

Determining the sample size for a cluster-randomised trial using knowledge elicitation: Bayesian hierarchical modelling of the intraclass correlation coefficient

Clinical Trials
2023, Vol. 20(3) 293–306
© The Author(s) 2023



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17407745231164569
journals.sagepub.com/home/ctj



Svetlana V Tishkovskaya¹ , Chris J Sutton², Lois H Thomas³
and Caroline L Watkins¹

Abstract

Background: The intraclass correlation coefficient is a key input parameter for sample size determination in cluster-randomised trials. Sample size is very sensitive to small differences in the intraclass correlation coefficient, so it is vital to have a robust intraclass correlation coefficient estimate. This is often problematic because either a relevant intraclass correlation coefficient estimate is not available or the available estimate is imprecise due to being based on small-scale studies with low numbers of clusters. Misspecification may lead to an underpowered or inefficiently large and potentially unethical trial.

Methods: We apply a Bayesian approach to produce an intraclass correlation coefficient estimate and hence propose sample size for a planned cluster-randomised trial of the effectiveness of a systematic voiding programme for post-stroke incontinence. A Bayesian hierarchical model is used to combine intraclass correlation coefficient estimates from other relevant trials making use of the wealth of intraclass correlation coefficient information available in published research. We employ knowledge elicitation process to assess the relevance of each intraclass correlation coefficient estimate to the planned trial setting. The team of expert reviewers assigned relevance weights to each study, and each outcome within the study, hence informing parameters of Bayesian modelling. To measure the performance of experts, agreement and reliability methods were applied.

Results: The 34 intraclass correlation coefficient estimates extracted from 16 previously published trials were combined in the Bayesian hierarchical model using aggregated relevance weights elicited from the experts. The intraclass correlation coefficients available from external sources were used to construct a posterior distribution of the targeted intraclass correlation coefficient which was summarised as a posterior median with a 95% credible interval informing researchers about the range of plausible sample size values. The estimated intraclass correlation coefficient determined a sample size of between 450 (25 clusters) and 480 (20 clusters), compared to 500–600 from a classical approach. The use of quantiles, and other parameters, from the estimated posterior distribution is illustrated and the impact on sample size described.

Conclusion: Accounting for uncertainty in an unknown intraclass correlation coefficient, trials can be designed with a more robust sample size. The approach presented provides the possibility of incorporating intraclass correlation coefficients from various cluster-randomised trial settings which can differ from the planned study, with the difference being accounted for in the modelling. By using expert knowledge to elicit relevance weights and synthesising the externally available intraclass correlation coefficient estimates, information is used more efficiently than in a classical approach, where the intraclass correlation coefficient estimates tend to be less robust and overly conservative. The intraclass

¹Lancashire Clinical Trials Unit, Faculty of Health and Care, University of Central Lancashire, Preston, UK

²Centre for Biostatistics, Division of Population Health, Health Services Research & Primary Care, School of Health Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK

³Faculty of Allied Health and Wellbeing, University of Central Lancashire, Preston, UK

Corresponding author:

Svetlana V Tishkovskaya, Lancashire Clinical Trials Unit, Faculty of Health and Care, University of Central Lancashire, Preston PR1 2HE, UK.
Email: s.tishkovskaya@gmail.com

correlation coefficient estimate constructed is likely to produce a smaller sample size on average than the conventional strategy of choosing a conservative intracluster correlation coefficient estimate. This may therefore result in substantial time and resources savings.

Keywords

Bayesian hierarchical model, cluster-randomised trial, intracluster correlation coefficient, knowledge elicitation, post-stroke incontinence, sample size determination

Background

Health interventions are often evaluated using cluster-randomised trials where clusters of individuals are randomly allocated to trial arms.¹ Typically, subjects within the same cluster have similar outcomes, not just because they are similar, but as they may share unmeasured cluster-level effects on outcome. The correlation of outcome measurements within a cluster, called the intracluster correlation coefficient (ICC or ρ),^{2,3} must be utilised in trial design and analysis. The ICC is a most commonly used measure of the similarity of clustered data.³ It compares within-group variance σ_w^2 with between-group variance σ_b^2 and for a continuous outcome is defined as

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

Similarities among individuals within clusters cause a loss of statistical power to detect a between-group difference. To maintain power, the target sample size must be inflated to allow for clustering: the sample size of an individually randomised trial is multiplied by the design effect $DEff = 1 + (m - 1) \times \rho$, where m is the average cluster size. If m is large, the inflation factor may substantially increase the sample size, even when ρ is small (as it often is in cluster-randomised trials).

As sample size is highly sensitive to the ICC, robust ICC estimates are required. Currently, obtaining a robust estimate of the ICC is a challenge in planning a cluster-randomised trial.^{1,3,4} An underestimated ICC will produce an underpowered study, whereas an inflated ICC will require more clusters and/or participants than necessary, leading to wasted resources and a potentially unethical trial.

The usual, and easiest, way of obtaining an ICC is using a single estimate from an existing source, such as a published similar study, published lists or databases of ICC estimates, or a pilot for the proposed study. Use of a single ICC value is not robust (often resulting in a trial being underpowered or overpowered causing wasted resources) and a sufficiently relevant value is seldom available.^{1,4-6}

An advanced approach might use multiple-estimate methods, combining different ICC estimates for the

particular outcome and cluster type. However, finding multiple sources highly relevant to the target outcome, population and cluster type is quite difficult. It is likely that available studies with ICCs will have only partial relevance to the planned trial. Furthermore, there is no recommended method of combining ICCs and simple approaches, such as using their mean, median or maximum, do not account for the specific characteristics of the studies, the degree of relevance or the uncertainty in each ICC estimate.

Imprecision in the ICC is usually expressed in terms of variance of the estimate, with different ways of calculating ICC variance proposed.¹ Power calculations for cluster-randomised trials typically use ICC estimates without taking their uncertainty into account.⁷ Accounting for imprecision in the ICC would improve the estimation of the required sample size and protect against trials having inadequate power due to a higher than expected ICC.

In this study, we adopted and extended the method of combining ICC values in the Bayesian framework⁸ suggested by Turner et al.⁹ who applied the Bayesian hierarchical approach to combine multiple relevant ICC estimates in a single model. The method can be applied to ICCs of varying relevance, thus relaxing the limitations of other approaches which require a high level of similarity across studies. Using this method, the input of each study is differentiated according to its degree of relevance. Relevance to target setting is expressed as weights for each study and for each study outcome. Turner et al.⁹ assigned categorised weights to studies and outcomes according to their relevance. Thus, the method allows incorporation of all available information, including less relevant data sources, into the model, allowing the latter to have less influence when combining ICC estimates.

We extended the approach suggested in Turner et al.⁹ by incorporating expert knowledge elicitation process into the Bayesian modelling to assist in assessing the relevance weights and deciding how much strength can be borrowed from each study. The motivation for developing the methodology presented here was a proposed cluster-randomised trial of the effectiveness of a systematic voiding programme for people on National Health Service (NHS) stroke units with stroke and urinary incontinence. The primary outcome measure was the International

Consultation on Incontinence Questionnaire–Urinary Incontinence Short Form (ICIQ–UI–SF) score, assessing incontinence symptom severity.¹⁰

A robust ICC estimate was not available to inform the proposed trial, although a feasibility trial, Identifying Continence OptioNs after Stroke (ICONS),¹¹ had been conducted. ICONS produced an imprecise ICC estimate due to having only 12 clusters with 413 participants and could therefore not be used as a reliable single source for the sample size calculation.

We applied a Bayesian approach to modelling and estimation of the ICC for determination of the sample size using multiple ICC estimates from previously published studies. The modelling automatically accounted for uncertainty in the synthesised ICCs and produced an informative posterior ICC distribution which we used to determine the sample size. We describe here the process used and provide a how-to guide on the proposed framework to assist researchers in exploring the utility of this approach for their own trials.

Methods

Model

To specify the model for the ICC, we assume, similar to Turner et al.,⁹ that each estimate $\hat{\rho}_m$ ($m = 1, \dots, s$) follows a Normal distribution around its true value ρ_m

$$\hat{\rho}_m \sim N(\rho_m, \text{Var}(\hat{\rho}_m))$$

For the distribution variance, we make use of Swiger et al.'s¹² formula for estimating the asymptotic variance of the ICC estimate

$$\begin{aligned} \text{Var}[\hat{\rho}_m] &= V(\rho_m, N_m, k_m) \\ &= \frac{2(N_m - 1)(1 - \hat{\rho}_m)^2 \left(1 + \left(\frac{N_m}{k_m} - 1\right) \hat{\rho}_m\right)^2}{\left(\frac{N_m}{k_m}\right)^2 (N_m - k_m)(k_m - 1)} \end{aligned}$$

where N_m is the total number of participants in the trial, and k_m is the number of clusters. Swiger's formula requires minimal information for calculating the ICC variance estimate and has been used in a number of methodological papers.¹ Turner et al.⁹ have also followed this method. Moreover, Swiger's formula has been compared with other methods for calculating variance, and it was concluded that all approaches lead to similar practical conclusions.⁷

Eliciting information for Bayesian hierarchical ICC modelling

Although a single robust ICC estimate was not available for the planned cluster-randomised trial, there were a number of studies available with ICC estimates that had varying degrees of relevance, both to the planned

trial population and its primary outcome. To identify a set of ICC estimates relevant to our setting, we conducted a structured literature search. To identify the studies, the trial team utilised the search strategy implemented in a paper by Sutton et al.¹³ reviewing the use and reporting of cluster-randomised trials in stroke. We conducted an electronic search of titles and abstracts in the PUBMED database, and titles, abstracts or keywords in the CENTRAL (Cochrane Trials) database to June 2014 using the search terms 'trial' AND 'stroke' AND (cluster-randomised OR group-randomised OR cluster randomised OR group randomised), to identify full papers reporting cluster-randomised trials of stroke-related interventions. Studies were included only if they reported estimated ICCs. Through the search, 16 studies relevant to the planned trial were identified, including the ICONS feasibility trial and 8 trials from the review by Sutton et al.¹³ Most of the selected trials evaluated interventions designed to reduce stroke incidence or improve stroke care; two of the trials were concerned with incontinence problems. A total of 34 ICC estimates were extracted, with some studies providing multiple ICCs. A summary of the data extracted from the selected studies is presented by Table 1. Supplemental Table 2 shows characteristics of the studies and ICCs.

To combine all 34 ICCs, we used the following model suggested by Turner et al.⁹

$$\begin{aligned} \hat{\rho}_{ml} &\sim N(\rho_{ml}, V(\rho_{ml}, N_{ml}, k_{ml})) \\ \text{logit}(\rho_{ml}) &\sim N\left(\mu_m, \frac{\sigma_w^2}{w_{ml}}\right) \\ \mu_m &\sim N\left(\mu, \frac{\sigma_b^2}{w_m}\right) \end{aligned}$$

here, ρ_{ml} is ICC for l th outcome within the m th study, N_{ml} and k_{ml} are the corresponding number of participants and the number of clusters, σ_w^2 and σ_b^2 are the within- and between-study variances, w_m and w_{ml} are the study and outcome weights, $m = 1, \dots, r$; $l = s_1, \dots, s_m$. In this model, the exchangeability is implied for both between and within studies so that the parameters may be considered as independently drawn from a common distribution.⁸ In the context of our model, exchangeability within each of the separate studies means that the estimates $\hat{\rho}_{ml}$ are distributed around an underlying value ρ_{ml} , with the ρ_{ml} assumed exchangeable within studies and Normally distributed on the logit-transformed scale around a study-specific mean μ_m . Logit transformation accounts for the permissible range of values [0, 1] for the ρ_{ml} . Exchangeability between studies means that the μ_m are assumed to be independently drawn from a common Normal distribution.

Similar to Turner et al.,⁹ we adjust for the varying relevance of the estimates by assigning weights to each

Table 1. Characteristics of studies with available relevant ICC estimates included in Bayesian modelling.

Study	Source	Intervention	Study population	Outcome	ICC estimate	Number of patients	Number of clusters
1	Thomas et al. ¹⁴	Systematic voiding programme.	Stroke patients with incontinence (UK).	Absence of incontinence at 12 weeks post-stroke.	0.00	413	12
2	Tannenbaum et al. ¹⁵	Three experimental continence interventions: (1) continence education; (2) evidence-based self-management; (3) combined continence education and self-management.	Women aged 60 years and older with untreated incontinence (UK).	Patient's global impression of improvement in continence questionnaire (PGI-I) measured at 3 months post-intervention.	0.05	259	71
3	Sackley et al. ¹⁶	Staff education on continence care and mobility care and mobility training.	UK care home residents.	Rivermead Mobility Index at baseline and 6 weeks post-intervention.	0.37	34	6
4	Sackley et al. ¹⁷	Occupational therapy provided to individuals and carer education.	UK care homes residents with moderate to severe stroke-related disability (Barthel Index (BI) score 4–15) except those with acute illness and those admitted for end-of-life care.	BI score at baseline. BI change to 3 months. BI change to 6 months. Global poor outcome at 3 months. Global poor outcome at 6 months.	0.26 0.18 0.2 0.14 0.09	173	12
5	Weir et al. ¹⁸	Computer-based decision support system to aid selection of long-term antithrombotic therapy.	UK hospital in-patients or out-patients with a clinical diagnosis of acute ischaemic stroke or TIA; first investigation of an event occurring within preceding 4 months.	Change in relative risk of ischaemic and haemorrhagic vascular events relative to the option of 'no antiplatelet or anticoagulant therapy'.	0.15	1952	16
6	De Luca et al. ¹⁹	The intervention group staff (physicians, nurses and drivers) training on and delivery of evidence-based prehospital emergency clinical pathway based on experiential learning approach. The training was focused on teaching the personnel to identify stroke symptoms.	People living in the community aged <80 years (Italy, acute care/community).	The proportion of eligible acute stroke patients correctly referred to stroke unit.	0.05	4895	20

(continued)

Table 1. continued

Study	Source	Intervention	Study population	Outcome	ICC estimate	Number of patients	Number of clusters
7	Dirks et al. ²⁰	Intervention to increase thrombolysis rates by creating local stroke teams, identifying barriers to thrombolysis delivery, setting goals and planned actions, and updating acute stroke guidelines.	Patients ≥ 18 years with acute stroke who were admitted to the hospital within 24 h from onset of symptoms (Netherlands, acute care/community).	Treatment with rtPA (recombinant tissue Plasminogen Activator).	0.0154	5515	12
8	Johnston et al. ²¹	Quality improvement in ischaemic stroke discharge orders comprising statin prescription; antihypertensive medication for those with hypertension; warfarin for all patients with atrial fibrillation (AF), except those with contraindication.	At least 40 years old, were the Kaiser Permanente Medical Care Plan (KPMCP) members with pharmacy benefits, and had been discharged alive to home or to a facility other than hospice (USA hospitals).	Composite binary variable comprising optimal treatment via all of: (1) documentation of filled statin prescription 6 m post-discharge; (2) achievement of controlled blood pressure 4–8 m post-discharge; (3) for those with AF, either documentation of a filled prescription for warfarin or an International Normalised Ratio blood test 6 m post-discharge or a contraindication to warfarin.	0.0038	3361	12
9	Jones et al. ²²	All nurses and health-care assistants on the participating stroke intervention units received a group teaching package to improve their understanding and clinical practice in patient positioning.	Patients on stroke rehabilitation units (UK hospitals): with stroke, dependent on another person to position limbs, with inability to move from sitting to standing without assistance.	Rivermead Mobility Index at 6 months post-stroke.	0.00	120	10
10	Lakshminarayan et al. ²³	(1) Audit and written feedback of baseline performance; (2) analysis of structural and knowledge barriers to stroke care identified by provider questionnaires; (3) use of clinical opinion leaders to deliver customised feedback to care providers; (4) use of hospital management leaders to overcome identified barriers to stroke care.	Stroke patients aged 30–84 years admitted through emergency room (US hospitals).	Three outcomes with associated ICCs, each is related to the provision of 3 or 4 indicators of quality of care:		2305	19

(continued)

Table 1. continued

Study	Source	Intervention	Study population	Outcome	ICC estimate	Number of patients	Number of clusters
11	McAlister et al. ²⁴	Educational lecture to patients with nonvalvular AF on reducing stroke risk, with self-administered booklet and individualised audiotape decision aid tailored to their personal stroke risk profile.	Adult patients with nonvalvular AF not living in institutions (Canada, Primary Care Practices).	acute care indicators. inpatient care indicators. discharge indicators. Change in proportion of patients taking antithrombotic therapy appropriate to their stroke risk 3 months post-intervention.	0.005 0.004 0.0007 0.02	434	102
12	Forster et al. ²⁵	Structured training programme for caregivers (the London Stroke Carers Training Course).	Patients at UK stroke units with a diagnosis of stroke, likely to return home with residual disability and with a caregiver providing support.	Self-reported extended activities of daily living at 6 months measured with the Nottingham Extended Activities of Daily Living scale.	0.027	928	36
13	Taylor et al. ²⁶	Structured goal elicitation using the Canadian Occupational Performance Measure.	Stroke patients admitted to inpatient rehabilitation services (New Zealand) with 'sufficient' cognition for goal setting and completing outcome assessment.	Quality of life at 12 weeks measured using the following tools: Schedule for Individualised Quality of Life (SEIQOL-DW).	0.40	41	4
14	Middleton et al. ²⁷	Treatment protocols to manage fever, hyperglycaemia and swallowing dysfunction with multi-disciplinary team building workshops to address implementation barriers.	Patients aged 18 years or older, who had a diagnosis of ischaemic stroke or intracerebral haemorrhage, and presented within 48 h of onset of symptoms to a participating Acute Stroke Unit (ASU), Australia.	The Medical Outcomes Study 36-item Short Form Health Survey (SF-36), Physical Component Summary (PCS) score. Functional Independence Measure. The Medical Outcomes Study 36-item Short Form Health Survey (SF-36), Mental Component Summary (MCS) score. Death and dependency 90 days after hospital admission. Functional dependency BI ≥ 95 , 90 days after hospital admission. Functional dependency BI ≥ 60 , 90 days after hospital admission. SF-36 PCS score, 90 days after hospital admission.	0.24 0.21 0.25 0.018 0.015 0.009 0.026	1696	19

(continued)

Table 1. continued

Study	Source	Intervention	Study population	Outcome	ICC estimate	Number of patients	Number of clusters
15	Power et al. ²⁸	Stroke 90/10, a quality improvement collaborative based on the Breakthrough Series model.	Patients admitted to stroke units at NHS hospital Trusts in the Northwest of England.	SF-36 MCS score, 90 days after hospital admission. Mean temp within 72 h in ASU. At least one temperature $\geq 37.5^{\circ}\text{C}$ in first 72 h. Mean glucose during first 72 h in ASU. Swallowing screening within 24 h of admission to ASU. Compliance with two evidence-based bundles of care: early hours and rehabilitation.	0.011 0.084 0.009 0.056 0.156	6592	24
16	Dregan et al. ²⁹	Remotely installed electronic decision support tools to promote intensive secondary prevention.	Patients ever diagnosed with acute stroke (Family Practices, UK).	Early hours bundle. Rehabilitation bundle. Systolic blood pressure.	0.066 0.197 0.032	11,391	106

ICC: intracluster correlation coefficient; TIA: transient ischemic attack; NHS: National Health Service.

study and to each outcome with an ICC. Study weights represent the degree of relevance to the study population and intervention; outcome weights reflect the degree of relevance to the planned trial's primary outcome. By assigning a lower weight, we decrease the influence that the corresponding estimate has on the constructed posterior ICC distribution.

In practice, however, it is unclear how to assign these weights. Guidance⁹ suggests that relevance weights should express a proportion of the total nonsampling variance that is not due to bias. This advice is difficult to implement in practice, and the authors emphasise that these weights are likely to be subjective.

To minimise subjectivity in defining weights, we considered using a knowledge elicitation process to derive study and outcome weights from expert prior opinion or beliefs. We performed a ranking exercise where eight members of the trial team, with relevant expertise, assigned weights for each trial and for each outcome. The expert team included investigators with the following expertise: health-service and stroke researchers, nurses, methodologists and a stroke-survivor (member of ICONS Patient and Public Involvement group).

To aid elicitation, we produced an information leaflet (see Supplemental Material) explaining the methodology. An Excel spreadsheet (Supplemental Table 1) was developed summarising the 16 relevant studies. It was carefully designed to focus the expert elicitation process on the key parameters required for comparing studies and assigning weights, and to avoid creating a burdensome exercise for the reviewers.

The eight experts were tasked with assigning study and outcome weights. The ranking exercise was set up in two stages. The first stage comprised a group exercise and was conducted face-to-face, in line with methodological developments suggesting that elicitation within group meetings is an efficient way of collecting expert opinion.³⁰ Reviewers were provided with an explanation of the method. Each expert was given an Excel spreadsheet with a summary of the extracted studies. The group was given an opportunity to work together on reviewing a few studies, with some interaction between experts. With the assistance of a project statistician facilitating the meeting, the experts discussed their initial estimates with the others to clarify the ranking exercise and establish a shared understanding of how to measure a relevance of outcome and study to the target setting. It has been shown that this first discussion round with interaction between experts is beneficial for the knowledge elicitation process as it generates sharing of knowledge and leads to improvements in response accuracy.³¹ At the second stage, the reviewers were asked to revise their ratings in light of the first round, rank all studies in their own time and return the individual judgements to a facilitator by email.

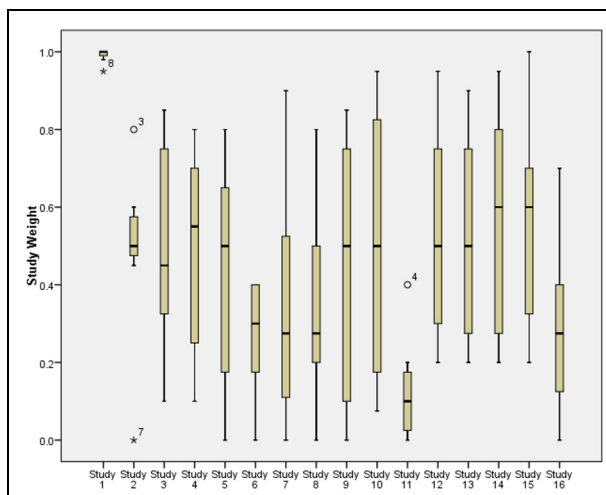


Figure 1. Boxplots showing the spread of reviewer responses about each study's weight. Study numbers are in same order as in Table 1. The narrowest boxplot relates to the ICONS feasibility trial.

The implemented knowledge elicitation process is a version of the probabilistic Delphi protocol³² recommended by the European Food Safety Authority³³ for work with multiple experts. It requires aggregation across the experts' final estimates; this is described in the following section.

Synthesising expert opinion

There was considerable diversity across the experts in the weights provided, with some reviewers having a tendency to assign much higher or much lower weights than others (Figure 1). The reviewers tended to be in better agreement for studies of lower relevance to the target study: the spread of responses was narrower for these studies when compared to the studies of higher relevance.

Before putting weights into the model, we assessed the reviewers' performance in a series of steps described in the Supplemental Material, by checking agreement and reliability of the responses focussing on study weights. We measured inter-rater agreement between eight reviewers R_i , $i = 1, \dots, 8$, and employed Reliability Analysis tools calculating Cronbach's alpha coefficient and inter-item correlations matrix. This analysis suggested that two reviewers (R_4 and R_5) had poor performance-related characteristics and therefore should be treated differently to other raters.

To produce a collective study weight and outcome weight to put into the Bayesian model, we followed the approach of mathematical aggregation³² where separate judgements elicited from the experts are combined into the aggregate estimate using a pooling rule. We used linear opinion pooling. To reflect differences in their reliability, each reviewer was assigned a certain weight,

Table 2. Bayesian hierarchical model specification.

Parameter	Proposal
$\hat{\rho}_{ml}$	$N(\rho_{ml}, V(\rho_{ml}, N_{ml}, k_{ml}))$
$\text{logit}(\rho_{ml})$	$N(\mu_m, \sigma_w^2/w_m)$
μ_m	$N(\mu, \sigma_b^2/w_m)$
μ	$N(0, 10\,000)$
σ_w	$U[0, 5]$
σ_b	$U[0, 5]$

N_{ml} and k_{ml} are the cluster size and number of clusters; vague priors $N(0, 10\,000)$ were assigned to mean μ for each model throughout; $U[0, 5]$ is a uniform distribution.

called an importance weight, which controlled the input of each reviewer into the pooled opinion. We employed the Rank Sum weight method.³⁴ In this approach, the weights are calculated as the individual ranks divided by the sum of the ranks:

$$\pi_j = \frac{(8 - \text{Rank}_j + 1)}{\text{Sum}(8 - \text{Rank}_j + 1)}$$

where Rank_j is the rank of the j th reviewer, $j = 1, \dots, 8$, producing normalised weights summing to one. The elicitation evaluation process and sensitivity analysis described below and in the Supplemental Materials suggested that downgrading the input of R_4 and R_5 would improve quality of the elicited weights. We therefore assigned R_4 and R_5 the lowest rank 8, with all others assigned rank 1. This calibration process produced lower importance weights $\pi_4 = \pi_5 = 0.02$ to reduce the impact of reviewers R_4 and R_5 on the pooled study and outcome weights, and $\pi_j = 0.16$ for all other reviewers. These importance weights were used in the modelling.

Bayesian hierarchical ICC modelling

All 34 ICC estimates from Table 1 were combined in the Bayesian hierarchical model (Table 2).

Studies included in the modelling had considerable variation in sample sizes, and consequently in ICC precision. In Figure 2, all 34 ICCs are plotted together with their 95% confidence intervals calculated using Swiger's formula.³⁵ Characteristics of the extracted weights are also shown in Supplemental Table 2.

The models were fitted within WinBUGS.³⁶

Results of Bayesian ICC modelling

Posterior ICC distribution

The constructed ICC distribution is summarised in Table 3. For the purpose of the sample size estimation, the ICC point estimate can be chosen by summarising the posterior distribution. The posterior median is the summary of location commonly used in practice.⁸ The

95% credible interval provides guidance regarding the probability of actually observing this particular ICC value and the adequacy of the planned sample size. For comparison, in a classical framework, the ICC estimate derived using the 34 ICCs from the identified studies could be calculated using one of the most commonly used approaches: median 0.05, mean 0.098, weighted mean 0.103 (using our outcome weights), maximum 0.4. These simple approaches do not take into account any differences between studies and their varying relevance to our target trial. Note that the maximum (0.4) is outside the 95% credible interval (0.00131–0.330) and is also likely to be overly conservative.

Sample size estimation

Modelling the ICC within a Bayesian framework provides the researcher with a full posterior distribution which can be summarised in a number of ways to provide the estimate for sample size calculation. We used the posterior median, although a range of posterior quantiles can be considered when designing a trial. For the post-stroke incontinence cluster-randomised trial, the sample size was chosen to provide at least 80% power with a 5% significance level to detect a minimally important between-group difference of 2.52 points³⁷ in mean ICIQ-UI-SF 3-month score, using an independent-samples t -test and a common standard deviation 8.32 computed from data collected for the ICONS feasibility trial.¹¹ An ICC was assumed to be less than or equal to the posterior median $\hat{\rho} = 0.0296$. The sample size was determined using PASS-16 sample size software.³⁸ It was assessed as realistic to recruit between 40 and 50 stroke units which requires the total sample size of $N = 480$ and $N = 450$ for $k = 20$ and $k = 25$ clusters per arm, with an average sample size per cluster $m = 12$ and $m = 9$, respectively. The power was 82%.

The advantage of using a Bayesian approach to determine sample size is that it provides a method which allows for imprecision in the ICC estimate, and informs researchers designing trials about the range of plausible sample size values, as opposed to the standard approach where sample size is calculated as a single number using a point ICC estimate. With a constructed posterior distribution of ICC, trials can be designed using a range of plausible ICC values, such as quantiles within the 95% credible interval; using values from the upper range of the credible interval would, however, probably be too conservative.

To incorporate the uncertainty about ICC into the sample size calculation, we used the posterior interquartile range of ICC to evaluate the range of plausible values of sample sizes which could be anticipated for the planned trial (Figure 3). The figure also shows sample sizes which would be obtained under the classical approach.

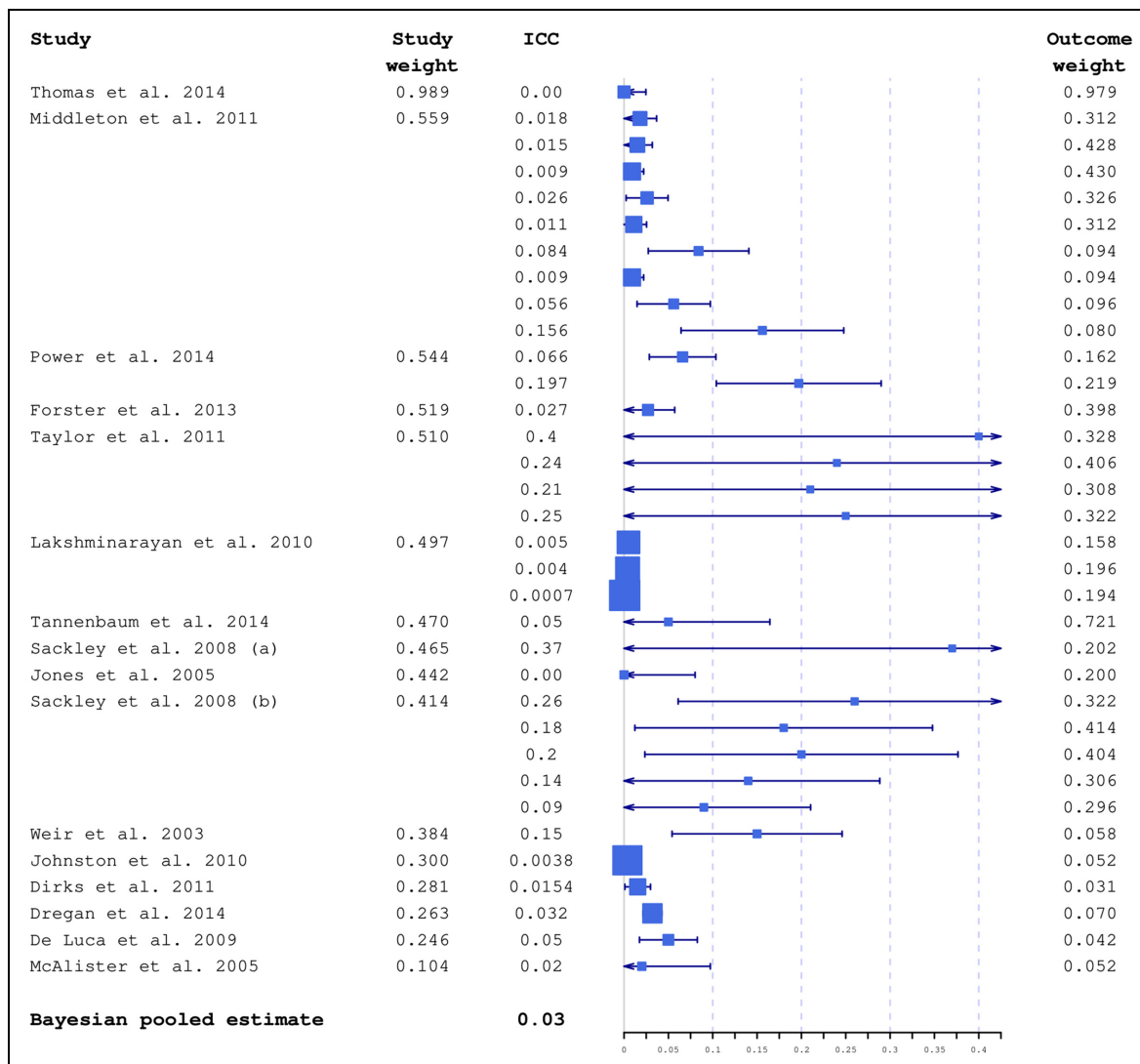


Figure 2. ICC estimates included in the modelling plotted together with 95% confidence intervals and average study and outcome weights. Box sizes are inversely proportional to variances. The studies are ordered by decreasing relevance to the planned study, based on estimated average study weight. The largest weights were from the ICONS feasibility trial.

Table 3. Summaries of posterior distributions constructed for the ICC and the model standard deviations, between-study S_b and within-study S_w .

Variable	Posterior mean (SD)	MC error	2.5% percentile	25% percentile	Posterior median	75% percentile	97.5% percentile
ICC	0.0607 (0.0937)	0.000145	0.00131	0.012	0.0296	0.0682	0.330
S_b	1.224 (0.409)	0.00369	0.633	0.940	1.156	1.43	2.217
S_w	0.345 (0.0874)	0.000413	0.206	0.284	0.335	0.395	0.546

ICC: intracluster correlation coefficient; SD: standard deviation; MC: Monte Carlo.

Figure 3 demonstrates that in this study the advantages of applying Bayesian modelling for ICC estimation are less apparent with large numbers of clusters. At $k = 60$, the sample sizes provided by the classical approach are the same as the sample size corresponding to the posterior median ICC ($N = 480$) due to the

small number of subjects per cluster. Figure 3 also illustrates the problem with low and moderate numbers of clusters in randomised controlled trials (RCTs): sample size becomes acceptable for practical purposes and reasonably stable only when the number of clusters is around 25 or greater.

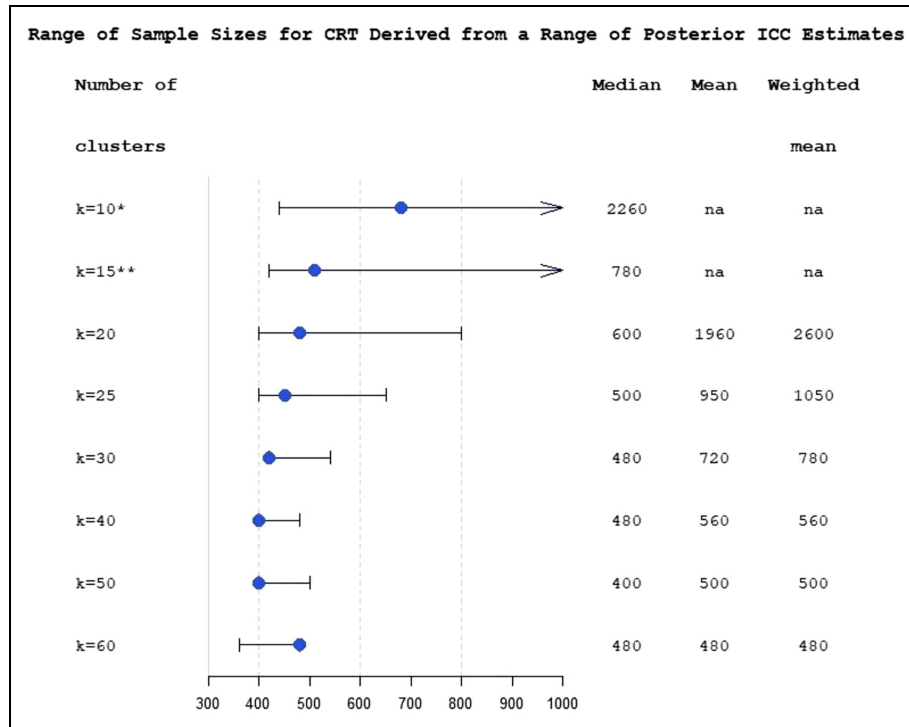


Figure 3. Range of sample sizes derived for different ICC values from posterior interquartile range of ICC estimate for the varying number of clusters at fixed levels (k from 10 to 60), for cluster-randomised trial with k equal size clusters per arm. The bullets are sample sizes calculated using posterior median ICC. The whiskers correspond to 25% and 75% posterior ICC quantiles. Median, mean and weighted mean columns show sample sizes calculated using a classical multi-estimate method.

All numbers correspond to at least 80% power achieved; *: 80% power is not achievable for ICC 75% quantile; **: sample size corresponding to ICC 75% quantile is 1440; na: 80% power is not achievable for this number of clusters.

Sensitivity analysis

To evaluate the sensitivity of the estimated ICC and sample size estimates to the model inputs and assumptions, we investigated sensitivity to (1) the choice of reviewers (and subsequently weights included in the modelling) by comparing three alternative versions of importance weights; (2) the choice of studies to be included into the modelling by investigating how focussing on most relevant studies (top 25%, 50% and 75%) would affect the results. Further details are in the Supplemental Material.

Sensitivity analysis demonstrated that the findings presented, and posterior ICC distribution in particular, showed limited sensitivity to the choice of the importance weights π_j and that the model fit is worse in scenarios where only the most relevant trials were included in the model (see Supplemental Table 3) suggesting that it is better to be overinclusive in terms of potential relevance. The model implemented with two-category Rank Sum importance weights and all 16 studies included demonstrated better fit.

Discussion

In this study, we employed a Bayesian framework that provides a flexible and informative way to handle ICC uncertainty and uses previously published or external

ICC estimates. We have presented an extension to the approach proposed by Turner et al.⁹ and described the implementation of a method to construct posterior distribution of the ICC using external information from available ICCs and expert knowledge.

The method suggested by Turner et al.⁹ has been used in several studies and practical applications,^{39,40} and has been extended to count data.⁴¹ However, it has not yet been adopted widely. The approach does require knowledge of a relatively advanced Bayesian technique, but another obstacle is the uncertainty in choosing study and outcome weights. The methodology we have proposed helps to overcome this problem. Using expert knowledge reduces subjectivity in choosing weights and improves informativity and robustness of the ICC estimate.

With the conventional approach, when an ICC estimate is imprecise or unreliable, researchers tend to choose a conservative ICC for their sample size calculation.^{1,42,43} This often leads to an unnecessarily inefficient trial, with more clusters than strictly necessary and hence greater overall trial costs. The strength of the suggested approach is that it provides justification for a robust and typically smaller ICC compared to the conventional approach, leading to sample size reductions and thus resulting in substantial efficiency savings for the proposed trial.

Using complex Bayesian models may require a greater investment of time and expertise. However, the benefit of using a Bayesian model to estimate the trial ICC is that it offers greater flexibility for combining available ICC estimates while incorporating uncertainty and information about the relevance of these estimates into the model.

As practical guidance, researchers wishing to explore the utility of the proposed approach would need to consider the following steps:

1. Identify and select ICCs for relevant outcome measures from previous relevant studies, through systematic review and exploring existing databases. The recommendation is to be highly inclusive in terms of potential relevance.
2. Summarise the selected studies with existing ICCs in a Summary Table similar to Supplemental Table 1.
3. Identify and invite experts in the topic relevant to the project, agree on elicitation technique, conduct training and task them with a ranking exercise to assign weights w_m to each study and w_{ml} to each outcome within the study using the Summary Table.
4. As an optional step, we recommend considering evaluation of the expert elicitation and agreement, and differentiate expert input into the model by introducing importance weights π_j , if required.
5. Aggregate the elicited weights and embed them into the Bayesian modelling of the targeted ICC.
6. Examine the sensitivity of the conclusions to the chosen model.
7. Choose an appropriate ICC estimate using the Bayesian posterior distribution of the ICC, and then use that to provide one or more estimates of the sample size.

The main strength of this study is that we have proposed a practical method of implementing the synthesis of externally available ICCs within a Bayesian framework using expert opinion.

The limitation of our approach is the validity and consistency of the reviewers' ratings. This can be mitigated by increasing the number of reviewers, conducting more targeted training prior to the ranking exercise, and applying different knowledge elicitation techniques. As an extension, strategies for evaluating the elicitation exercise (including ranking reviewers, agreement and coherence checks, calibration)³⁰ can be embedded within the elicitation process, together with examining sensitivity of the conclusions to the used models.⁴⁴ A more advanced approach would be to set up calibration questions, where experts are asked questions where the truth is known.⁴⁵ The choice of the reviewers in this study was a convenience purposive expert sample, and it could be improved in further practical implementation of the method.

Evaluating the elicited beliefs can be affected by a range of biases.^{30,46} Johnson et al.⁴⁶ developed a conceptual framework outlining the belief-elicitation process. They emphasised that elicitation methods should be evaluated in respect to such measurement properties as *validity*, *reliability*, *responsiveness* and *feasibility*, with *validity* and *reliability* being a prerequisite for meeting methodological standards. For this study, evaluation of *validity* was limited as there was no gold standard for the elicitation of the required probability weights. *Reliability*, and inter-rater reliability in particular, was evaluated using appropriate measures of association. The *responsiveness* was not applicable in this study and the property of *feasibility* was not directly evaluated, although the elicitation process was designed to minimise required time, costs and need for equipment. Further research on developing methodological strategies to evaluate measurement properties would help to reduce the influence of potential biases on the weight elicitation in the proposed framework.

The methodology described in this article is proposed in the context of continuous outcome data, as was Swiger's approach. However, Swiger's formula can easily be extended to construct interval estimates for the ICC in the setting of binary outcome clustering by replacing the appropriate quantities in the formula with the binary outcome equivalents,⁷ expanding the practical applicability of the method.

The proposed methodology reduces the impact of uncertainty in the ICC on the design. A next step towards more robust study design could be calculating a mean power (assurance)⁴⁷⁻⁴⁹ using the ICC distribution produced by our method. Choosing a sample size to achieve a desired assurance, rather than to achieve a desired power, conditional on an assumed point estimate of the ICC, would help to protect further against loss of power, although the produced sample size would be typically larger.⁵⁰

There are a range of knowledge elicitation tools and techniques which can influence effectiveness of the elicitation. Using online graphical elicitation tool can provide an accessible and intuitive framework for eliciting the information and would naturally produce prior distributions for statistical models.⁴⁴ Applying leading elicitation protocols – such as Cooke, SHELF or probabilistic Delphi – can minimise bias and improve accuracy in multiple expert judgements.^{32,33} Future implementation of the method could be expanded to use elicitation software, such as the Sheffield Elicitation Framework and associated web-based version,^{51,52} to gather expert knowledge in the form of probability distributions for unknown quantities.

Acknowledgements

The authors acknowledge input to the paper of the following people: Michael Leathley (systematic search of the relevant

papers, summarising the extracted studies and reviewing of the selected papers), Marian Brady, Bernie Carter, Cliff Panton and Brenda Roe (review of the selected papers).

Author contributions

C.J.S. and S.V.T. made substantial contributions to the conception. S.V.T. wrote the first draft of the manuscript, was responsible for all analyses and interpretation of results. C.J.S. contributed to interpretation of results. All authors contributed to collection and interpretation of data, reviewed and approved the submitted manuscript.

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: S.V.T.'s work on this study was partially supported by Research Catalyst Fund of University of Central Lancashire.

ORCID iD

Svetlana V Tishkovskaya  <https://orcid.org/0000-0003-3087-6380>

Data availability statement

All data generated or analysed during this study are included in this published article as supplemental information files.

Supplemental material

Supplemental material for this article is available online.

References

- Eldridge S and Kerry S. *A practical guide to cluster randomised trials in health services research*. New York: Wiley, 2012.
- Killip S, Mahfoud Z and Pearce K. What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *Ann Fam Med* 2004; 2(3): 204–208.
- Kerry SM and Bland JM. The intracluster correlation coefficient in cluster randomisation. *BMJ* 1998; 316: 1455–1460.
- Rutterford C, Copas A and Eldridge S. Methods for sample size determination in cluster randomized trials. *Int J Epidemiol* 2015; 44(3): 1051–1067.
- Campbell MJ, Donner A and Klar N. Developments in cluster randomized trials and statistics in medicine. *Stat Med* 2007; 26: 2–19.
- Donner A, Brown KS and Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989. *Int J Epidemiol* 1990; 19(4): 795–800.
- Turner RM, Prevost AT and Thompson SG. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Stat Med* 2004; 23: 1195–1214.
- Gelman A, Carlin JB, Stern HS, et al. *Bayesian data analysis*. 3rd ed. Abingdon: Taylor & Francis, 2013.
- 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.
- 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: 322–330.
- Thomas LH, French B, Sutton CJ, et al. *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*. Southampton: NIHR Journals Library, 2015.
- Swiger LA, Harvey WR, Everson DO, et al. The variance of intraclass correlation involving groups with one observation. *Biometrics* 1964; 20: 818–826.
- Sutton CJ, Watkins CL and Dey P. Illustrating problems faced by stroke researchers: a review of cluster-randomized controlled trials. *Int J Stroke* 2013; 8(7): 566–574.
- Thomas LH, Watkins CL, Sutton CJ, et al. Identifying continence options after stroke (ICONS): a cluster randomised controlled feasibility trial. *Trials* 2014; 15: 509.
- Tannenbaum C, Agnew R, Benedetti A, et al. Effectiveness of continence promotion for older women via community organisations: a cluster randomised trial. *BMJ Open* 2013; 3: e004135.
- Sackley CM, Rodriguez NA, van den Berg M, et al. A phase II exploratory cluster randomized controlled trial of a group mobility training and staff education intervention to promote urinary continence in UK care homes. *Clin Rehabil* 2008; 22(8): 714–721.
- Sackley C, Wade DT, Mant D, et al. Cluster randomized pilot controlled trial of an occupational therapy intervention for residents with stroke in UK care homes. *Stroke* 2006; 37(9): 2336–2341.
- Weir CJ, Lees KR, MacWalter RS, et al. Cluster-randomized, controlled trial of computer-based decision support for selecting long-term anti-thrombotic therapy after acute ischaemic stroke. *QJM* 2003; 96(2): 143–153.
- De Luca A, Toni D, Lauria L, et al. An emergency clinical pathway for stroke patients – results of a cluster randomised trial (isrctn41456865). *BMC Health Serv Res* 2009; 9: 14.
- Dirks M, Niessen LW, van Wijngaarden JD, et al. Promoting thrombolysis in acute ischemic stroke. *Stroke* 2011; 42: 1325–1330.
- Johnston SC, Sidney S, Hills NK, et al. Standardized discharge orders after stroke: results of the quality improvement in stroke prevention (QUISP) cluster randomized trial. *Ann Neurol* 2010; 67(5): 579–589.
- Jones A, Tilling K, Wilson-Barnett J, et al. Effect of recommended positioning on stroke outcome at six months: a randomized controlled trial. *Clin Rehabil* 2005; 19(2): 138–145.

23. Lakshminarayan K, Borbas C, McLaughlin B, et al. A cluster-randomized trial to improve stroke care in hospitals. *Neurology* 2010; 74: 1634–1642.
24. McAlister FA, Man-Son-Hing M, Straus SE, et al. Impact of a patient decision aid on care among patients with nonvalvular atrial fibrillation: a cluster randomized trial. *CMAJ* 2005; 173: 496–501.
25. Forster A, Dickerson J, Young J, et al. A structured training programme for caregivers of inpatients after stroke (TRACS): a cluster randomised controlled trial and cost-effectiveness analysis. *Lancet* 2013; 382: 2069–2076.
26. Taylor WJ, Brown M, William L, et al. A pilot cluster randomized controlled trial of structured goal-setting following stroke. *Clin Rehabil* 2012; 26(4): 327–338.
27. Middleton S, McElduff P, Ward J, et al. Implementation of evidence-based treatment protocols to manage fever, hyperglycaemia, and swallowing dysfunction in acute stroke (QASC): a cluster randomised controlled trial. *Lancet* 2011; 378: 1699–1706.
28. Power M, Tyrrell PJ, Rudd AG, et al. Did a quality improvement collaborative make stroke care better? A cluster randomized trial. *Implement Sci* 2014; 9: 40.
29. Dregan A, van Staa TP, McDermott L, et al. Point-of-care cluster randomized trial in stroke secondary prevention using electronic health records. *Stroke* 2014; 45(7): 2066–2071.
30. O'Hagan A, Buck CE, Daneshkhan A, et al. *Uncertain judgements: eliciting experts' probabilities*. New York: Wiley, 2006.
31. Hanea AM, McBride MF, Burgman MA, et al. Classical meets modern in the IDEA protocol for structured expert judgement. *J Risk Res* 2018; 21: 417–433.
32. O'Hagan A. Expert knowledge elicitation: subjective but scientific. *Am Stat* 2019; 73: 69–81.
33. European Food Safety Authority. Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA J* 2014; 12: 3734.
34. Stillwell WG, Seaver DA and Edwards W. A comparison of weight approximation techniques in multiattribute utility decision making. *Organ Behav Hum Perform* 1981; 28: 62–77.
35. Donner A and Wells G. A comparison of confidence interval methods for the intraclass correlation coefficient. *Biometrics* 1986; 42: 401–412.
36. Lunn DJ, Thomas A, Best N, et al. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput* 2000; 10: 325–337.
37. Nystrom E, Sjostrom M, Stenlund H, et al. ICIQ symptom and quality of life instruments measure clinically relevant improvements in women with stress urinary incontinence. *Neurourol Urodyn* 2015; 34: 747–751.
38. NCSS Statistical Software. *PASS 16 power analysis and sample size software*. Kaysville, UT: NCSS Statistical Software, 2018.
39. Wells GA, Kelly S, Elliott J, et al. Direct oral anticoagulants for the treatment of venous thromboembolic events: a systematic review and network meta-analysis, <https://www.ottawaheart.ca/sites/default/files/uploads/documents/Researchers/gwells-doac-vte-scientific-report-2015-2016.pdf>
40. Wills K. *Uncertainty in estimating between-teacher variation for value-added modeling: a Bayesian perspective*. Seattle, WA: University of Washington, 2014.
41. Clark AB and Bachmann MO. Bayesian methods of analysis for cluster randomized trials with count outcome data. *Stat Med* 2010; 29: 199–209.
42. Hade EM, Murray DM, Pennell ML, et al. Intraclass correlation estimates for cancer screening outcomes: estimates and applications in the design of group-randomized cancer screening studies. *J Natl Cancer Inst Monogr* 2010; 2010(40): 97–103.
43. 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.
44. Mason AJ, Grieve RD, Richards-Belle A, et al. A framework for extending trial design to facilitate missing data sensitivity analyses. *BMC Med Res Methodol* 2020; 20: 66.
45. Colson AR and Cooke RM. Expert elicitation: using the classical model to validate experts' judgments. *Rev Environ Econ Policy* 2018; 12: 113–132.
46. Johnson SR, Tomlinson GA, Hawker GA, et al. Methods to elicit beliefs for Bayesian priors: a systematic review. *J Clin Epidemiol* 2010; 63(4): 355–369.
47. O'Hagan A, Stevens JW and Campbell MJ. Assurance in clinical trial design. *Pharm Stat* 2005; 4: 187–201.
48. Grieve AP. *Hybrid frequentist/Bayesian power and Bayesian power in planning clinical trials*. Boca Raton, FL: CRC, 2022.
49. 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.
50. Chen DG, Fraser MW and Cuddeback GS. Assurance in intervention research: a Bayesian perspective on statistical power. *J Soc Soc Work Res* 2018; 9: 159–173.
51. Morris DE, Oakley JE and Crowe JA. A web-based tool for eliciting probability distributions from experts. *Environ Model Softw* 2014; 52: 1–4.
52. Oakley JE and O'Hagan A. SHELF: the Sheffield elicitation framework (version 2.0), <http://tonyohagan.co.uk/shelf> (2010, accessed 17 June 2020).