University of Central Lancashire
UCLan

Central Lancashire Online Knowledge (CLoK)

Title	Hybrid sample size calculations for cluster randomised trials using assurance
Type	Article
URL	https://clok.uclan.ac.uk/id/eprint/48625/
DOI	https://doi.org/10.1177/17407745241312635
Date	2025
Citation	Williamson, S. Faye, Tishkovskaya, Svetlana and Wilson, Kevin J. (2025) Hybrid sample size calculations for cluster randomised trials using assurance. Clinical Trials, 22 (5). pp. 517-526. ISSN 1740-7745
Creators	Williamson, S. Faye, Tishkovskaya, Svetlana and Wilson, Kevin J.

It is advisable to refer to the publisher's version if you intend to cite from the work.
https://doi.org/10.1177/17407745241312635

For information about Research at UCLan please go to http://www.uclan.ac.uk/research/

All outputs in CLoK are protected by Intellectual Property Rights law, including Copyright law. Copyright, IPR and Moral Rights for the works on this site are retained by the individual authors and/or other copyright owners. Terms and conditions for use of this material are defined in the http://clok.uclan.ac.uk/policies/

Check for updates

Article

CLINICAL TRIALS

Clinical Trials
1–10
© The Author(s) 2025

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17407745241312635
journals.sagepub.com/home/ctj


Hybrid sample size calculations for cluster randomised trials using assurance

S. Faye Williamson[1], **Svetlana V Tishkovskaya**[2] and **Kevin J Wilson**[3]

Abstract

Background/Aims: Sample size determination for cluster randomised trials is challenging because it requires robust estimation of the intra-cluster correlation coefficient. Typically, the sample size is chosen to provide a certain level of power to reject the null hypothesis in a two-sample hypothesis test. This relies on the minimal clinically important difference and estimates for the overall standard deviation, the intra-cluster correlation coefficient and, if cluster sizes are assumed to be unequal, the coefficient of variation of the cluster size. Varying any of these parameters can have a strong effect on the required sample size. In particular, it is very sensitive to small differences in the intra-cluster correlation coefficient. A relevant intra-cluster correlation coefficient estimate is often not available, or the available estimate is imprecise due to being based on studies with low numbers of clusters. If the intra-cluster correlation coefficient value used in the power calculation is far from the unknown true value, this could lead to trials which are substantially over- or under-powered.

Methods: In this article, we propose a hybrid approach using Bayesian assurance to determine the sample size for a cluster randomised trial in combination with a frequentist analysis. Assurance is an alternative to traditional power, which incorporates the uncertainty on key parameters through a prior distribution. We suggest specifying prior distributions for the overall standard deviation, intra-cluster correlation coefficient and coefficient of variation of the cluster size, while still utilising the minimal clinically important difference. We illustrate the approach through the design of a cluster randomised trial in post-stroke incontinence and compare the results to those obtained from a standard power calculation.

Results: We show that assurance can be used to calculate a sample size based on an elicited prior distribution for the intra-cluster correlation coefficient, whereas a power calculation discards all of the information in the prior except for a single point estimate. Results show that this approach can avoid misspecifying sample sizes when the prior medians for the intra-cluster correlation coefficient are very similar, but the underlying prior distributions exhibit quite different behaviour. Incorporating uncertainty on all three of the nuisance parameters, rather than only on the intra-cluster correlation coefficient, does not notably increase the required sample size.

Conclusion: Assurance provides a better understanding of the probability of success of a trial given a particular minimal clinically important difference and can be used instead of power to produce sample sizes that are more robust to parameter uncertainty. This is especially useful when there is difficulty obtaining reliable parameter estimates.

Keywords

Assurance, Bayesian design, cluster randomised trials, expected power, hybrid approach, intra-cluster correlation, minimal clinically important difference, sample size determination

[1]Biostatistics Research Group, Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK
[2]Lancashire Clinical Trials Unit, University of Central Lancashire, Preston, UK
[3]School of Mathematics, Statistics & Physics, Newcastle University, Newcastle upon Tyne, UK

Corresponding author:
S. Faye Williamson, Biostatistics Research Group, Population Health Sciences Institute, Newcastle University, 4th Floor Ridley Building 1, Queen Victoria Road, Newcastle upon Tyne NE1 7RU, UK.
Email: faye.williamson@newcastle.ac.uk

Background

Cluster randomised trials (CRTs) are a type of randomised controlled trial (RCT) in which randomisation is at the cluster-level, rather than the individual-level as in standard RCTs. This means that 'groups' of individuals (e.g. general practices, schools or communities) are randomly allocated to different interventions (e.g. vaccination programmes or behavioural interventions). A common reason for implementing this design is to mitigate the risk of contamination or where individual randomisation is not feasible. Other justifications are detailed in the study by Eldridge and Kerry.[1]

Individuals within a cluster are likely to share similar characteristics (e.g. demographics), as well as be exposed to extraneous factors unique to the cluster (e.g. delivery of the intervention by the same healthcare professional). Consequently, outcomes from members of the same cluster are often correlated, which can be quantified by the intra-cluster correlation coefficient (ICC). This lack of independence reduces the statistical power compared to a standard RCT of the same size, meaning that the sample size needs to be inflated to allow for the clustering effect.

Various methods for sample size determination in CRTs exist,[2,3] which all rely on estimation of the ICC. In practice, ICC estimates are typically based on pilot studies, but these are often too small to provide precise and reliable estimates.[4] An alternative simple approach is to use a conservative estimate of the ICC (e.g. the upper confidence interval limit) in the sample size calculation.[5] However, this can lead to over-powered and unnecessarily large trials. A more reliable method is to combine ICC estimates from multiple sources, such as previous trials or databases listing ICC estimates,[6] and use information on patterns in ICCs.[7] This raises further issues such as how to effectively combine the ICC estimates, how to adequately reflect their varying degrees of relevance to the planned trial and how to capture the uncertainty in the individual ICC estimates.[8] It was suggested to consider integrating over a range of possible ICC values, determined by confidence intervals obtained using methods in the study by Ukoumunne,[9] to provide an 'average' sample size with respect to the ICC. However, this does not consider the uncertainty present in other design parameters, such as the treatment effect and variability of the outcome measures. Furthermore, it assumes that each value of the ICC is equally likely. Other approaches to deal with uncertainty in the ICC include sample size re-estimation[10,11] and robust designs, such as maximin designs.[12] For the latter approach, a range rather than a prior is used for the ICC.

Utilising a Bayesian approach for the trial design, in which prior distributions are assigned to the unknown design parameters such as the ICC, could further circumvent these issues and is particularly useful in settings where ICC estimates are not readily available. In the CRT literature, prior distributions for the ICC have been proposed based on subjective beliefs[13] and single or multiple ICC estimates,[14] which may be weighted by relevance of outcomes and patient population.[15] These are used to estimate a distribution for the power of the planned trial for a given sample size. Within the Bayesian framework, uncertainty in other design parameters can be incorporated into the sample size calculation in a similar way, and the relative likelihood of different parameter values is encompassed through specification of the prior distribution. For example, Sarkodie et al.[16] assigned a prior to the overall standard deviation, in addition to the ICC, then described a 'hybrid' approach to determine the sample size required to attain a desired 'expected power', defined as a weighted average of the probability that the null hypothesis is rejected (with weights determined by the priors).

Hybrid approaches, which combine a Bayesian design with a frequentist analysis of the final trial data, have gained increasing popularity, particularly with respect to standard RCTs.[17,18] In this article, we adopt a hybrid approach by using the Bayesian concept of 'assurance' to determine the sample size for a two-arm parallel-group CRT with a Wald test for the analysis. In contrast to traditional frequentist power, which represents a conditional probability that the trial is a success, given the values chosen for the design parameters and the hypothesised treatment effect, assurance typically refers to the 'unconditional' probability that the trial will be 'successful'.[19] We modify this definition by conditioning on the minimal clinically important difference (MCID) instead of assigning a prior distribution to, and integrating over, the treatment effect as is standard practice.[17,20] This is more representative of the design stage of a trial, in which the treatment effect is typically fixed a priori by investigators. Moreover, this ensures that the assurance will tend to one as the sample size increases so can be used analogously to traditional power, thus aiding interpretation.

A key consideration when applying a Bayesian design is how to specify suitable prior distributions. In contrast to the study by Sarkodie et al.,[16] which assumes independent priors on the ICC and standard deviation, we suggest a joint prior distribution for these parameters, as described in the Methods section. In addition, we account for the fact that many CRTs have unequal cluster sizes by defining a prior distribution on the coefficient of variation of cluster size. This is often overlooked in standard sample size calculations for CRTs.[21,22]

Our approach is motivated by a parallel-group CRT, Identifying Continence OptioNs after Stroke (ICONS), outlined in the 'Results' section. We illustrate the effects of redesigning this trial using the entire ICC prior distribution to inform sample size

determination via an assurance calculation, rather than relying on a single point estimate from this distribution as in Tishkovskaya et al.[23] The impacts of varying the ICC prior distributions on the chosen sample size are evaluated. We perform sensitivity analyses on other design parameters in an additional simulation study provided in the Appendix.

Jones et al.[24] summarise the current state of play regarding the use of Bayesian methods in CRTs. In doing so, they highlight the 'need for further Bayesian methodological development in the design and analysis of CRTs ... in order to increase the accessibility, availability and, ultimately, use of the approach'. This article is, therefore, a timely contribution.

Methods

Analysis for CRTs

Suppose that we are designing a two-arm, parallel-group CRT assuming 1:1 randomisation of clusters and normally distributed outcomes. A common analysis following the trial is to use a linear mixed-effects model. That is, if Y_{ij} is the response for individual $i = 1, \ldots, n_j$ in cluster $j = 1, \ldots, J$, then

$$Y_{ij} = \alpha + X_j\delta + c_j + e_{ij}, \tag{1}$$

where α is an intercept term; X_j is a binary variable that takes the value 1 if cluster j is allocated to the treatment arm and 0 if it is allocated to the control arm, so that δ represents the treatment effect; $c_j \sim \mathrm{N}(0, \sigma_b^2)$ is a random cluster effect with σ_b^2 denoting the between-cluster variation and $e_{ij} \sim \mathrm{N}(0, \sigma_w^2)$ is the individual-level error with σ_w^2 denoting the within-cluster variation.

The ratio of the variability between clusters σ_b^2 to the total variability $\sigma^2 = \sigma_b^2 + \sigma_w^2$ determines the extent to which clustering induces correlations between outcomes for individuals in the same cluster. This is referred to as the ICC, $\rho = \sigma_b^2/\sigma^2$.[25]

The superiority of the treatment is assessed via a hypothesis test of $H_0 : \delta \leqslant 0$ versus $H_1 : \delta > 0$. Using a Wald test, assuming asymptotic normality, the test statistic is $Z = \hat{\delta}/\sqrt{\mathrm{Var}\left(\hat{\delta}\right)}$, where $\hat{\delta}$ is the estimate of δ and $\mathrm{Var}\left(\hat{\delta}\right) = 4\sigma^2[1 + \{(\nu^2 + 1)\bar{n} - 1\}\rho]/J\bar{n}$,[12] where $\bar{n}$ is the average sample size per cluster and ν is the coefficient of variation of cluster size, that is, the ratio of the standard deviation of cluster sizes to the mean cluster size.

Choosing a sample size using assurance

The power of the one-sided Wald test for significance level α can be approximated[4] by

$$P(n \mid \delta, \psi) = \Phi\left(\delta\sqrt{\frac{J\bar{n}}{4\sigma^2[1 + \{(\nu^2 + 1)\bar{n} - 1\}\rho]}} - z_{1-\alpha}\right), \tag{2}$$

where $z_{1-\alpha}$ is the $100(1-\alpha)\%$ percentile of the standard normal distribution and $\psi = (\sigma, \rho, \nu)$ is the vector of 'nuisance' parameters, excluding the treatment effect. For a two-sided Wald test, $z_{1-\alpha}$ would be replaced by $z_{1-\alpha/2}$. For equal cluster sizes, the power function would take the same form as equation (2), with $\nu = 0$ and $\bar{n} = n_j = n$.

In a standard power calculation, the sample size would be chosen as the smallest value which gives 80% or 90% power, based on values for $\theta = (\delta, \psi)$. The treatment effect δ could be specified as the MCID or an estimate based on a pilot study, similar historical trials or expert knowledge. The values used for ψ are typically estimates.

Alternatively, we can use assurance to choose the sample size. Whereas the power is conditioned on the chosen estimates for ψ and possibly δ, the assurance represents the 'unconditional' probability that an RCT will achieve a successful outcome.[26] Assurance has been used almost exclusively when the value to be used for δ is an estimate. In this case, suppose that the CRT is a success if the null hypothesis is rejected by the Wald test for δ. Rather than using point estimates for θ, we could assign a prior distribution $\pi(\theta)$ to it and define the assurance $A(n)$ as the power, averaged over the uncertainty in θ:

$$\begin{aligned} A(n) &= \int_\theta \Pr\left(H_0 \textit{rejected} \mid \theta\right)\pi(\theta)d\theta, \\ &= \int_\theta P(n \mid \theta)\pi(\theta)d\theta. \end{aligned} \tag{3}$$

One disadvantage of the assurance is that it tends to $\Pr(\delta > 0)$ under $\pi(\delta)$ as the sample size increases. That is, unlike power, there may be no sample size for which the assurance is above the typical thresholds of 80% or 90%. Kunzmann et al.[17] avoid this by conditioning the prior distribution for δ on $\delta > 0$ in the assurance calculation. In this article, we consider the following alternative approach.

The assurance in equation (3) assumes that we choose δ in the sample size calculation based on a priori considerations of the likelihood of the treatment effect. Instead, we consider the assurance in conjunction with a trial planned using the relevance argument, that is, using the MCID for δ, δ_M. In this case, there is no need to define a prior distribution for δ, and the assurance reduces to:

$$A(n \mid \delta_M) = \int_{\psi} P(n \mid \delta_M, \psi)\pi(\psi)d\psi.$$

The advantage of this is that the assurance will now tend to 1 as the sample size increases.

To evaluate the assurance in practice, we sample values of $(\psi_j)_{j=1,\ldots,S}$ from the prior distribution $\pi(\psi)$ for some large number of samples S, and use Monte Carlo simulation to approximate the assurance as

$$\tilde{A}(n \mid \delta_M) = \frac{1}{S}\sum_{j=1}^{S} P\left(n \mid \delta_M, \psi_j\right),$$
$$\approx \frac{1}{S}\sum_{j=1}^{S} \Phi\left(\delta_M \sqrt{\frac{J\bar{n}}{4\sigma_j^2[1 + \{\left(\nu_j^2 + 1\right)\bar{n} - 1\}\rho_j]}} - z_{1-\alpha}\right). \quad (4)$$

Specification of priors

To evaluate the assurance, we are required to specify a prior distribution for ψ. This simplifies to specifying marginal prior distributions for each parameter if they can be assumed independent. Given that σ^2 and ρ are both functions of σ_w^2 and σ_b^2, it is unlikely that σ and ρ can be assumed independent. Therefore, we consider a joint prior distribution for (σ, ρ) and a marginal prior distribution for ν. In order for the assurance to be a meaningful representation of the probability that the null hypothesis is rejected, these prior distributions should be informative, representing the current state of knowledge about the possible parameter values. This is an elicitation problem, and information to specify the priors can be obtained from relevant past data, expert knowledge or a combination (an example of this is given in the 'Results' section).

Since the coefficient of variation can only take positive values, a gamma distribution $\nu \sim Gamma(a_\nu, b_\nu)$ is a sensible choice for its prior distribution. The hyperparameters a_ν and b_ν could be chosen based on previous studies, via modelling or by eliciting expert knowledge.[4]

One way to specify a joint prior distribution for (σ, ρ) is to assign independent priors to σ_b^2 and σ_w^2, which will induce a correlation between ρ and σ^2. If we sample values of σ_b and σ_w from their priors, we can obtain samples from the joint prior of (σ, ρ). Typical choices of prior distributions for σ_b^2 and σ_w^2 are (inverse) gamma distributions because they provide conjugacy.

An alternative approach, relevant to our application, is to specify the joint distribution between ρ and σ directly. For example, we can utilise a bivariate copula to encode the dependence between the parameters. A bivariate copula is a joint distribution function on $[0, 1]^2$ with standard uniform marginal distributions.[27] It can be used to construct a joint prior for ρ and σ via

$$\pi_{\rho,\sigma}(\rho, \sigma) = \pi_\rho(\rho)\pi_\sigma(\sigma)c(u, v),$$

where π_ρ and π_σ are marginal prior distributions, $c(u, v)$ is the bivariate copula density function evaluated at $u = F_\rho(\rho)$ and $v = F_\sigma(\sigma)$ for prior cumulative distribution functions (CDFs) F_ρ and F_σ. One simple choice is the Gaussian copula:

$$c(u, v) = \frac{\partial^2}{\partial u \partial v}\Phi_\gamma\left(\Phi^{-1}(u), \Phi^{-1}(v)\right),$$

where Φ_γ is the CDF of the bivariate standard normal distribution with correlation γ, and Φ^{-1} is the inverse univariate standard normal CDF. The advantage of this structure is that it allows specification of the marginal prior distributions for ρ and σ separately to their dependence, which is given by γ.

Results

The ICONS post-stroke incontinence CRT

The approach developed in this article is motivated by a planned parallel-group CRT, 'Identifying Continence OptioNs after Stroke' (ICONS), which investigates the effectiveness of a systematic voiding programme in secondary care versus usual care on post-stroke urinary incontinence for people admitted to NHS stroke units.[28] The primary outcome is the severity of urinary incontinence at three months post-randomisation, measured using the International Consultation on Incontinence Questionnaire.[29] Although a feasibility trial, ICONS-I[30] was conducted, the resulting ICC estimate was of low precision and could not be used as a reliable single source to inform the planning of the proposed trial.

ICONS, therefore, considered a Bayesian approach to combine multiple ICC estimates from 16 previous related trials. The opinions of eight experts regarding the relevance of the previous ICC estimates were elicited[31] and used to assign weights to each study and each outcome within a study. The elicited study and outcome weights were combined using mathematical aggregation[32] and incorporated into a Bayesian hierarchical model following the method by Turner et al.[15] The resulting constructed ICC distribution had a posterior median of $\hat{\rho} = 0.0296$ with a 95% credible interval of $(0.00131, 0.330)$. Details of the expert elicitation process and modelling are described in the study by Tishkovskaya et al.[23]

For the ICONS CRT, the sample size was chosen to give 80% power with a 5% significance level to detect $\delta_M = 2.52$ using a two-tailed independent-samples t-test and a common standard deviation σ of 8.32 obtained

from the ICONS-I feasibility trial. The ICC was assumed to be less than or equal to $\hat{\rho} = 0.0296$. It was assessed as realistic to recruit between 40 and 50 stroke units, which required total sample sizes of $N = 480$ and $N = 450$, respectively, and an average sample size per cluster of $n = 12$ and $n = 9$, respectively. The original sample size calculation assumed equally sized clusters (i.e. $\nu = 0$). However, if we consider unequal cluster sizes with $\nu = 0.49$ (obtained from ICONS-I) and apply the Wald test, the required sample sizes remain the same.

Redesigning the ICONS CRT using assurance

We consider assurance as an alternative to power to determine the sample size for the ICONS CRT. This seems like a more natural approach given the uncertainty in the ICC and the extensive elicitation and modelling that was conducted to construct the ICC posterior distribution (which forms the prior distribution for the assurance-based sample size calculation). Moreover, assurance incorporates the full ICC distribution into the sample size calculation, rather than relying on a single point estimate from it as in the power calculation.

We consider the following two forms of assurance.

Assurance based on the ICC prior only. In the first case, we fix (σ, ν) using the point estimates obtained from ICONS-I ($\hat{\sigma} = 8.32$ and $\hat{\nu} = 0.49$) and only consider the assurance with respect to the ICC. We sample $S = 10,000$ values of ρ from its distribution (see Figure 1) and approximate the assurance using equation (4).

To obtain an assurance of 80%, the resulting average sample sizes per cluster are $\bar{n} = 17$ for $J = 40$ clusters ($J/2 = 20$ per arm) and $\bar{n} = 11$ for $J = 50$ clusters ($J/2 = 25$ per arm), requiring total sample sizes of $N = 680$ and $N = 550$, respectively (see Table 1). Thus, the inclusion of uncertainty in the ICC results in a larger sample size than when using the posterior median ICC, but provides a more realistic and robust study design. Compared to the classical approach, the total sample size attained is smaller for the smaller number of clusters.

The left-hand side plot of Figure 2 illustrates the trade-off between cluster size and assurance/power, for $J = 40$ clusters ($J/2 = 20$ per arm). The power calculation based on the median from the elicited prior distribution of ρ is represented by the red curve and the assurance with a prior on ρ only by the black curve. We see that the assurance requires a larger sample size than power when the target lies above 0.5. We also include the power curve corresponding to the commonly used approach of taking the median of the 34 ICC estimates (blue line). For a target power of 0.8 (horizontal line),

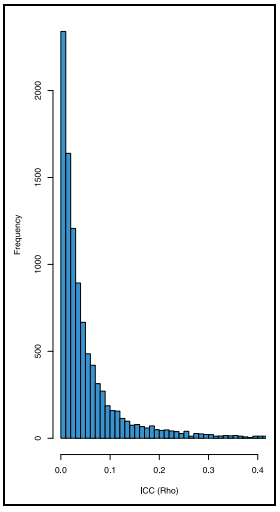

Figure 1. Histogram of 10,000 samples of the ICC, ρ.

Figure 2 shows that this method requires a larger sample size per cluster than the aforementioned methods.

We illustrate the effect of changing $\nu = (0, 0.1, \ldots, 1)$ on the assurance in the right-hand side plot of Figure 2. The red curve corresponds to $\nu = 0.5$, the top curve to $\nu = 0$ and the bottom curve to $\nu = 1$. As ν increases, the assurance decreases for a given cluster size. We see that the estimate of ν has a relatively strong effect on the assurance, and hence the required sample size. This implies that it needs to be estimated accurately, or its uncertainty should be taken into account in the assurance calculation.

Assurance based on the prior for ψ. In the second case, we obtain the sample size required using an assurance calculation, which averages over a prior distribution on σ and ν, as well as the ICC. Using the data from ICONS-I, we give σ and ν gamma marginal prior distributions, centred at their estimated values of 8.32 and 0.49, respectively. The standard deviations of the prior distributions are chosen to represent a belief that σ is very likely to be in the range $[5, 11]$ and ν is very likely to be in the range $[0.3, 0.7]$. Specifically, $\sigma \sim Gamma(a_\sigma, b_\sigma)$ and $\nu \sim Gamma(a_\nu, b_\nu)$, where $a_\cdot = m_\cdot^2/v_\cdot$ and $b_\cdot = m_\cdot/v_\cdot$, $m_\sigma = 8.32, v_\sigma = 1^2$ and $m_\nu = 0.49, v_\nu = 0.066^2$.

To incorporate the dependence between ρ and σ, we utilise the Gaussian copula with $\gamma = 0.43$. This is chosen to be consistent with the correlation between ρ and σ that would result from independent prior distributions on the between- and within-group standard deviations of $\sigma_b \sim Gamma(0.6, 0.5)$ and $\sigma_w \sim Gamma(83.5, 10.4)$, respectively. The hyperparameters of these two gamma prior distributions are chosen to provide the correct marginal means and variances for ρ and σ. To sample values of ρ and σ from their joint prior distribution, we repeat the following steps:

Table 1. Summary of sample sizes obtained for the ICONS CRT based on power and assurance calculations.

Method	Priors	Total number of clusters, J	Mean cluster size, $\bar{n}$	Total sample size, N
Power (classical approach)	NA	50	12	600
		40	18	720
		30	37	1110
Power (based on posterior median)	NA	50	9	450
		40	12	480
		30	19	570
Power (conservative values)	NA	50	23	1150
		40	57	2280
		30	>100	>3000.
Assurance	ρ	50	11	550
		40	17	680
		30	30	900
Assurance	$\psi = (\sigma, \rho, \nu)$	50	12	600
		40	18	720
		30	35	1050

ICONS: Identifying Continence OptioNs after Stroke; CRT: cluster randomised trial.

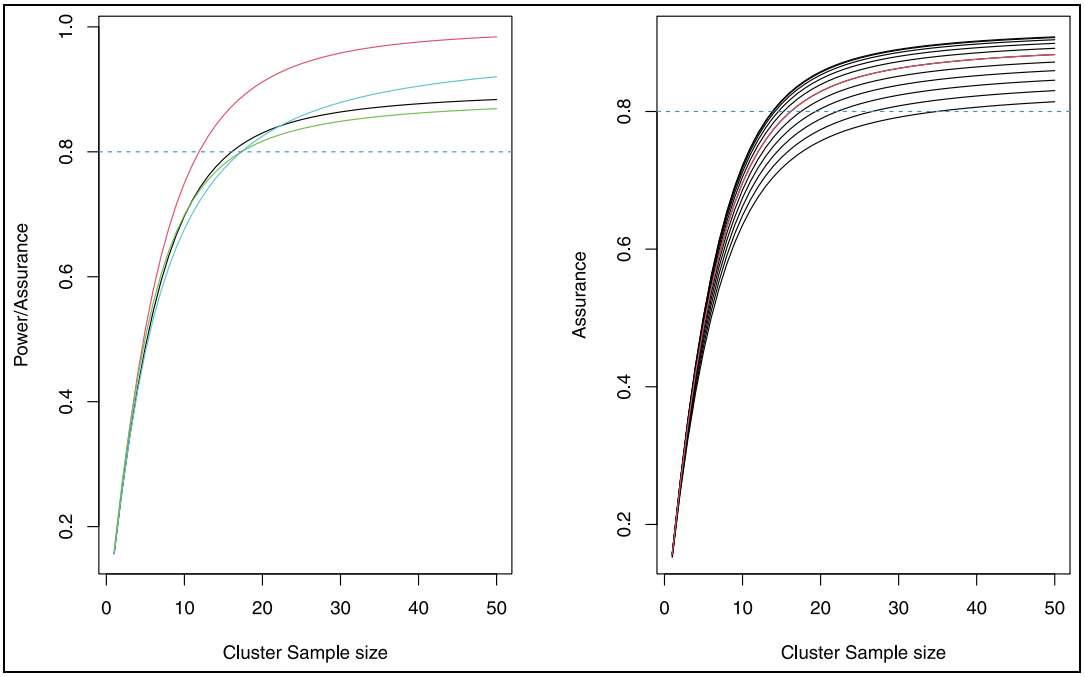

Figure 2. Power and assurance curves for the ICONS CRT (left). The power using the posterior median ICC is red, the power using the median ICC from the 34 ICC estimates is light blue, the assurance with a prior only on the ICC is black and the assurance with a prior on all of the nuisance parameters ψ is green. The effect of varying the coefficient of variation ν on the assurance (right). ν varies between 0 (top curve) and 1 (bottom curve), with the red line at $\nu = 0.5$. The horizontal line indicates the desired power/assurance. Each plot corresponds to $J = 40$ ($J/2 = 20$ per arm).

1. Sample (x_i, y_i), $i = 1, \ldots, S$ from $\mathrm{N}_2(0, R)$, where R is the prior correlation matrix with diagonal elements 1 and off-diagonal elements $\gamma = 0.44$.
2. Calculate (ρ_i, σ_i) as $\left(F_\rho^{-1}(\Phi(x)), F_\sigma^{-1}(\Phi(y))\right)$.

The quantile function F_σ^{-1} is that of the relevant normal distribution. The empirical quantile function F_ρ^{-1} is used for ρ, based on the 10,000 prior samples.

The resulting joint prior distribution for (σ, ρ) and marginal prior distribution for ν are illustrated in Figure 3. We see that the marginal prior for ρ remains as in Figure 1, but the samples are positively correlated with the values of σ.

The resulting average cluster sample sizes for an assurance of 80% are $\bar{n} = 18$ for $J = 40$ clusters ($J/2 = 20$ per arm) and $\bar{n} = 12$ for $J = 50$ clusters ($J/2 = 25$ per arm), requiring total sample sizes of

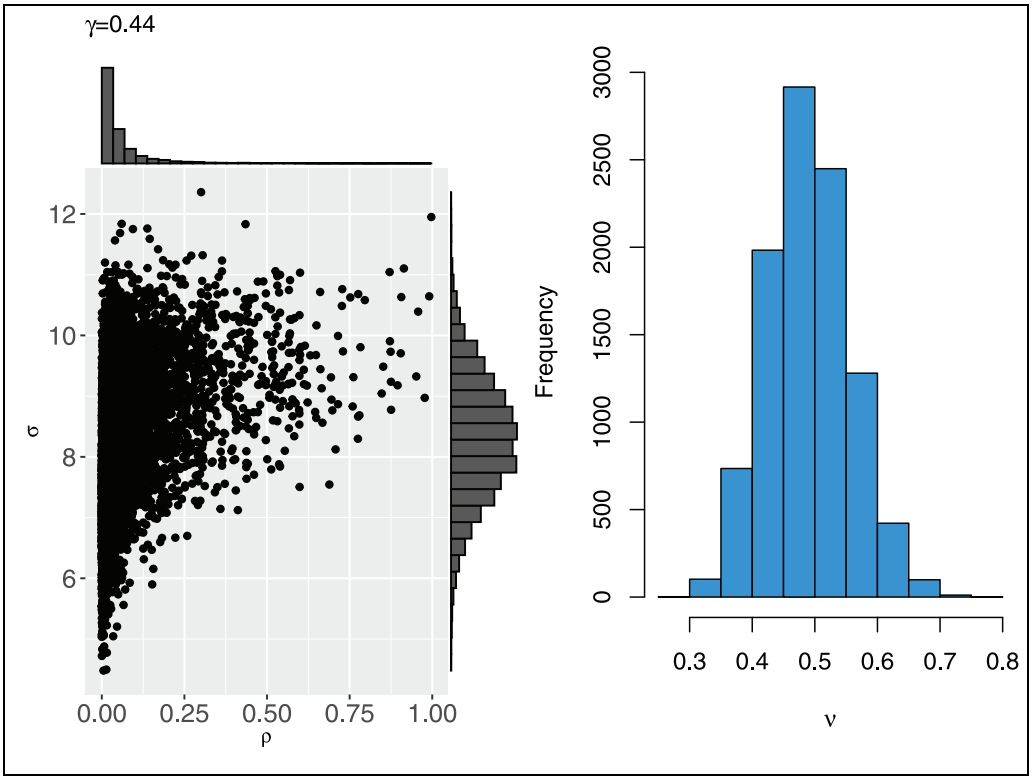

Figure 3. The joint prior distribution between ρ and σ (left) and the marginal prior distribution for ν (right), based on 10,000 samples.

N = 720 and N = 600, respectively. By incorporating uncertainty on σ and ν, as well as ρ, the sample size increases only slightly, as illustrated in the left-hand side plot of Figure 2 (green line). To achieve a target assurance of 80% (dashed horizontal line), the average sample size required per cluster increases from 17 to 18 when $J = 40$; an increase in total sample size of approximately 5%.

Table 1 summarises the sample sizes required to attain a target power/assurance of 80% for the various approaches applied to the ICONS trial. 'Classical approach' refers to the multiple-estimate method of taking the median of the ICC estimates without taking the relevance of the different studies into account. Relative to the classical approach that is often used in practice, the total sample size required when using the assurance-based method remains the same while incorporating uncertainty on all three parameters. 'Conservative values' refers to using conservative values for each of (σ, ρ, ν), which we take as the upper quartile values from each of their marginal design priors. In this case, we obtain sample sizes that are more than double those attained via any other approach.

We include the solutions for a smaller number of clusters, $J = 30$. Apart from power using the posterior median, which uses a relatively small ICC value, we see assurance resulting in the smallest sample sizes for this case.

Sensitivity analysis for the ICC prior

In the above, we consider the ICC prior distribution based on all eight reviewers and all 16 relevant studies. In this section, we investigate the sensitivity of the assurance-based sample size (with priors on ψ) to varying assumptions on the reviewers and relevant studies, and compare this to the sensitivity of the sample sizes from power calculations (using the posterior median ICC).

To recognise uncertainty in the individual reviewers' responses, and in how these responses were pooled, the mathematical aggregation was refitted with alternative reviewer importance weights: equal weights of 0.125 for all reviewers and using a rank sum approach.[23] For the rank sum approach, we use Cronbach's alpha score and assign ranks to each reviewer according to this score. In addition, we rerun the Bayesian hierarchical model for only the top 4 (25%), 8 (50%) and 12 (75%) most relevant studies. We refer to the five variations of the original ICC prior distribution as: equal weights, differentiated weights, top 4, top 8 and top 12.

The differentiated weights prior (red) and equal weights prior (green) are provided alongside the original prior (black) in the left-hand side plot of Figure 4. The top 4 prior (red), top 8 prior (green) and top 12 prior (blue) are given alongside the original prior (black) in the right-hand side plot of Figure 4. In both plots, the prior medians are given by vertical dashed lines.

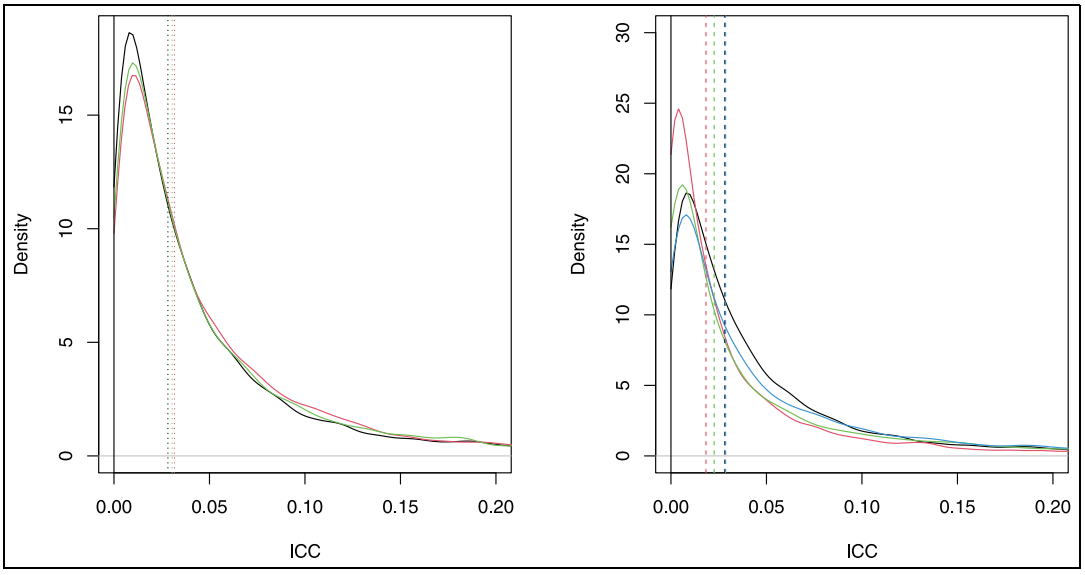

Figure 4. Left: The densities of the differentiated weights (red), equal weights (green) and original ICC prior (black). Right: The densities of the top 4 (red), top 8 (green), top 12 (blue) and original ICC prior (black). The prior medians are represented by vertical dashed lines.

Table 2. The average sample size per cluster $\bar{n}$ and the total sample size N required for the ICONS CRT using assurance (with priors on ψ) and power based on the original ICC estimate/prior and five alternative estimates/priors when $J = 50$ and $J = 40$.

ICC Estimate/Prior	J=50				$J = 40$			
	Assurance		Power		Assurance		Power	
	$\bar{n}$	N	$\bar{n}$	N	$\bar{n}$	N	$\bar{n}$	N
Original	12	600	9	450	18	720	12	480
Differentiated weights	13	650	10	500	20	800	13	520
Equal weights	13	650	9	450	19	760	13	520
Top 4	12	600	8	400	18	720	11	440
Top 8	14	700	9	450	23	920	12	480
Top 12	15	750	9	450	24	960	12	480

The power is based on the posterior median of the ICC.
ICONS: Identifying Continence OptioNs after Stroke; CRT: cluster randomised trial; ICC: intra-cluster correlation coefficient.

We see that the ICC prior remains similar to the original prior whether differentiated weights or equal weights are used, although both alternative weightings assign more probability to the ICC taking larger values. There is a larger change when using the top 4, top 8 or top 12 studies. In each case, the alternative prior is more diffuse than the original prior. Relatively large changes in the prior can cause only small changes in the prior median (e.g. the original prior compared to the top 12 prior). The effects of the alternative ICC priors on the sample sizes are shown in Table 2.

We see smaller changes in sample sizes for $J = 50$ than $J = 40$ using assurance. Overall, we observe larger changes in sample size using assurance than power based on the prior median of the ICC. This illustrates the risk with using just the median; it takes no account of the prior probability that the ICC could be relatively large, so has the potential to systematically underestimate the required sample size. In contrast, the assurance-based sample size is sensitive to the entire ICC prior distribution, particularly the upper tail.

To illustrate this point, compare the original ICC prior (black) to the top 12 prior (blue) in the right-hand side of Figure 4. They have substantially different priors, resulting in large differences in sample sizes required under assurance (600 versus 750 when $J = 50$, respectively). However, their prior medians are almost identical, resulting in identical sample size requirements under power (450 when $J = 50$).

In the Appendix, we further evaluate the properties of the hybrid approach compared to power via a simulation study.

Conclusion

A standard sample size calculation requires pre-specification of parameters that are unknown at the design stage of a trial. Unique to sample size calculations for typical CRTs is the ICC, which requires robust estimation to avoid over- or under-powering the trial. Unnecessarily high ICC values, for example, lead to inefficient trials, increasing the number of clusters and/or participants and overall trial costs. In practice, parameter uncertainty is typically not considered, which can be problematic given the sensitivity of the sample size to small differences in the ICC.

This article proposes an alternative approach to sample size determination for CRTs using the Bayesian concept of assurance to incorporate parameter uncertainty into the design. The advantage of this approach is that it yields designs that provide adequate power across the likely range of parameter values and is, therefore, more robust to parameter misspecification. This is particularly important when there is difficulty obtaining a reliable ICC estimate, as in the ICONS post-stroke incontinence CRT used to motivate this work. Another approach in this context is to perform an interim analysis for sample size re-estimation. The approach proposed in this article could be used in combination with sample size re-estimation to provide further robustness to the design of CRTs.

We assign prior distributions to the ICC, overall standard deviation and coefficient of variation of the cluster size, while setting the treatment effect equal to the MCID in line with standard practice. We consider a joint prior for the ICC and standard deviation to model the dependency between these parameters. In the motivating case study, we use the entire ICC prior distribution elicited from expert opinion and data from previous studies to inform the sample size. Further work could consider using a commensurate prior to synthesise multiple sources of pre-trial information on the ICC, as in literature.[33]

Sensitivity analyses of the assurance-based sample size to different ICC priors showed that different behaviour of the prior, particularly in the upper tail, can have quite a strong effect on the required sample size. Using a point estimate from this prior, for example the median, can miss this overall behaviour and result in sample sizes which are systematically too small, based on current knowledge about the ICC. Additional sensitivity analyses conducted on the overall standard deviation showed that the greater the uncertainty expressed in the prior, the more robust the assurance-based sample size is (see Appendix).

Uncertainty in the treatment effect can also be incorporated into the assurance calculation in a similar way. This may be appropriate for non-inferiority trials, for example, where the non-inferiority margin is fixed in advance and the treatment difference can be considered a nuisance parameter.

In line with regulatory requirements, we have maintained a frequentist analysis to present a hybrid framework. Further work could consider a fully Bayesian approach by using assurance when the success criterion is based on the posterior distribution of the treatment effect.[13]

The hybrid approach presented in this article can be applied to avoid incorrectly powered studies resulting from ill-estimated model parameters, to mitigate the impact of uncertainty in the ICC and other nuisance parameters, and to incorporate expert opinion or historical data when designing a CRT. The approach proposed has been outlined in the case that the number of clusters is fixed and we aim to determine the total sample size for the trial. The approach would also allow the reverse process – to calculate the necessary number of clusters given a fixed total sample size.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research presented in this paper used data from ICONS-I trial funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research scheme (grant number: RP-PG-0707-10059).

ORCID iDs

S. Faye Williamson https://orcid.org/0000-0002-1800-6625
Svetlana V. Tishkovskaya https://orcid.org/0000-0003-3087-6380
Kevin J. Wilson https://orcid.org/0000-0001-9450-862X

Supplemental material

Supplemental material for this article is available online.

References

1. Eldridge S and Kerry S. *A practical guide to cluster randomised trials in health services research*. Hoboken, NJ: John Wiley & Sons, 2012.
2. Rutterford C, Copas A and Eldridge S. Methods for sample size determination in cluster randomized trials. *Int J Epidemiol* 2015; 44(3): 1051–1067.
3. Gao F, Earnest A, Matchar DB, et al. Sample size calculations for the design of cluster randomized trials: a summary of methodology. *Contemp Clin Trials* 2015; 42: 41–50.

4. Eldridge SM, Costelloe CE, Kahan BC, et al. How big should the pilot study for my cluster randomised trial be? *Stat Methods Med Res* 2016; 25(3): 1039–1056.
5. Browne RH. On the use of a pilot sample for sample size determination. *Stat Med* 1995; 14: 1933–1940. DOI: 10.1002/sim.4780141709.
6. Moerbeek M and Teerenstra S. *Power analysis of trials with multilevel data*. Boca Raton, FL: CRC Press, 2015.
7. Korevaar E, Kasza J, Taljaard M, et al. Intra-cluster correlations from the CLustered OUtcome Dataset bank to inform the design of longitudinal cluster trials. *Clin Trials* 2021; 18(5): 529–540.
8. Lewis J and Julious SA. Sample sizes for cluster-randomised trials with continuous outcomes: accounting for uncertainty in a single intra-cluster correlation estimate. *Stat Methods Med Res* 2021; 30(11): 2459–2470.
9. Ukoumunne OC. A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Stat Med* 2002; 21(24): 3757–3774.
10. Lake S, Kammann E, Klar N, et al. Sample size re-estimation in cluster randomization trials. *Stat Med* 2002; 21: 1337–1350.
11. van Schie S and Moerbeek M. Re-estimating sample size in cluster randomised trials with active recruitment within clusters. *Stat Med* 2014; 33: 3253–3268.
12. van Breukelen GJ and Candel MJ. Efficient design of cluster randomized and multicentre trials with unknown intraclass correlation. *Stat Methods Med Res* 2015; 24(5): 540–556.
13. Spiegelhalter DJ. Bayesian methods for cluster randomized trials with continuous responses. *Stat Med* 2001; 20(3): 435–452.
14. Turner RM, Toby Prevost A and Thompson SG. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Stat Med* 2004; 23(8): 1195–1214.
15. Turner RM, Thompson SG and Spiegelhalter DJ. Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clin Trials* 2005; 2(2): 108–118.
16. Sarkodie SK, Wason JM and 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 intracluster correlation. *Clin Trials* 2023; 20(1): 59–70.
17. Kunzmann K, Grayling MJ, Lee KM, et al. A review of Bayesian perspectives on sample size derivation for confirmatory trials. *Am Stat* 2021; 75(4): 424–432.
18. Grieve A. *Hybrid frequentist/Bayesian power and Bayeisian power in planning clinical trials*. Boca Raton, FL: CRC Press, 2022.
19. Chen DG and Ho S. From statistical power to statistical assurance: it's time for a paradigm change in clinical trial design. *Commun Stat Simul Comput* 2017; 46(10): 7957–7971.
20. Ciarleglio MM, Arendt CD and Peduzzi PN. Selection of the effect size for sample size determination for a continuous response in a superiority clinical trial using a hybrid classical and Bayesian procedure. *Clin Trials* 2016; 13(3): 275–285.
21. Eldridge SM, Ashby D and Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol* 2006; 35(5): 1292–1300.
22. Zhan D, Xu L, Ouyang Y, et al. Methods for dealing with unequal cluster sizes in cluster randomized trials: a scoping review. *PLoS ONE* 2021; 16(7): e0255389. DOI: 10.1371/journal.pone.0255389.
23. Tishkovskaya SV, Sutton CJ, Thomas LH, et al. Determining the sample size for a cluster-randomised trial: Bayesian hierarchical modelling of the intracluster correlation coefficient. *Clin Trials* 2023; 20(3): 293–306. DOI: 10.1177/17407745231164569.
24. Jones BG, Streeter AJ, Baker A, et al. Bayesian statistics in the design and analysis of cluster randomised controlled trials and their reporting quality: a methodological systematic review. *Syst Rev* 2021; 10(91): 1–14.
25. Kerry SM and Bland JM. The intracluster correlation coefficient in cluster randomisation. *BMJ* 1998; 316(7142): 1455–1460. DOI: 10.1136/bmj.316.7142.1455. https://www.bmj.com/content/316/7142/1455.1
26. O'Hagan A and Stevens JW. Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Med Decis Making* 2001; 21(3): 219–230.
27. Nelson R. *An introduction to copulas*. New York: Springer-Verlag, 2006.
28. Thomas LH, French B, Sutton CJ, et al; on behalf of the ICONS project team and the ICONS patient, public and carer involvement groups. Identifying Continence OptioNs after Stroke (ICONS): an evidence synthesis, case study and exploratory cluster randomised controlled trial of the introduction of a systematic voiding programme for patients with urinary incontinence after stroke in secondary care. *Programme Grants Appl Res* 2015; 3(1): 1–644.
29. Avery K, Donovan J, Peters TJ, et al. ICIQ: a brief and robust measure for evaluating the symptoms and impact of urinary incontinence. *Neurourol Urodyn* 2004; 23(4): 322–330.
30. Thomas LH, Watkins CL, Sutton CJ, et al. Identifying continence options after stroke (ICONS): a cluster randomised controlled feasibility trial. *Trials* 2014; 15(509): 1–15.
31. O'Hagan A. Expert knowledge elicitation: subjective but scientific. *Am Stat* 2019; 73(suppl 1): 69–81. DOI: 10.1080/00031305.2018.1518265.
32. O'Hagan A, Buck CE, Daneshkhah A, et al. *Uncertain judgements: eliciting experts' probabilities*. Hoboken, NJ: John Wiley & Sons, 2006.
33. Zheng H, Jaki T and Wason JM. Bayesian sample size determination using commensurate priors to leverage preexperimental data. *Biometrics* 2023; 79(2): 669–683. DOI: 10.1111/biom.13649.