American Journal of Biostatistics 1 (1): 1-8, 2010
ISSN 1948-9889

Design Based on Intra-Class Correlation Coefficients

Gheorghe Doros and Robert Lew
Department of Biostatistics, Boston University, Boston, MA, 02118 USA

Abstract: Problem statement: Reliability studies are concerned with the study of “consistency” or “repeatability” of measurements. Often times (but not always) the reliability coefficients are Intra-class Correlation Coefficients (ICC). Depending on the design or the conceptual intent of the study there are three types of intra-class correlation coefficients, termed intra-class correlation coefficients Case 1, 2 and 3, for measuring the reliability of a single interval measure. While methods for sample size calculations for intra-class correlation coefficients in Case 1 are available and implemented in Power Analysis and Sample Size System (PASS); to our knowledge, no methods based on intra-class correlation coefficients in Case 2 and 3 are available. **Objective:** Develop a method for calculating the size of a reliability study based on intra-class correlation coefficients Case 1 and 2. **Approach:** A practical method for computing sample size using simulations was proposed. We proposed to compute sample size based on the expected width of the confidence interval. For a given target value of the intra-class correlation coefficient, the proposed method chooses the design assures a 95% confidence interval with average length shorter than a pre-specified value. The applicability of the proposed method in practice for intra-class coefficients Case 2 was supported by demonstrating three invariance properties of the proposed confidence intervals. **Results:** Tables with sample size requirements were derived and displayed. A program for carrying out the calculations was developed in R. The method was used to size a trial aimed to study the reliability of a scale that measures the cleanness of the colon at the time of colonoscopy. **Conclusion:** A simple method for sample size calculation for intra-class correlation coefficient Case 1 and 2 based on the average length of confidence intervals was proposed. The proposed was implemented by the authors in R (freely available software). Three invariance properties of the confidence intervals for the intra-class correlation coefficients Case 2 were studied by simulations. These properties are an important tool when considering the design of this type of studies.

Key words: Intra-class correlation coefficients, sample size based on confidence interval, reliability studies

INTRODUCTION

Intra-class correlation coefficients are commonly used in epidemiology, psychology, sociology and medicine as reliability coefficients. Their area of application, however, extends beyond these research fields. Reliability studies are concerned with the study of “consistency” or “repeatability” of measurements. Often times (but not always) the reliability coefficients are Intra-class Correlation Coefficients (ICC). Depending on the design or the conceptual intent of the study (Shrout and Fleiss, 1979) describe three types of intra-class correlation coefficients for measuring the reliability of a single interval mea-sure, which they term intra-class correlation coefficients Case 1, 2 and 3. While methods for sample size calculations for intra-class correlation coefficients in Case 1 are available and implemented in PASS (Power Analysis and Sample Size System) (Walter *et al*., 1998; Winer, 1991); to our knowledge, no methods based on intra-class correlation coefficients in Case 2 and 3 are available. Available approaches to calculate sample size for intra-class correlation coefficients ρ are based on power of tests for hypothesis H_0: $\rho = \rho_0$ which start from the premise that an hypothesis test will be used. However, often in reliability studies the focus is on estimation of the intra-class correlation coefficients, not on testing. Thus, using power as criterion for sample size calculation is not an appropriate approach. The criterion used in sample size calculations should be a measure of quality of the estimator. Designs based on confidence intervals have been proposed before (Beal, 1989; Cochran and Cox, 1957; Daly, 1991; Greenland, 1988; McHugh and Le, 1984; O’Neill, 1984) however, not for ICC

Corresponding Author: Gheorghe Doros, Department of Biostatistics, Boston University, 801 Mass. Ave. Room 331, Boston, MA, 02118 USA

coefficients case 2 and 3. In this study we propose a method for calculating the size of a study based on confidence intervals of intra-class correlation coefficients. We apply our method to intra-class correlation coefficients in case 1 and 2.

Intra-class correlation coefficients: A generic definition of an intra-class correlation coefficient ρ is:

$$\frac{\text{'True Variance'}}{\text{'Observed Variance'}}$$

Where:

‘True Variance’	=	The variability between the targets
‘Observed Variance’	=	The total variance-true variance plus other variance

In many cases, but not always, intra-class correlation coefficients are reliability coefficients. Depending on the design or the conceptual intent of the study (Shrout and Fleiss, 1979) describe three classes of intra-class correlation coefficients to measure reliability, which they term Case 1, 2 and 3. In each case n randomly chosen targets are rated by k raters, with the distinction that for Case 1-each target is rated by different raters, for Case 2-the same raters rate each target and for Case 3-all possible raters rate each target. For each of the three cases above, Shrout and Fleiss (1979) further distinguish two cases: First in which the aim is to estimate the reliability of a single rating and the second in which the aim is to estimate the reliability of the average several ratings, thus resulting in a total of six ICC’s. The above mentioned authors coined the notation ICC(l,m) for these coefficients, where m is number of repeated measurements from the same rater on same target and l = 1, 2, 3 is the case. In this study we will only refer to ICC(1,1) and ICC(2,1). For a different categorization of various ICCs also (McGraw and Wong, 1996).

Variances entering the definition of ICC(1,1) and ICC(2,1) are estimated with a one-way and two-way random effects ANOVA, respectively. For Case 1 it is assumed that the j^{th} observation y_{ij} for target i (i = 1, 2,…, n; j = 1, 2,…, k) satisfies:

$$y_{ij} = \mu + t_i + \in_{ij} \qquad (1)$$

where, μ is the population mean of the measurements; the random target effects t_i and the measurement errors $\in_{ij}$ are independent, normally distributed random variables with mean 0 and variance σ^2_T and σ^2_W, respectively. We assume the targets are randomly sampled from some population of interest. For Case 2 it is assumed that the j^{th} observation y_{ij} for target i (i = 1, 2,…, n; j = 1, 2,…, k) satisfies:

$$y_{ij} = \mu + t_i + r_j + \in_{ij} \qquad (2)$$

where, μ is the population mean of the measurements; the random target effects t_i, the random rater effects r_i and the measurement errors $\in_{ij}$ are independent normally distributed random variables with mean 0 and variance σ^2_T, σ^2_J and σ^2_E, respectively. We assume the targets and raters are randomly sampled from some populations of interest. The Analysis Of Variance (ANOVA) tables for the models in Eq. 1 and 2 are summarized in Table 1.

Following the notation in (Shrout and Fleiss, 1979), several key results on the estimation of ICC(1,1) and ICC(2,1) are summarized next.

ICC(1,1) coefficient: Coefficient. The interclass correlation coefficient ICC(1,1) is defined as:

$$\rho = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_W} \qquad (3)$$

As a measure of inter-rater reliability, the intra-class correlation above is the proportion of total variance in observed scores accounted for by the subject-to-subject variability in the true, unobserved scores. A consistent (biased) estimate of this coefficient is (Shrout and Fleiss, 1979):

$$\hat{\rho} = \frac{(\text{BMS} - \text{WMS})/k}{(\text{BMS} - \text{WMS})/k + \text{WMS}}$$

Confidence interval: A 100(1-ff)% confidence interval for this parameter is presented by Fleiss and Shrout (1978) and can be expressed as:

$$\text{CI}(\alpha) = \{L(\alpha) < \rho < U(\alpha)\}$$

with the lower bound:

$$L(\alpha) = \frac{\rho_0 / q_F(1-\alpha/2, n-1, n(k-1)) - 1}{\rho_0 / q_F(1-\alpha/2, n-1, n(k-1)) + k - 1}$$

Table 1: ANOVA table for case 1 and 2

			Expected MS	
Source of variation	DF	MS	Case 1 One-way random effects	Case 2 Two-way random effects
Between targets	n-1	BMS	$k\sigma^2_T + \sigma^2_W$	$k\sigma^2_T + \sigma^2_E$
Within targets	n(k-1)	WMS	σ^2_W	-
Between raters	k-1	JMS	-	$n\sigma^2_J + \sigma^2_E$
Residual	(n-1)(k-1)	EMS	-	σ^2_E

and the upper bound:

$$U(\alpha) = \frac{\rho_0 / q_F(1-\alpha/2, n-1, n(k-1)) - 1}{\rho_0 / q_F(1-\alpha/2, n-1, n(k-1)) + k - 1}$$

Where:

ρ_0 = BMS/WMS

q_F = (1-α, l, m) is the 100(1-α)% percentile of an F distribution with l and m degrees of freedom

A study of various confidence intervals for ICC(1,1) is presented in Donner and Wells (1986).

ICC(2,1) coefficient:

Coefficient: The intra-class correlation coefficient ICC(2,1) is defined as:

$$\rho = \frac{\sigma^2{}_T}{\sigma^2{}_T + \sigma^2{}_J + \sigma^2{}_E} \tag{4}$$

As a measure of inter-rater reliability, the intra-class correlation above is the proportion of total variance in observed scores accounted for by the subject-to-subject variability in the true, unobserved scores. A consistent (biased) estimate of this coefficient is (Shrout and Fleiss, 1979):

$$\hat{\rho} = \frac{(BMS - EMS)/k}{(BMS - EMS)/k + (JMS - EMS)/n + EMS}$$

Confidence interval: Constructing a confidence interval for this parameter is more difficult because it is a function of three independent quantities. An approximate 100(1-α)% confidence interval for this parameter is provided by Fleiss and Shrout12 and it has complicated form. We first set-up some notation. Let q_F (1-α, l, m) be the 100(1-α)% percentile of an F distribution with l and m degrees of freedom, let:

$$\rho_0 = JMS/EMS$$

and let:

$$\delta = \frac{(k-1)(n-1)[\hat{\rho}\rho_0 + n(1+(k-1)\hat{\rho}) - k\hat{\rho}]^2}{(n-1)k^2\hat{\rho}^2\rho_0^2 + [n(1+(k-1)\hat{\rho}) - k\hat{\rho}]^2}$$

With this notation, the confidence interval can be expressed as:

$$CI(\alpha) = \{L(\alpha) < \rho < U(\alpha)\}$$

Where:

$$L(\alpha) = \frac{n[BMS - q_F(1-\alpha/2, n-1, \delta)EMS]}{q_F(1-\alpha/2, n-1, \delta)\begin{bmatrix} kJMS + \\ (kn+k-n)EMS \end{bmatrix} + nBMS} \tag{5}$$

and

$$U(\alpha) = \frac{n[q_F(1-\alpha/2, \delta, n-1)BMS - EMS]}{kJMS + (kn+k-n)EMS + nq_F\begin{pmatrix} 1-\alpha/2, \\ \delta, n-1 \end{pmatrix}BMS} \tag{6}$$

The confidence intervals proposed in (Cappelleri and Ting, 2003)] improve slightly on coverage over the intervals proposed by Fleiss and Shrout (1978), however, the improvement is not great and we will not be using them here. An approach to inference for ICC(2,1) based on a generalized variable model have been proposed by Tian and Cappelleri (2004).

Sample size calculations based on ICC(2,1): In reliability studies, the main aim is not testing but accurate estimation. For this reason sample size calculations should focus on precision rather than power (Beal, 1989; Daly, 1991; Greenland, 1988). Design based on confidence intervals have been proposed before in (Beal, 1989; Cochran and Cox, 1957; Daly, 1991; Greenland, 1988; McHugh and Le, 1984; O'Neill, 1984)] but not for ICC(2,1) coefficients.

Confidence intervals and sample size: Designing a study is an involved process. Sample size justification is part of the design. The traditional approach to sample size calculations is based on power. This approach starts from the premise that a test of hypothesis will be used. Hypothesis testing and confidence intervals are closely related. For instance, when a 100(1-α)% confidence interval is constructed, all values in the interval are considered plausible values for the parameter being estimated. Values outside the interval are rejected as relatively implausible. If the value of the parameter specified by the null hypothesis is contained in the 100(1-α)% interval, then the null hypothesis cannot be rejected at the 100α% level. If the value specified by the null hypothesis is not in the interval, then the null hypothesis can be rejected at the 100% level. Over the past decades many editorial boards have recommended that in the analysis of data, confidence intervals be used rather than hypothesis tests whenever confidence intervals are warranted. In spite of this shift in the presentation of the data many studies continue to use power or minimal detectable difference as the criteria for sample size calculations. Estimates of sample size based on confidence interval can be quite different from the estimates based on power depending on the criteria used for sample size calculation (Greenland, 1988). Different criteria for sample size calculation based on confidence intervals have been proposed. For example, Cochran and Cox (1957), McHugh and Le (1984) and O'Neill (1984) propose using expected width of confidence interval, while Beal (1989), Daly (1991) and Greenland (1988) propose different concepts of 'power' for confidence interval in calculating sample size. Beal (1989) proposes the use of the conditional probability that the half length of a confidence interval is smaller that a preset value given that the interval includes the true value be used as a criteria for sample size calculations. Greenland (1988) introduces the concept of 'discriminatory power' for calculating sample size based on confidence interval and Daly (1991) proposes to translate the standard methods for calculating sample size based on power into the confidence interval framework. Kupper and Hafner (1989) for an evaluation of the performance of sample size calculations based on confidence intervals. Finally, many authors agree that when the main aim of the study is estimation the average length of the confidence interval can be used as an criterion for sample size calculations (Beal, 1989; Daly, 1991; Greenland, 1988; Maxwell and Kelley, 2007).

MATERIALS AND METHODS

We will concentrate on sample size calculations based on ICC(2,1). The method proposed by Walter *et al.* (1998) applies only for the designs where the intra-class correlation coefficient is derived from a one-way ANOVA, hence not for a design based on ICC(2,1). The authors state "A treatment of sample size calculation in the two-way ANOVA cases, taking rater effects explicitly into account seems needed,...". Kraemer (1976) considers reliability in a two-way ANOVA framework, however the definition for the intra-class correlation coefficient is different from the traditional definition; in her definition the rater variability is excluded.

We propose to compute sample size based on the expected width of the confidence interval. The proposed method can be briefly outlined as follows: For a given target value of ρ design a study (choose n and k) that will assure a 95% confidence interval with average length shorter than a pre-specified value Δ. This method is readily applicable for ICC(1,1) where the only parameter needed is the target ρ; however, for ICC(2,1) besides the target ρ we would also need guess-estimates for two of the three variance components. Assuming that extra information (estimates) can be elicited, we can use Monte Carlo simulations to simulate from the distribution of the 'approximate' confidence bounds presented in Eq. 5 and 6 and calculate the appropriate sample size using the following steps:

- Simulate data $y_{i,j}$ according to the model in Eq. 2 and calculate BMS, JMS and EMS
- calculate the expected length of the confidence intervals for a variety of n and k values
- Choose the design (n and k) which results in a confidence interval with expected length smaller than a value Δ

The method as outlined above is rarely practical. The amount of information needed is rarely available. Moreover, even if all the information were available it is not clear that the proposed method has any desirable properties: for example, the property that the length of confidence interval decreases with the increasing amount of information (i.e., n and k).

RESULTS

Next, several properties of the confidence interval of ICC(2,1) are investigated. Armed with these properties, a new method is proposed. More precisely, through a simulation study we examine whether a scaling invariance property holds for the approximate confidence interval and whether the expected length of the confidence interval decreases with the increase of either n or k or ρ. Also, we examine the dependence of the average length of the confidence interval on the ratio $r_{T,E}(\sigma^2_T/\sigma^2_E)$.

Properties of confidence intervals for ICC(2,1): Properties of the confidence interval for the ICC(2,1) are examined through an extensive simulation study. We considered scenarios in which $2\leq k\leq 10$, $k\leq n\leq 30$, $\rho\geq 0.6$ and $\rho/(1-\rho)\leq r_{T,E}\leq 40$. Note that, for a given ρ, the ratio $r_{T,E}$ is lower bounded by $\rho/(1-\rho)$ because $0<\sigma^2_T/\sigma^2_E = (1/\rho-1)r_{T,E}-1$. For each combination of n, k, ρ and $r_{T,E}$, in order to calculate the expected length of the approximate confidence interval proposed by Fleiss and Shrout (1978), we first simulate data $y_{i,j}$ according to the model in Eq. 2, then calculate BMS, JMS and EMS, then calculate the approximate bounds according to the formulas in Eq. 5 and 6 and then average over the simulation results. For each scenario we use 1,000 simulations. The simulations are run using the R package (The R Development Core Team, 2005).

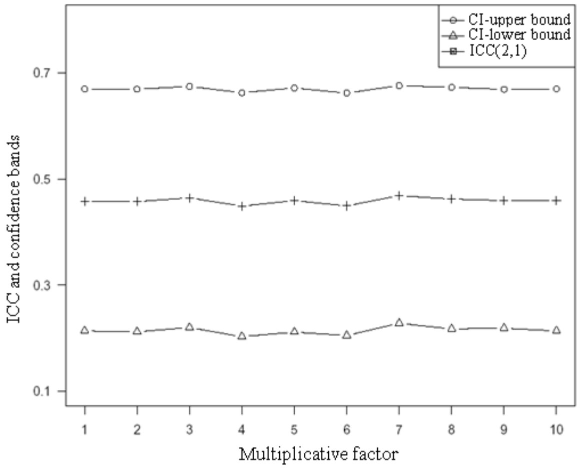

Fig. 1: Scale invariance (n = 30, k = 4, $r_{T,E}$ = 4, $r_{J,E}$ = 3)

Scaling property: The formula in Eq. 4 defining ICC(2,1) can be written as:

$$\rho = \frac{\sigma^2_T / \sigma^2_E}{\sigma^2_T / \sigma^2_E + \sigma^2_J / \sigma^2_E + 1}$$

that is, this coefficient depends on σ^2_T, σ^2_J and σ^2_E only through the ratios $r_{T,E}$ and $r_{J,E}$(= σ^2_T/σ^2_E). The question is whether the expected confidence interval inherit the above scaling property, that is, will different triplets σ^2_T, σ^2_J and σ^2_E with the same ratios ($r_{T,E}$ and $r_{J,E}$) result in the same expected confidence intervals? Figure 1 shows the results for n = 30, k = 4, $r_{T,E}$ = 4 and $r_{J,E}$ = 3. Similar results were found in all other scenarios we considered. Simulations support this scaling property for the confidence intervals. In other words, the average confidence interval depends on σ^2_T, σ^2_J and σ^2_E only through the ratios $r_{T,E}$ and $r_{J,E}$.

Dependence on $r_{T,E}$: Turning now to the dependence of the average confidence interval on $r_{T,E}$, for given n, k and ρ, simulations will be run for a set of $r_{T,E}$ values. Figure 2 shows the average length of confidence intervals as a function of $r_{T,E}$ for n = 20 and k = 7. Similar results were found in all other scenarios we considered. The conclusion is that the average length increases with the variance ratio $r_{T,E}$, however the increase is not big and it plateaus.

Monotonicity: The next task is to assess if the average length of the confidence interval decreases with the increase of either the number targets n, or the number of raters k or the magnitude of the coefficient ρ. Figure 3 shows the average length of confidence interval for n = 10; 20 and 30, k = 4; 7 and 10 and ρ = 0:7; 0:75; 0:8; 0:85 and 0:9. The average over a wide range of $r_{T,E}$ values has been used to construct this graph.

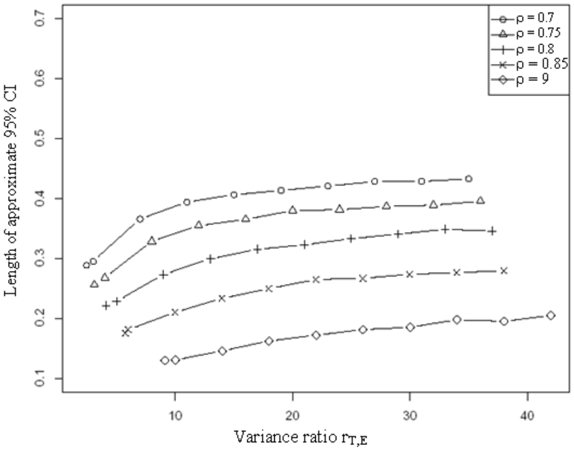

Fig. 2: Average length of confidence intervals as a function of $r_{T,E}$ (n = 20, k = 7)

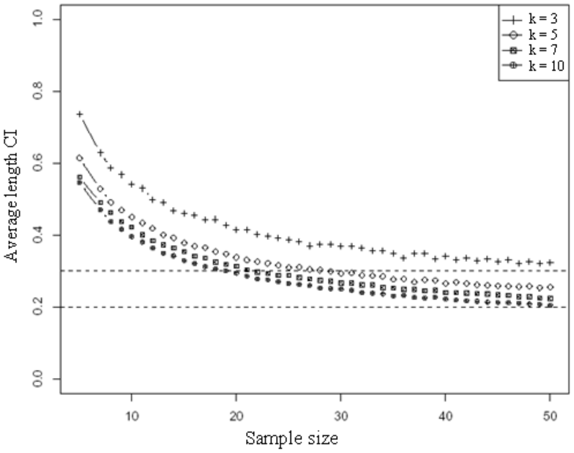

Fig. 3: Monotonicity-length of confidence interval averaged over a range of variance ratios

These results indicate that, increasing the number of targets or/and raters will provide a narrower confidence interval and, everything else being constant, the average length of the confidence interval for a value ρ_1 will be smaller than the average length of the confidence interval corresponding to a value $\rho_0<\rho_1$.

The scale invariance property of the average approximate confidence interval reduces the dimensionality of the problem. In other words, instead of having to input into the calculation two parameters (two of the σ^2_T, σ^2_J, or σ^2_E), the input of only one (one of the $r_{T,E}$ or $r_{J,E}$) is necessary. The average length of the confidence interval increases with the ratio $r_{T,E}$. When one of these ratios can be elicited from a pilot study the sample size can be calculated using simulations;

in practice, however, it is often difficult to obtain prior estimates of these ratios (Walter *et al*., 1998). Even when this is possible, the estimates come with big uncertainty as they are either obtained from small pilot studies, or they are adopted from studies on different measures and assumed to hold approximately true for the measure under study. Therefore, when these estimates are not available, the average confidence interval for a range of $r_{T,E}$ values will be calculated. Thus, a more practical approach is the following: for a target ρ and fixed k, calculate n that assures an average confidence interval length that is smaller than a value Δ following the steps:

- Choose a range of likely values for $r_{T,E}$ and chose a set of values (i.e., a grid) that spans this range
- For a target ρ and the set of values for $r_{T,E}$ as above simulate data $y_{i,j}$ according to model in Eq. 2 and calculate BMS, JMS and EMS
- Calculate the expected length of the confidence intervals for a variety of n and k and then average over the grid of $r_{T,E}$ values
- Choose the design (n and k) that gives a confidence interval with length smaller than Δ

Based on our experience this method will result in a conservative estimate of the sample size. When this approach is used we would recommend using Δ between 0.3 and 0.5. Smaller values will often result is sizes (n and k) that are too large to be practical. A script in R is available from the author upon request.

Example: A clean bowel is crucial to a successful colonoscopy. A dirty colon can preclude the doctor seeing polyps. That is why prior to a colonoscopy procedure, the patients are asked to drink a bowel cleaning medicine. Lai *et al.* (2007) propose a scale that measures how clean the colon is at the time of colonoscopy. A pilot study was conducted for a proposed Bowel Preparation Scale to inform the design of a reliability study to asses the reliability of the scale. Results from a pilot study suggested a reliability coefficient of around 0.7 ($\hat{r}_{T,E} = 3.6$, $\hat{r}_{J,E} = 0.4$). The main goal is determine how many colonoscopies and raters are needed to ensure on average length for a 95% confidence interval of 0.3 or less. The results of the calculations are shown in Fig. 4. By increasing the number of raters (k) from 5-7 or from 7-10 we do not gain much. In other words, most benefit in power is obtained by increasing n; a similar result holds for ICC(1,1) (Walter *et al.*, 1998). As the principal investigator was able secure ten raters we choose k = 10.

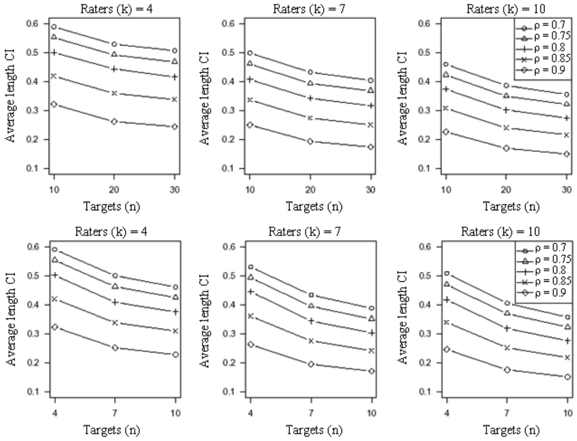

Fig. 4: Example-average length for confidence intervals

The minimum number of colonoscopies needed to assure an average length for the 95% confidence interval smaller than Δ = 0.3 was determined to be 19. If we average over a grid of values between 3.5 and 6, with Δ = 0.3 and k = 10 raters, we would need n = 25 colonoscopies. The latter estimate for the number of colonoscopies might be more realistic given the uncertainty around the estimate for $r_{T,E}$.

DISCUSSION

The design for reliability studies have not been well studied. In Kraemer (1976) addresses the problem of sample size calculation for a design based on ICC derived from a two-way ANOVA, however, the author's definition for the intra-class correlation coefficient is different from the traditional definition. The method proposed by Walter *et al*. (1998) applies only for the designs where the intra-class correlation coefficient is derived from a one-way ANOVA. Hence a method to deal with ICC(2,1) was needed.

A method for sample size calculation for intra-class correlation coefficient ICC(2,1) based on the average length of confidence intervals is proposed. A target ρ and an estimate for one of the ratios $r_{T,E}$ or $r_{J,E}$ has to be specified. In this study we chose to concentrated on the parameterization (ρ; $r_{T,E}$) however the method could be easily be applied for cases when prior information is available on $r_{J,E}$ in which case the parameterization (ρ; $r_{J,E}$) is more appropriate. The length of the confidence interval varies the most with the target ρ and decreases with the increase of either the number targets n, or the number of raters k. Our study shows that the length of the confidence intervals varies little with $r_{T,E}$. In the absence of this information, an average over a range of plausible values of $r_{T,E}$ may be obtained.

Table 2: Number of targets (n≥k) for average length of confidence interval to be bounded by Δ

	Confidence level					
	α = 0.1			α = 0.05		
Δ	ρ = 0.6	0.7	0.8	0.6	0.7	0.8
k = 3:						
0.2	>50	48	28	>50	>50	38
0.3	32	24	14	45	32	20
0.4	18	14	14	26	20	14
k = 5:						
0.2	50	37	23	>50	>50	32
0.3	24	18	14	32	24	16
0.4	14	14	5	19	16	14
k = 7:						
0.2	44	32	20	>50	46	28
0.3	20	16	14	28	22	16
0.4	14	14	7	16	14	14
k = 10:						
0.2	40	30	20	>50	42	28
0.3	18	16	10	26	21	14
0.4	10	10	10	16	14	10

This method will result in conservative estimates of the size of the study. An R script is available from the author upon request.

Our approach is also applicable to designs based on ICC(1,1) and a table for required number of targets to achieve an expected length of a confidence interval bounded by Δ is displayed in Table 2.

We do not address in this study the problem of sample size calculation for ICC(3,1) based on confidence interval. In Case 3 the raters are modeled as fixed effects and a confidence interval for ICC(3,1) is a function of the raters' effects. Thus, to implement a similar approach one would need to elicit prior information on these effects which is not practical.

The computer time to run the simulations to determine the sample size for ICC(2,1) can be shortened if in the first step we skip the data generation and start by simulating values for BMS, JMS and EMS from properly scaled independent chi-squared distributions with appropriate degrees of freedom. The approach proposed by Beal (1989) could also be used as a criteria for sample size calculations based on confidence intervals for intra-class correlation coefficients, however we do not pursue this here. Also, in some cases investigators might be interested in powering a study to be able distinguish between two values. For example, Landis and Koch (1977) propose classifying reliability based on the magnitude of a reliability coefficient ρ as follows: ρ = 0 is defined as 'non-existing', ρ between 0 and 0.2 'slight', ρ between 0.2 and 0.4 'fair', ρ between 0.4 and 0.6 'moderate', ρ between 0.6 and 0.8 'substantial', ρ between 0.8 and 1.0 'almost perfect'. Given this classification, an investigator might be hoping for almost perfect reliability (i.e., ρ≥0.8), however he/she would want to be able to tell that the reliability is substantial (i.e., ρ≥0.6). Our approach could be adapted to this situation by using the methods by Greenland (1988).

CONCLUSION

A simple method for sample size calculation for intra-class correlation coefficient Case 1 and 2 based on the average length of confidence intervals was proposed. The proposed was implemented by the authors in R (freely available software). Three invariance properties of the confidence intervals for the intra-class correlation coefficients Case 2 were studied by simulations. These properties are an important tool when considering the design of this type of studies.

ACKNOWLEDGEMENT

Many thanks to Dr. Daniela Cojocaru and Dr. John de Figueiredo for their useful comments on the manuscript.

REFERENCES

Beal, S.L., 1989. Sample size determination for confidence intervals on the population mean and on the difference between two population means. Biometrics, 45: 969-977. http://www.ncbi.nlm.nih.gov/pubmed/2790131

Cappelleri, J. and N. Ting, 2003. A modified marge-sample approach to approximate interval estimation for a particular intraclass correlation coefficient. Stat. Med., 22: 1861-1877. http://www.ncbi.nlm.nih.gov/pubmed/12754721

Cochran, W. and G. Cox, 1957. Experimental Designs. 2nd Edn., Wiley, New York, ISBN: 0-47-116203-5, pp: 617.

Daly, L.E., 1991. Confidence intervals and sample sizes: Don't throw out all your old sample size tables. Br. Med. J., 302: 333-336. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1668990

Donner, A. and G. Wells, 1986. A comparison of confidence interval methods for the intra-class correlation coefficient. Biometrics, 42: 401-412. http://www.jstor.org/pss/2531060

Fleiss, J. and P. Shrout, 1978. Approximate interval estimation for a certain intraclass correlation coefficient. Psychometrika, 43: 259-262. http://ideas.repec.org/a/spr/psycho/v43y1978i2p259-262.html

Greenland, S., 1988. On sample-size and power calculations for studies using confidence intervals. Am. J. Epidemiol., 128: 231-237. http://cat.inist.fr/?aModele=afficheN&cpsidt=7131394

Kraemer, H., 1976. The small sample non-null properties of Kendall's coefficient of concordance for normal populations. J. Am. Stat. Assoc., 71: 608-613. http://www.jstor.org/pss/2285590

Kupper, L. and K. Hafner, 1989. How appropriate are popular sample size formulas? Am. Stat., 43: 101-105. http://www.jstor.org/pss/2684511

Lai, E., G. Doros, O. Fix and B. Jacobson, 2007. The Boston bowel preparation scale: A valid and reliable instrument for colonoscopy-oriented research. Gastrointest. Endosc., 55: 361-361. http://linkinghub.elsevier.com/retrieve/pii/S0016510707014253

Landis, J. and G. Koch, 1977. The Measurement of observer agreement for categorical data. Biometrics, 33: 159-174. http://www.ncbi.nlm.nih.gov/pubmed/843571

Maxwell, S.E., J. Rausch and K. Kelley, 2007. Sample size planning for statistical power and accuracy in parameter estimation. Annu. Rev. Psychol., 59: 537-563. http://www.ncbi.nlm.nih.gov/pubmed/17937603

McGraw, K. and S. Wong, 1996. Correction to McGraw and Wong. Psychol. Methods, 1: 390.

McGraw, K.O. and S.P. Wong, 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods, 1: 30-46. http://psycnet.apa.org/index.cfm?fa=main.doiLanding&uid=1996-03170-003

McHugh, R.B. and C.T. Le, 1984. Confidence estimation and the size of a clinical trial. Controll. Clin. Trials, 5: 157-163. http://www.ncbi.nlm.nih.gov/pubmed/6744887

O'Neill, R.T, 1984. Samples sizes for estimation of the odds ratio in unmatched casecontrol studies. Am. J. Epidemiol., 120: 145-153. http://www.ncbi.nlm.nih.gov/pubmed/6741915

Shrout, P.E. and J.L. Fleiss, 1979. Intraclass correlations: Uses in assessing rater reliability. Psychol. Bull., 86: 420-428. http://www.ncbi.nlm.nih.gov/pubmed/18839484

The R Development Core Team, 2005. A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. http://cran.r-project.org/doc/manuals/refman.pdf

Tian, L. and J. Cappelleri, 2004. A new approach for interval estimation and hypothesis testing of a certain intraclass correlation coefficient: The generalized variable method. Stat. Med., 23: 2125-2135. http://www.ncbi.nlm.nih.gov/pubmed/15211607

Walter, S.D., M. Eliasziw and A. Donner, 1998. Sample size and optimal designs for reliability studies. Stat. Med., 17: 101-110. http://www.ncbi.nlm.nih.gov/pubmed/9463853

Winer, B.J., 1991. Statistical Principles in Experimental Design. 3rd Edn., McGraw-Hill, New York, ISBN: 0070709823, pp: 1057.