Blinded	Hybrid ($s = 1.00$)	0	33.1	22.3	44.6
Blinded	Hybrid ($s = 0.01$)	δ	99.6	0.0	0.4
Blinded	Hybrid ($s = 0.10$)	δ	31.9	13.1	55.0
Blinded	Hybrid ($s = 1.00$)	δ	27.6	12.6	59.8
Unblinded	Frequentist	0	26.7	25.6	47.7
Unblinded	Frequentist	δ	27.9	39.4	32.7
Unblinded	Hybrid ($s = 0.01$)	0	99.9	0.0	0.1
Unblinded	Hybrid ($s = 0.10$)	0	36.6	23.6	39.7
Unblinded	Hybrid ($s = 1.00$)	0	32.4	22.8	44.8
Unblinded	Hybrid ($s = 0.01$)	δ	99.9	0.0	0.1
Unblinded	Hybrid ($s = 0.10$)	δ	36.7	23.6	39.7
Unblinded	Hybrid ($s = 1.00$)	δ	32.4	22.9	44.7

Note: Here, $m = \rho = 0.059$ is assumed.

increasing as the prior becomes non-informative. Despite the relative variability increase that results from a large prior SD, the variability in these scenarios is still lower than in the corresponding frequentist approach. The horizontal dashed line representing the oracle sample size illustrates that as the prior becomes less informative, the hybrid framework tends to more often overestimate the required sample size. This could be a possible utility of the hybrid framework, as most people favour trials with higher-than-required sample sizes over those with insufficient sample sizes, though this would depend on the degree of overestimation. We elaborate on the proportion of trials for which the sample sizes are correctly estimated, overestimated or underestimated in the next section.

3.3 | Proportion of correctly powered, underpowered, and overpowered trials for correctly specified priors

In this section, we evaluate the performance of the SSRE procedures based on the proportion of trials that were correctly powered, overpowered, or underpowered. Again, we focus on the case where $m = \rho$. The results presented in Table 2 indicate that in the frequentist framework, trials are often either underpowered or overpowered, with approximately

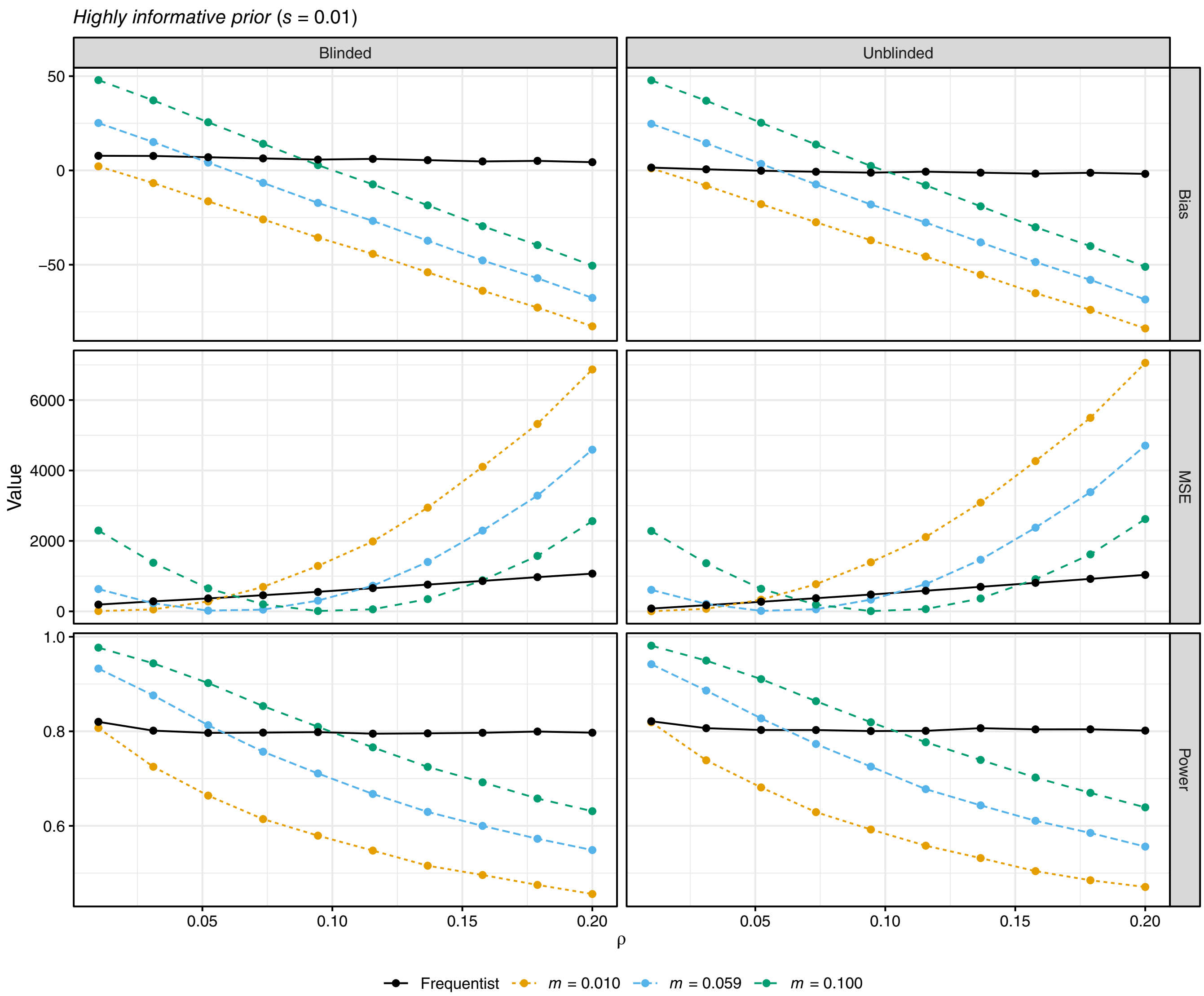

FIGURE 4 The bias, mean square error (MSE), and power of the frequentist and hybrid methods are shown as a function of the intra-cluster correlation (ρ). Results are faceted by the use of blinded versus unblinded sample size re-estimation. For the hybrid approach, all combinations of $m = 0.01, 0.059, 0.1$ and $s = 0.01$ are considered.

only 28% of trials correctly powered. In the hybrid framework, employing a highly informative prior unsurprisingly leads to correctly powered trials almost 100% of the time. However, for weakly and non-informative priors, trials are most often overpowered for the considered scenarios. Significantly, depending on the interim model, we observe around a 10% increase in the proportion of correctly powered trials when using the weakly informative prior compared to the frequentist approach. The proportion of correctly powered trials tends to be similar to that of the frequentist framework when using non-informative priors, although some marginal gains are noted.

3.4 | Impact of prior misspecification on SSRE performance

The results from Table 2 correspond to when $m = \rho$. In practice, this is unlikely to be the case as SSRE is specifically utilised in scenarios where the ICC is subject to considerable uncertainty. Accordingly, in what follows, we evaluate the performance of a range of SSRE methods across possible values of the ICC, specifically $\rho \in [0.01, 0.2]$. A framework is considered efficient if it has a low (preferably positive) bias and MSE. Given that a low bias may still be associated with

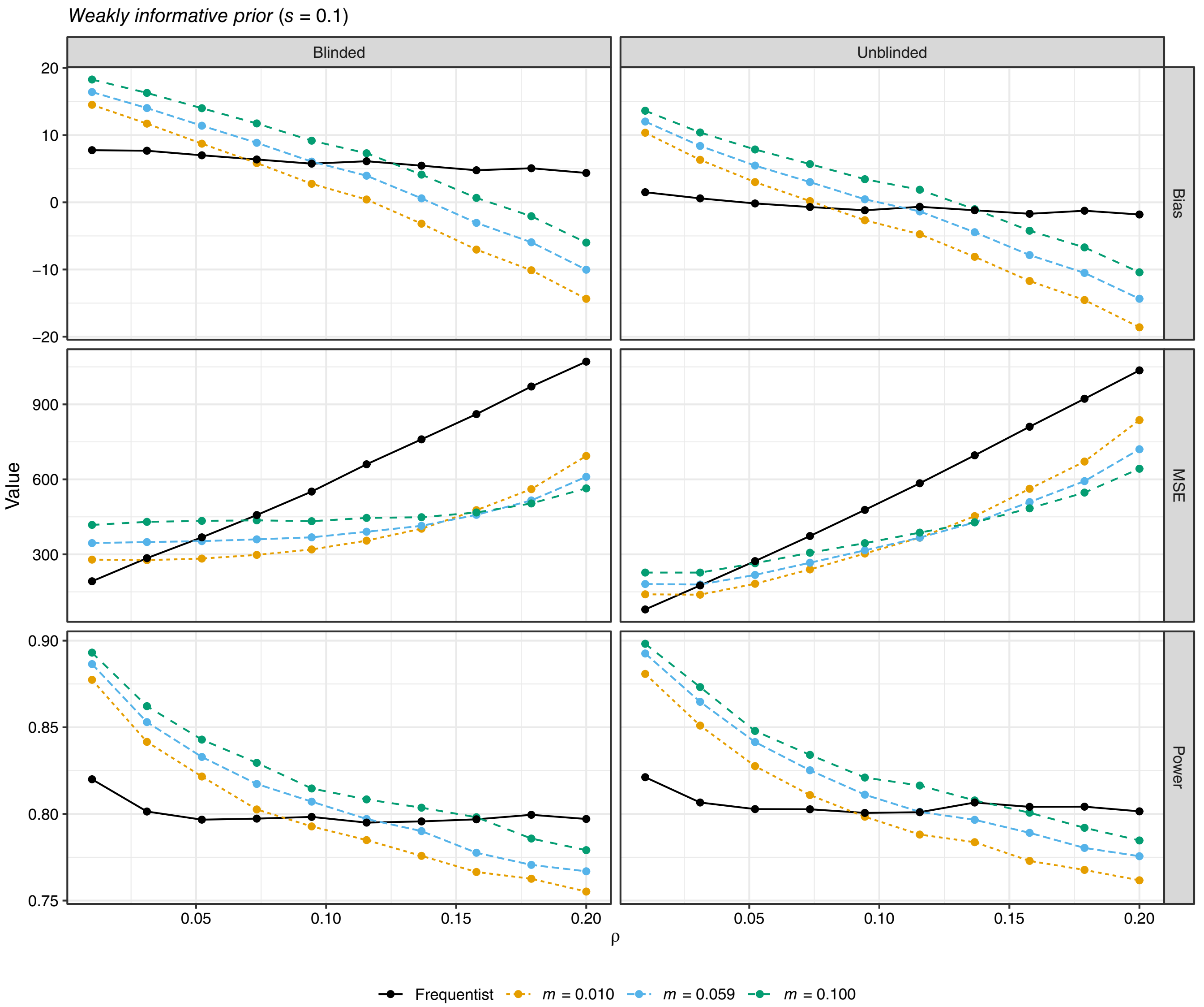

FIGURE 5 The bias, mean square error (MSE), and power of the frequentist and hybrid methods are shown as a function of the intra-cluster correlation (ρ). Results are faceted by the use of blinded versus unblinded sample size re-estimation. For the hybrid approach, all combinations of $m = 0.01, 0.059, 0.1$ and $s = 0.10$ are considered.

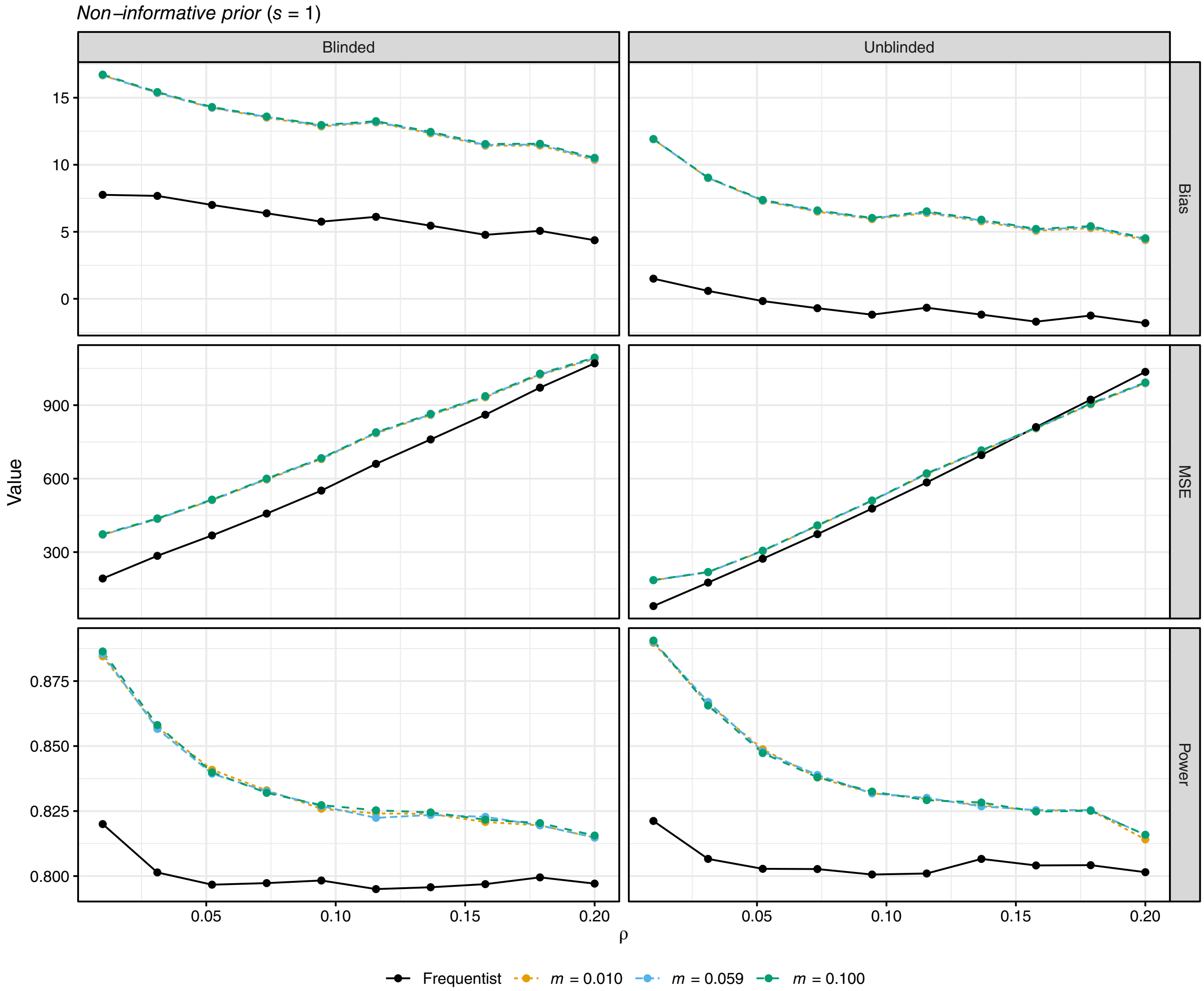

FIGURE 6 The bias, mean square error (MSE), and power of the frequentist and hybrid methods are shown as a function of the intra-cluster correlation (ρ). Results are faceted by the use of blinded versus unblinded sample size re-estimation. For the hybrid approach, all combinations of $m = 0.01, 0.059, 0.1$ and $s = 1.00$ are considered.

high variability in the re-estimated sample size, we place a higher emphasis on frameworks with low MSE. Recall a design exhibits zero bias if the re-estimated sample size is equal to the Oracle sample size on average. Note that in the Supplementary Material, we further demonstrate the effect of prior misspecification on SSRE performance by examining the proportions of correctly powered, underpowered, and overpowered trials, as defined in Section 3.3. The results for the bias, MSE, and power are presented in Figures 4,5 and 6.

The frequentist framework seems relatively stable in terms of the performance measures across the considered values of ρ. As expected, the final sample sizes in the frequentist approach are unbiased for the unblinded model, and subject only to small bias for the blinded model.

For highly informative priors, the hybrid approach is only approximately unbiased in terms of the re-estimated sample size when the prior mean is equal to ρ, for both blinded and unblinded models. Thus, the final sample size in the hybrid SSRE is considerably underestimated when the value of the ICC is larger than the prior mean. Given that an underestimated sample size results in an underpowered trial and vice versa, the negative relationship between the bias and ICC is also observed in the power. When compared to the frequentist, the hybrid techniques offer a lower MSE if the ICC is within a specific range, with the range dependent on the values of m. For example, when $m = 0.01$, the hybrid method has a lower MSE if the ICC is less than 0.05 for both blinded and unblinded models. Whereas, when $m = 0.059$, a lower MSE is observed in the hybrid framework if $\rho \in [0.03, 0.12]$ in the blinded model and if $\rho \in [0.03, 0.1]$ in the unblinded model. When $m = 0.1$, the hybrid method can reduce the MSE if $\rho \in [0.06, 0.17]$ in the blinded model and if $\rho \in [0.06, 0.15]$ in the unblinded model.

When using weakly informative priors, the interplay between bias and power is also exhibited in the same way as when using highly informative priors. However, the power curves for weakly informative priors are less steep, provided the ICC is not very small. As a result, there is approximately a 5% loss or gain in power compared to the desired power over a wide range of ICCs, specifically, when $\rho \in [0.025, 0.2]$. Concerning the MSE, when $m = 0.01$, the hybrid framework performs better when $\rho \geq 0.03$ in the blinded model and when $\rho \geq 0.025$ in the unblinded model. With $m = 0.059$, the hybrid framework is better than the frequentist in terms of MSE when $\rho \geq 0.05$ in the blinded model and when $\rho \geq 0.03$ in the unblinded model. When a larger prior mean is selected ($m = 0.1$), the hybrid method becomes more powerful when the ICC is greater than 0.07 in the blinded model and greater than 0.05 in the unblinded model. We note that when using weakly informative priors, the unblinded approaches do indeed overestimate the required number of clusters when $\rho = m$. This is a consequence of (a) the weight still given to larger values of the ICC in the posterior distribution for ρ and (b) the differences induced by placing a requirement on expected power versus frequentist power. This bias is a fact that should be noted when choosing a suitable design, but in general it appears small. We note also that as a consequence of the discretization of the number of required clusters, it is possible to have a small negative bias and still achieve the desired power.

For non-informative priors, the MSE aligns closely with the frequentist approach in the unblinded model and exhibits a slightly worse performance in the blinded model for hybrid designs when compared to the frequentist approach. Observed gains over the frequentist approach in terms of power are a result of an increased positive bias in the hybrid approach. As noted in Figure 2, the prior means have less impact on the final sample size if the prior is non-informative, evidenced by the lack of differentiation in the hybrid lines in Figure 6.

4 DISCUSSION

SSRE using a frequentist approach mostly abates the difficulties of obtaining precise estimates of the ICC during the trial design stage; yet, it has some practical issues. Notable among these issues is a large variation in the re-estimated sample size, a consequence of variability around the re-estimated ICC.[18] In this paper, we have demonstrated how a hybrid approach to SSRE could address this issue whilst effectively controlling the type I error rate.

Previous studies have demonstrated that the ICC is often low. Sarkodie et al,[11] based on 34 trials, found that assumed ICCs were positively skewed on the interval [0.002, 0.5], with a median value of 0.05. Similarly, Offorha et al,[22] in their analysis of 86 trials in health services research, reported that the observed ICC for the primary outcome had a mean of 0.06 and an IQR of ($0.001 - 0.060$). Given this evidence, priors that assign low weights to large ICC values may often be considered highly plausible in CRT settings. According to Figure 1, both highly and weakly informative priors become candidates for use in practice as they eliminate what may likely be an over-reaction at the re-estimation point from the observation of an extremely large ICC value. However, the use of highly informative priors is discouraged as it defeats the purpose of the SSRE procedure, with the re-estimated number of clusters approximately constant regardless of the interim data. As expected, the use of a non-informative prior results in a design that performs similarly to a frequentist approach. Therefore, we propose that the use of a weakly informative prior in practice may result in a SSRE procedure with advantageous performance. Indeed, such weakly informative priors were demonstrated to be robust in terms of the MSE over a wide range of ICC values commonly observed in practice.

The finding that a weakly informative prior effectively facilitates the SSRE process is an important benefit of the hybrid technique since some information (eg, in the form of routinely gathered data or expert opinion) may often exist in practice for prior construction. Nonetheless, it is important to note that the low MSE from the use of the weakly informative prior can come at the cost of requiring more clusters than necessary (equivalently, at the cost of more power than was desired). Since higher power isn't necessarily always optimal (given, eg, financial considerations), it is imperative to determine through simulation the optimal re-estimation procedure for a given trial. This procedure should aim to achieve the desired power while maintaining a very low MSE. One possible way of striking a balance between the MSE and power in practice could be to set the EP requirement to a value lower than that of the frequentist power requirement. This approach may increase the likelihood of achieving an efficient MSE across a wide range of ρ while maintaining a power similar to that of the frequentist SSRE procedure.

The results also show that in both frameworks, the interim ICC estimates from the blinded model are biased (overestimated) when there is a non-zero treatment effect ($\mu = \delta$). Previous studies have proposed methods to adjust for this bias. Specifically, available methods to remove this bias include (a) using block randomisation, but this approach has been demonstrated to introduce large variability in the re-estimated sample size, which directly contradicts the problem we aim to address in this paper or (b) 'guessing' the treatment effect and subtracting the bias under this effect. This approach only removes the bias in the situation where the stated effect is correctly guessed, and has actually been argued to be disadvantageous as a small positive bias in the estimated variance parameters can help the re-estimation procedure achieve the desired power. See Friede and Kieser for further details.[1] For this reason, we did not consider attempting to adjust for the bias in the interim ICC estimate under the blinded model. Similarly, we did not modify the final analysis to account for the interim analysis as the observed type I error rate inflation was small.

A practical consideration for SSRE designs is the choice of sample size for the interim analysis. Studies have shown that estimates from pilot studies, which typically employ small sample sizes are frequently imprecise.[3,23] To some researchers, 40 clusters are inadequate to yield precise estimates of the ICC in the frequentist framework.[24] When there is such uncertainty around the assumed ICC, the hybrid approach should be preferred over the frequentist. This is because even if the best guess estimate (m) based on existing data is inaccurate, a weakly informative prior may often still be advantageous since the performance of the SSRE procedure does not depend heavily on the accuracy of the prior mean.

This study has some limitations. First, the use of a truncated normal prior makes it challenging to target small values of ρ when s is fixed in a way that renders the prior weakly informative or non-informative. As noted previously, this can lead to an overestimation of the sample size on average for small values of ρ, regardless of the chosen value of m in such scenarios. In this study, we have employed the truncated normal prior without adjusting for these biases as a proof of concept. In practice, trial designers might consider varying both s and m if they believe ρ is sufficiently small or consider using a different prior altogether. While we believe that an extension of this approach to different priors and outcome data might yield similar results, future studies detailing the results and complexities of such models could be helpful. While we believe that an extension of this approach to different priors and outcome data might yield similar results, future studies detailing the results and complexities of such models could be helpful. Another limitation was the consideration of only one CRT design (the parallel group). Hence, inferences from this study cannot be generalised to other CRT designs.

Additionally, though some studies have defined SSRE in the context of updating cluster sizes,[18] we have focused on updating the number of clusters instead, as this generally has a higher impact on power. This is consistent with several previous studies which sought increases in power through increasing the number of clusters.[25-27] However, we acknowledge that due to logistical constraints, it may in some instances be more difficult to add more clusters than to increase the number of participants per cluster.[27] In such scenarios, the methods discussed here could be readily extended to re-estimate the cluster size; we illustrate how this can be achieved in the Supplementary Material. Whilst the methodology can be easily extended, we note though that due to the diminishing returns in the power gain from increasing the cluster sizes, it may not always be possible even under re-estimation to achieve the desired power.

Another limitation of the proposed methodology is the assumption of a normal outcome with Fisher's formula then used to estimate the variance of the ICC estimate. Nevertheless, as highlighted by Turner et al,[9] this method for estimating the variance of the ICC can be extended to, for example, binary outcomes. This would facilitate performing re-estimation for a trial that assumes a binary outcome. Additionally, we assumed a linear mixed model. Nonetheless, we see no compelling rationale for why the results would differ greatly if, for example, a Generalized Estimating Equation based analysis was assumed.

In conclusion, the MSE of the hybrid approach becomes similar to the frequentist approach when using a completely uninformative prior, whereas a highly informative prior does better if the prior is correct (and is poor otherwise). Utilising a weakly informative prior performs well, demonstrating robustness in terms of the MSE over a wide range of ICC values typically observed in practice. A simulation study can be useful to assess when a hybrid approach may offer utility in terms of overcoming known issues with the frequentist SSRE approach.

ACKNOWLEDGEMENTS

JMSW is funded by a NIHR Research Professorship (NIHR301614).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Github at https://github.com/sks2023/article_codes.

ORCID

Samuel K Sarkodie https://orcid.org/0000-0002-9296-8216
James MS Wason https://orcid.org/0000-0002-4691-126X

REFERENCES

1. Friede T, Kieser M. Blinded sample size re-estimation in superiority and noninferiority trials: bias versus variance in variance estimation. *Pharm Stat*. 2013;12(3):141-146. doi:10.1002/pst.1564
2. Campbell MK, Grimshaw JM, Elbourne DR. Intracluster correlation coefficients in cluster randomized trials: empirical insights into how should they be reported. *BMC Med Res Methodol*. 2004;4:1-5. doi:10.1186/1471-2288-4-9
3. Ip EH, Wasserman R, Barkin S. Comparison of intraclass correlation coefficient estimates and standard errors between using cross-sectional and repeated measurement data: the safety check cluster randomized trial. *Contempor Clin Trials*. 2011;32(2):225-232. doi:10.1016/j.cct.2010.11.001
4. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health*. 2004;94(3):423-432. doi:10.2105/AJPH.94.3.423
5. Wu S, Crespi CM, Wong WK. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contempor Clin Trials*. 2012;33(5):869-880. doi:10.1016/j.cct.2012.05.004
6. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JA, Burney PG. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technol Assessm*. 1999;3(5):1-98. doi:10.3310/hta3050
7. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. *BMJ (Online)*. 2012;345(7881):1-21. doi:10.1136/bmj.e5661
8. Ukoumunne OC. A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Stat Med*. 2002;21(24):3757-3774. doi:10.1002/sim.1330
9. Turner RM, Prevost AT, Thompson SG. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Stat Med*. 2004;23(8):1195-1214. doi:10.1002/sim.1721
10. Jones BG, Streeter AJ, Baker A, Moyeed R, Creanor S. Bayesian statistics in the design and analysis of cluster randomised controlled trials and their reporting quality: a methodological systematic review. *Systemat Rev*. 2021;10(1):1-14. doi:10.1186/s13643-021-01637-1
11. Sarkodie SK, Wason JMS, Grayling MJ. A hybrid approach to comparing parallel-group and stepped-wedge cluster-randomized trials with a continuous primary outcome when there is uncertainty in the intra-cluster correlation. *Clin Trials*. 2022;20(1):59-70. doi:10.1177/17407745221123507
12. Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Stat Med*. 1986;5(1):1-13. doi:10.1002/sim.4780050103
13. O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. *Pharmaceut Stat*. 2005;4(3):187-201. doi:10.1002/pst.175
14. Kunzmann K, Grayling MJ, Lee KM, Robertson DS, Rufibach K, Wason JMS. A review of bayesian perspectives on sample size derivation for confirmatory trials. *Am Stat*. 2021;75(4):424-432. doi:10.1080/00031305.2021.1901782
15. Grayling MJ, Mander AP, Wason JM. Blinded and unblinded sample size reestimation procedures for stepped-wedge cluster randomized trials. *Biometr J*. 2018;60(5):903-916. doi:10.1002/bimj.201700125
16. Lake S, Kammann E, Klar N, Betensky R. Sample size re-estimation in cluster randomization trials. *Stat Med*. 2002;21(10):1337-1350. doi:10.1002/sim.1121
17. Schie v S, Moerbeek M. Re-estimating sample size in cluster randomised trials with active recruitment within clusters. *Stat Med*. 2014;33(19):3253-3268.
18. Hemming K, Martin J, Gallos I, Coomarasamy A, Middleton L. Interim data monitoring in cluster randomised trials: practical issues and a case study. *Clin Trials*. 2021;18(5):552-561. doi:10.1177/17407745211024751
19. Fisher RA. *Statistical Methods for Research Workers*. CT: Darien; 1970.
20. Hankonen N, Heino MT, Araujo-Soares V, et al. 'Let's move it'–a school-based multilevel intervention to increase physical activity and reduce sedentary behaviour among older adolescents in vocational secondary schools: a study protocol for a cluster-randomised trial. *BMC Public Health*. 2016;16(1):1-15. doi:10.1186/s12889-016-3094-x
21. U.S. Food & Drug Administration. Placebos and Blinding in Randomized Controlled Cancer Clinical Trials for Drug and Biological Products Guidance for Industry. 2019. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/placebos-and-blinding-randomized-controlled-cancer-clinical-trials-drug-and-biological-products.
22. Offorha BC, Walters SJ, Jacques RM. Statistical analysis of publicly funded cluster randomised controlled trials: a review of the national institute for health research journals library. *Trials*. 2022;23(1):1-16. doi:10.1186/s13063-022-06025-1
23. Eldridge SM, Costelloe CE, Kahan BC, Lancaster GA, Kerry SM. How big should the pilot study for my cluster randomised trial be? *Stat Methods Med Res*. 2016;25(3):1039-1056. doi:10.1177/0962280215588242
24. 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
25. Koepsell TD, Wagner EH, Cheadle AC, et al. Selected methodological issues in evaluating community-based health promotion and disease prevention programs. *Ann Rev Public Health*. 1992;13(1):31-57.
26. Thompson SG, Pyke SDM, Hardy RJ. The design and analysis of paired cluster randomized trials: an application of meta-analysis techniques. *Stat Med*. 1997;16(18):2063-2079. doi:10.1002/(SICI)1097-0258(19970930)16:18¡2063::AID-SIM642¿3.0.CO;2-8

27. Van Breukelen GJ, Candel MJ. Calculating sample sizes for cluster randomized trials: we can keep it simple and efficient! *J Clin Epidemiol*. 2012;65(11):1212-1218. doi:10.1016/j.jclinepi.2012.06.002

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Sarkodie SK, Wason JM, Grayling MJ. A hybrid approach to sample size re-estimation in cluster randomized trials with continuous outcomes. *Statistics in Medicine*. 2024;43(24):4736-4751. doi: 10.1002/sim.10205