Confidence intervals for intraclass correlation coefficients in variance components models

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–18

DOI: 10.1177/0962280214522787
smm.sagepub.com
SAGE

Confidence intervals for intraclass correlation coefficients in variance components models

Nino Demetrashvili,[1,2] Ernst C Wit[2] and Edwin R van den Heuvel[1,2]

Abstract

Confidence intervals for intraclass correlation coefficients in agreement studies with continuous outcomes are model-specific and no generic approach exists. This paper provides two generic approaches for intraclass correlation coefficients of the form $\sum_{q=1}^{Q}\sigma_q^2/(\sum_{q=1}^{Q}\sigma_q^2+\sum_{p=Q+1}^{P}\sigma_p^2)$. The first approach uses Satterthwaite's approximation and an *F*-distribution. The second approach uses the first and second moments of the intraclass correlation coefficient estimate in combination with a Beta distribution. Both approaches are based on the restricted maximum likelihood estimates for the variance components involved. Simulation studies are conducted to examine the coverage probabilities of the confidence intervals for agreement studies with a mix of small sample sizes. Two different three-way variance components models and balanced and unbalanced one-way random effects models are investigated. The proposed approaches are compared with other approaches developed for these specific models. The approach based on the *F*-distribution provides acceptable coverage probabilities, but the approach based on the Beta distribution results in accurate coverages for most settings in both balanced and unbalanced designs. A real agreement study is provided to illustrate the approaches.

Keywords

agreement study, ANOVA, REML, *F*-distribution, Beta distribution

1 Introduction

1.1 Motivating example

Radiotherapy is part of the treatment of patients with head and neck cancer. Before patients undergo irradiation, the organs at risk (e.g. submandibular glands) need to be contoured to deliver the radiation at the appropriate spot. However, variation in contouring is an obstacle for optimal patient treatment. To investigate this delineation process, an agreement study was conducted.[1] One aspect in this study was to assess the agreement among oncologists on

[1]Department of Epidemiology, University Medical Center Groningen, University of Groningen, RB Groningen, the Netherlands
[2]Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, AK Groningen, the Netherlands

Corresponding author:
Nino Demetrashvili, University Medical Center Groningen, University of Groningen, the Netherlands.
Email: n.demetrashvili@umcg.nl

measurements of organ volumes. Five oncologists observed each of six patients at two time points, at the beginning of treatment and approximately six months later. For one particular organ at risk, a specific three-way mixed effects model is used to describe the observed volumes

$$y_{ijk} = \mu + \alpha_i + b_j + c_k + (\alpha b)_{ij} + (\alpha c)_{ik} + (bc)_{jk} + e_{ijk} \tag{1}$$

where y_{ijk} is the volume (cm^3) of subject $j = 1, \ldots, J$ observed by oncologist $k = 1, \ldots, K$ at time point $i = 1, \ldots, I$ with μ the overall mean, α_i the fixed effect of time, $b_j \sim N(0, \sigma_S^2)$ a random effect for subject, $c_k \sim N(0, \sigma_O^2)$ a random effect for oncologist, $(\alpha b)_{ij} \sim N(0, \sigma_{TS}^2)$ an interaction effect of time and subject, $(\alpha c)_{ik} \sim N(0, \sigma_{TO}^2)$ an interaction effect of time and oncologist, $(bc)_{jk} \sim N(0, \sigma_{SO}^2)$ an interaction effect of subject and oncologist, and $e_{ijk} \sim N(0, \sigma_R^2)$ the residual. The constraint $\alpha_1 + \alpha_2 = 0$ on the fixed effects is set for identifiability. All random effects are assumed to be mutually independent. The intraclass correlation coefficient (ICC) was used to measure the agreement

$$\text{ICC} = \frac{\sigma_S^2 + \sigma_{TS}^2}{\sigma_S^2 + \sigma_O^2 + \sigma_{TS}^2 + \sigma_{TO}^2 + \sigma_{SO}^2 + \sigma_R^2} \tag{2}$$

The variability in the numerator is related to changes in the volume of the organ at risk for subjects and does not depend on observer variability. The six subjects were treated in between the two time points and the change over time was assumed to be the result of treatment. The ICC in equation (2) can be understood as the correlation coefficient between any pair of observers for one subject at the same time point, i.e. $\text{ICC} = \text{corr}(Y_{ijk_1}, Y_{ijk_2})$.

We fitted the three-way mixed effects model to the dataset (with 60 observations) for the left submandibular gland and estimated the variance components (see Table 1). The estimate for the ICC is $\widehat{\text{ICC}} = 0.61$, as it was earlier reported.[1] However, a confidence interval on this estimate was not reported.

A literature search on confidence intervals for ICCs demonstrated that this topic is mainly discussed for one-way and two-way random or mixed effects models. For our three-way mixed effects model we could not find an approach. Moreover, there is no closed-form generic approach that could handle any type of variance components models for balanced and unbalanced designs. In the remaining part of the paper we will provide and evaluate two generic closed-form approaches for constructing confidence intervals for ICCs that have the form (2).

1.2 Background

The concept of ICC originated from genetics, in which it was used to judge the correlation among family members with respect to biological characteristics.[2–4] This concept was extended to social and medical sciences. Initially, the ICC was introduced for categorical measurements.[5–7] Later, this

Table 1. Restricted maximum likelihood estimates of variance components in three-way mixed effects model.

$\hat{\sigma}_S^2$	$\hat{\sigma}_O^2$	$\hat{\sigma}_{TS}^2$	$\hat{\sigma}_{TO}^2$	$\hat{\sigma}_{SO}^2$	$\hat{\sigma}_R^2$
1.409	0.755	2.946	0.488	0.655	0.927

concept was discussed in the context of continuous measurements.[8,9] Several other authors provided different forms of ICCs for one-way and two-way variance components models.[10–12] For these variance components models, various approaches on the construction of confidence intervals were established.[13] In generalizability theory[14–16] higher-order variance components models are typically used, but these studies investigate reliability measures for average values of items in questionnaires instead of individual observations. To our knowledge, only limited work has been conducted on confidence intervals on these reliability indices.

One challenge in the construction of a confidence interval for the ICC in equation (2) is the estimation method. Variance components can be estimated[17] via the method of moments (MM), maximum likelihood (ML), and restricted maximum likelihood (REML). The MM allows negative estimates of variance components, possibly leading to negative ICCs. Albeit theoretically acceptable, negative ICCs in the form of equation (2) are meaningless. This issue can be resolved by using the ML or REML estimation method since the negative estimates of variance components from the MM essentially are substituted by zero. However, this leads to another problem. The distribution of the ICC is unknown when all variance components in the numerator of the ICC are at the boundary of the parameter space, making it difficult or impossible to construct confidence intervals. For balanced one-way random effects models an exact probability of a zero estimate of the between group variance component was provided.[17] From this probability it is shown that the REML estimates are less likely to lie on the boundary of the parameter space than the ML estimates. This and other discussed properties,[17,18] justify the use of REML as the preferred method of estimation.

Another challenge is the lack of an exact (closed-form) method for the construction of confidence intervals for ICCs in general. Exact confidence intervals are given for one-way random effects models with equal group sizes[19] and unequal group sizes.[20] For balanced designs, under assumptions of normality and independence of the random effects, the ICC for one-way random effects models using the mean squares results into a function of a random variable that has a central F-distribution. Such form does not hold true for unbalanced and higher-order models using analysis of variance (ANOVA). The lack of such a general distributional form for ICCs makes the use of exact methods for construction of confidence intervals impossible or at least complicated. The existing approach[20] for unbalanced one-way random effects models is exact, but it is computationally challenging. Such difficulty led to construction of approximate closed-form confidence intervals for unbalanced one-way random effects models. There are essentially two principles: one that uses an approximation of the F-statistic and another that uses large sample approximations to the variance of the estimator of ICC. These approximate methods have been compared[21,22] for different settings. Focusing on cluster randomized controlled trials,[22] the methods based on the F-statistic provide better coverage than those based on large sample approximations. Focusing on family studies,[21] the approach based on large sample variance approximations[23] was preferred.

For balanced two-way random and mixed effects models, most methods[9,24] use some form of Satterthwaite's approximation[25] with the chi-square distribution for linear combinations of mean squares. However, challenges are associated with Satterthwaite's approximation. If the estimated degrees of freedom of independent chi-squared random variables differ greatly, then Satterthwaite's approach can produce liberal confidence intervals (see Burdick and Graybill[13], pp. 29–30). Two alternative approaches have been proposed. One of them involves the confidence intervals of ICC for a two-factor nested design in large samples.[26] Another alternative is the modified large-sample (MLS) approach, which was proposed[27] for confidence intervals of nonnegative linear combinations of the variance components. In this approach, the confidence limits under large-sample normal

theory are modified so that they become exact for small or moderate sample sizes. Later, an extension of the MLS-type approach was proposed[28] on linear combinations of variance components that are unrestricted in sign. The MLS approach was extended[29,30] to ICC, but none of these approaches investigated ICCs for three-way mixed effects models. As shown earlier,[30,31] the MLS-type approach is preferred over Satterthwaite's bounds because they better maintain the stated confidence level. However, the MLS-type approach is based on the MM estimates instead of the preferred REML method.

Recently, an approximate confidence interval was proposed[32] on a particular ICC for a balanced three-way random effects model using the MM estimates. Their approach is similar to Satterthwaite's methods used in two-way random and mixed effects models. The approach based on Satterthwaite's approximation is a general principle, but requires tedious calculations every time another model is needed. Furthermore, it is not always applicable for unbalanced designs, since the mean squares are not uniquely defined. In unbalanced design, none of the above described methods are general enough to be used in any type of variance components model.

We propose two general, REML-based closed-form approaches to construct an approximate confidence interval on the ICC of interest for a continuous response. These methods are applicable to any variance components model with balanced or unbalanced designs under assumption of normality. The first approach assumes that the ICC is a ratio of sums of jointly independent chi-squared distributed random variables. Consequently, it is approximated with a function of an F-distributed variable using Satterthwaite's chi-square approximation. This approach is somewhat similar to the existing approaches for balanced one-way and two-way random effects models. The second approach is based on an approximation of the ICC with a Beta distribution. Such approximation is based on various theoretical ideas and solves the challenges associated with Satterthwaite's approximation.

2 Generic closed-form methods for approximate confidence intervals on ICCs

Consider a variance components model for a particular agreement study, with mixed or only random effects, where one subset $q = 1, \ldots, Q$ of variance components $\sigma_1^2, \sigma_2^2 \ldots, \sigma_Q^2$ is unrelated to the observer process and another disjoint subset $p = Q+1, \ldots, P$ of variance components $\sigma_{Q+1}^2, \sigma_{Q+2}^2 \ldots, \sigma_P^2$ does represent the part of the observer variability. If we define the combined, observer unrelated and related variance components by $\sigma_G^2 = \sum_{q=1}^{Q} \sigma_q^2$ and $\sigma_E^2 = \sum_{p=Q+1}^{P} \sigma_p^2$, respectively, then the ICC in its general form can be defined as

$$\mathrm{ICC} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2} \tag{3}$$

The ICC in equation (3) has the form of the ICC for a one-way random effects model, where $P = 2$ and $Q = 1$.[13] It also has the form of the ICC in our motivating example (2). The ICC represents the correlation coefficient of two particular observations from the variance components model, but there are other forms of ICCs,[11,12] that do not fit our general form.

Let the estimates of the variance components $\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_P^2$ be obtained by REML. Summing up the respective variance component estimates results in the variance component estimates of the unrelated $\hat{\sigma}_G^2$ and related observer variability $\hat{\sigma}_E^2$. An estimate of the ICC is now obtained by substituting the variance component estimates in equation (3). The variance–covariance estimates of the variance components estimates $\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_P^2$ are based on the Fisher information matrix and

can be obtained easily with several software packages. Denote the estimated standard errors for the variance component estimates by $\hat{\tau}_1, \ldots, \hat{\tau}_P$ and let the estimated covariances be denoted by $\hat{\tau}_{1,2}, \ldots, \hat{\tau}_{1,P}, \hat{\tau}_{2,3}, \ldots, \hat{\tau}_{P-1,P}$. The variances of $\hat{\sigma}^2_G$ and $\hat{\sigma}^2_E$ are respectively determined by

$$\begin{aligned} \tau^2_{\hat{\sigma}^2_G} &= \sum_{q=1}^{Q} \tau_q^2 + 2 \sum_{q=1}^{Q-1} \sum_{r=q+1}^{Q} \tau_{qr} \\ \tau^2_{\hat{\sigma}^2_E} &= \sum_{p=Q+1}^{P} \tau_p^2 + 2 \sum_{p=Q+1}^{P-1} \sum_{r=p+1}^{P} \tau_{pr} \end{aligned} \tag{4}$$

An approximate variance of the $\widehat{ICC}$ can be obtained through a first-order Taylor expansion as shown below

$$\hat{\tau}^2_{\widehat{ICC}} \approx \frac{\hat{\sigma}^4_E}{(\hat{\sigma}^2_G + \hat{\sigma}^2_E)^4} \hat{\tau}^2_{\hat{\sigma}^2_G} + \frac{\hat{\sigma}^4_G}{(\hat{\sigma}^2_G + \hat{\sigma}^2_E)^4} \hat{\tau}^2_{\hat{\sigma}^2_E} - \frac{2\hat{\sigma}^2_G \hat{\sigma}^2_E}{(\hat{\sigma}^2_G + \hat{\sigma}^2_E)^4} \hat{\tau}_{\hat{\sigma}^2_G \hat{\sigma}^2_E} \tag{5}$$

with $\hat{\tau}_{\hat{\sigma}^2_G \hat{\sigma}^2_E}$ being the covariance of the estimators $\hat{\sigma}^2_G$ and $\hat{\sigma}^2_E$. Note that we have used the derivatives of the ICC with respect to the variance components σ^2_G and σ^2_E, i.e. the partial derivatives are $\partial(ICC)/\partial(\sigma^2_G) = \sigma^2_E/(\sigma^2_G + \sigma^2_E)^2$ and $\partial(ICC)/\partial(\sigma^2_E) = -\sigma^2_G/(\sigma^2_G + \sigma^2_E)^2$.

The variance of the $\widehat{ICC}$ for unbalanced one-way random effects model was discussed earlier.[33] It is different from our variance, although both use a first-order Taylor approximation. In earlier work,[33] the ICC was written as a ratio of two different linear combinations of the within and between sums of squares. This is another representation of the ICC estimate for the one-way random effects model but such form is more difficult to generalize to higher-order models. Furthermore, this representation typically uses moment estimates, while we selected the REML estimates for the computation of the ICC. Constructing the confidence intervals on the ICC using the normal distribution and the variance in equation (5) would not take into account that the distribution of $\widehat{ICC}$ is skewed. Thus, instead of using the normal distribution, the distribution of $\widehat{ICC}$ should be approximated with a skewed distribution. We will use the *F*- and Beta distributions.

2.1 *F*-approach

The *F*-approach is based on Satterwhaite's approximation of the sum of variance components $\hat{\sigma}^2_G$ and $\hat{\sigma}^2_E$ separately. In balanced variance components models, these variance components are linear combinations of weighted independent chi-square distributed variables. Thus, Satterthwaite's approach is indeed applicable even though it originated on linear combinations of mean squares.[9] It determines the degrees of freedom df_G and df_E such that $df_G \hat{\sigma}^2_G/\sigma^2_G \sim \chi^2_{df_G}$ and $df_E \hat{\sigma}^2_E/\sigma^2_E \sim \chi^2_{df_E}$. Essentially, it provides the numbers of degrees of freedom such that the first two moments of the variance components estimates $\hat{\sigma}^2_G$ and $\hat{\sigma}^2_E$ are equal to the first two moments of a chi-square distributed random variable. The degrees of freedom are given by

$$\begin{aligned} df_G &= 2(\mathrm{E}\hat{\sigma}^2_G)^2/\tau^2_{\hat{\sigma}^2_G} \\ df_E &= 2(\mathrm{E}\hat{\sigma}^2_E)^2/\tau^2_{\hat{\sigma}^2_E} \end{aligned} \tag{6}$$

with E the expected value, $\tau^2_{\hat{\sigma}^2_G}$ the variance of $\hat{\sigma}^2_G$, and $\tau^2_{\hat{\sigma}^2_E}$ the variance of $\hat{\sigma}^2_E$. The degrees of freedom in equation (6) are estimated by substituting the numerators with $\hat{\sigma}^4_G$ and $\hat{\sigma}^4_E$, respectively, and

replacing the variances $\tau^2_{\hat{\sigma}^2_G}$ and $\tau^2_{\hat{\sigma}^2_E}$ with an appropriate estimate. It was shown[34] that the approach of Satterthwaite applied to the sums of variance components works better when these variance estimates do not include the covariance terms $\hat{\tau}_{1,2}, \ldots, \hat{\tau}_{1,P}, \hat{\tau}_{2,3}, \ldots, \hat{\tau}_{P-1,P}$. Thus, following earlier work,[34] the estimated variances in equation (6) for the F-approach are not of the form (4), but are given by

$$\hat{\tau}^2_G = \sum_{q=1}^{Q} \hat{\tau}^2_q, \quad \hat{\tau}^2_E = \sum_{q=Q+1}^{P} \hat{\tau}^2_q \tag{7}$$

The estimated numbers of degrees of freedom are then given by $\widehat{df}_G = 2\hat{\sigma}^4_G/\hat{\tau}^2_G$ and $\widehat{df}_E = 2\hat{\sigma}^4_E/\hat{\tau}^2_E$, respectively. If we now further assume that the two estimates $\hat{\sigma}^2_G$ and $\hat{\sigma}^2_E$ are also independent, then the estimate $\widehat{\text{ICC}}$ is approximately distributed according to

$$\frac{\sigma^2_G F_{\widehat{df}_G, \widehat{df}_E}}{\sigma^2_G F_{\widehat{df}_G, \widehat{df}_E} + \sigma^2_E} \tag{8}$$

where $F_{N,D}$ is a random variable having an F-distribution with N and D degrees of the freedom for the numerator and denominator, respectively. Based on this approximate distribution, the symmetric approximate $100\%(1-\alpha)$ confidence interval on the ICC in equation (3) is given by the lower and upper confidence limits of the form

$$\begin{aligned} LCL_F &= \frac{\hat{\sigma}^2_G F^{-1}_{\widehat{df}_G, \widehat{df}_E}(\alpha/2)}{\hat{\sigma}^2_G F^{-1}_{\widehat{df}_G, \widehat{df}_E}(\alpha/2) + \hat{\sigma}^2_E} \\ UCL_F &= \frac{\hat{\sigma}^2_G F^{-1}_{\widehat{df}_G, \widehat{df}_E}(1-\alpha/2)}{\hat{\sigma}^2_G F^{-1}_{\widehat{df}_G, \widehat{df}_E}(1-\alpha/2) + \hat{\sigma}^2_E} \end{aligned} \tag{9}$$

with $F^{-1}_{N,D}(q)$ the q^{th} quantile of an F-distribution with N and D degrees of freedom for the numerator and denominator, respectively.

The estimated number of degrees of freedom $\widehat{df}_G$ can in principle be close to zero. To avoid computational issues for $F^{-1}_{\widehat{df}_G, \widehat{df}_E}$ when $\widehat{df}_G$ is very small, we avoid values of $\widehat{df}_G$ below one, by taking $\widehat{df}_G = \max(1, 2\hat{\sigma}^4_G/\hat{\tau}^2_G)$.[34] The construction of a confidence interval on the ICC when the estimate σ^2_G is equal to zero is postponed to Section 2.3.

The F-approach in equation (9) does not result into the exact method for balanced one-way random effects models. The exact method[19] uses the ratio of the mean squares of the between groups and the within groups. The confidence limits on the ICC for exact method are written in the form $(F/F_L - 1)/(F/F_L + n - 1)$ instead of equation (9), with n being the number of subjects within each group and F_L a quantile of the F-distribution having $m - 1$ and $m(n-1)$ degrees of freedom (m is the number of groups).

2.2 Beta-approach

In this approach, we approximate the distribution of the ICC estimate with a Beta distribution. There are several rationales behind such an approximation. The first argument is that the support

of the Beta distribution coincides with the range of the ICC $\in [0, 1]$.[35] Secondly, the Beta distribution can assume different shapes and in particular it can be highly skewed, both to the right when the ICC estimate is close to zero or to the left when the ICC estimate is close to one. The final and most important argument for the use of a Beta distribution is based on a familiar theorem (see Knight[36], p. 64). It states that the ratio $X_1/(X_1 + X_2)$ has a Beta distribution with parameters $a > 0$ and $b > 0$, when X_1 and X_2 are independent, X_1 has a gamma distribution with parameters $a > 0$ and $c > 0$, and X_2 has a gamma distribution with parameters $b > 0$ and $c > 0$. The ICC estimate (3) has the same form as $X_1/(X_1 + X_2)$, although the estimates $\hat{\sigma}_G^2$ and $\hat{\sigma}_E^2$ might not be independent and not necessarily gamma distributed. On the other hand, in the previous section we approximated the distribution of these variance component estimates with chi-square distributions, which are specific gamma distributions. The extension to gamma distributions seems a logical step.

The mean and variance of the Beta distribution with parameters $a > 0$ and $b > 0$ are given by

$$\begin{aligned} \mu &= \frac{a}{a+b} \\ \sigma^2 &= \frac{ab}{(a+b)^2(a+b+1)} \end{aligned} \tag{10}$$

If the mean and variance of the Beta distribution are estimated from the data, $\hat{\mu}$ and $\hat{\sigma}^2$, then the MM estimates for a and b are

$$\begin{aligned} \hat{a} &= \frac{\hat{\mu}[\hat{\mu}(1-\hat{\mu}) - \hat{\sigma}^2]}{\hat{\sigma}^2} \\ \hat{b} &= \frac{(1-\hat{\mu})[\hat{\mu}(1-\hat{\mu}) - \hat{\sigma}^2]}{\hat{\sigma}^2} \end{aligned} \tag{11}$$

To approximate the distribution of the ICC estimate, the mean μ will be estimated with $\hat{\mu} = \widehat{\text{ICC}}$ and the variance σ^2 will be estimated with the approximate variance of $\widehat{\text{ICC}}$ given in equation (5).

The Beta distribution is not always unimodal.[35] When both parameters a and b are below one, the beta density has a *U*-shaped form. This form seems less appropriate for the approximation of the distribution of the ICC estimate. This means that we exclude parameter estimates that are both below one. Thus, whenever the ICC $= a/(a + b)$ is below 0.5, we would like to use the mass of the beta density on the left tail and we use a decreasing density in ICC values. Thus, to avoid the *U*-shaped density, we change it into a *J*-shaped density with more mass closer to zero by changing the parameter b to one ($b = 1$), but keeping the parameter a at its value below one. Alternatively, when the ICC is larger than 0.5, a is increased to one ($a = 1$) and b remains at its value below one. These approaches are illustrated in the left (ICC ≤ 0.5) and right (ICC > 0.5) parts of Figure 1. It should be noted that we do allow parameter estimates that would result in J-shaped Beta densities, i.e. $(\hat{a} - 1)(\hat{b} - 1)$ can be negative.

The parameters a and b in the Beta distribution are positive, but the estimates in equation (11) can become negative when the approximate variance of the ICC estimate in equation (5) is larger than $\widehat{\text{ICC}}(1 - \widehat{\text{ICC}})$. This means that we cannot estimate the Beta distribution with our approach in equation (11). In such case we set the estimate for b equal to one ($\hat{b} = 1$) when the ICC estimate is smaller than or equal to 0.5. The estimate for a is then obtained by $\hat{a} = \widehat{\text{ICC}}/(1 - \widehat{\text{ICC}})$, which is the

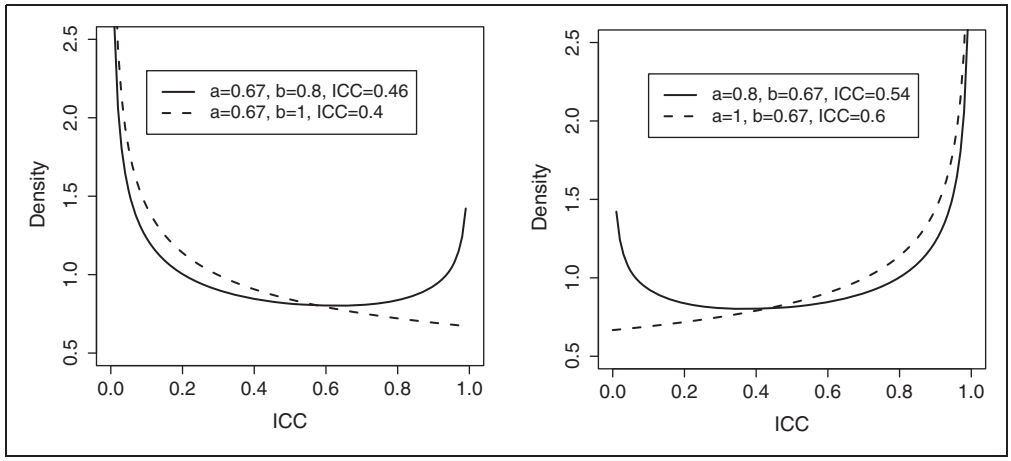

Figure 1. Beta densities. Left: $\widehat{ICC} \leq 0.5$; Right: $\widehat{ICC} > 0.5$.

first moment estimate when $b = 1$. Alternatively, the estimate for a is set equal to one ($\hat{a} = 1$) when the ICC estimate is larger than 0.5. The estimate for b then becomes $\hat{b} = (1 - \widehat{ICC})/\widehat{ICC}$.

Based on these choices of parameter estimates, the quantiles of the Beta distribution are obtained. Thus, the approximate 100% $(1 - \alpha)$ confidence interval on the ICC in equation (3) using the Beta distribution is given by the lower and upper confidence limits

$$\begin{aligned} LCL_B &= B^{-1}_{\hat{a},\hat{b}}(\alpha/2) \\ UCL_B &= B^{-1}_{\hat{a},\hat{b}}(1 - \alpha/2) \end{aligned} \tag{12}$$

The calculation of the quantiles of the Beta distribution involves some numerical issues that should be addressed. When one of the parameters gets close to zero many software packages cannot calculate the quantiles. To address this numerical issue, we never use parameter estimates below the value 0.01. This choice is pragmatic and suitable for most cases. Indeed, for $\hat{a} = 0.01$ and $\hat{b} = 1$, the quantile $B^{-1}_{\hat{a},\hat{b}}(0.025) \approx 0$ and for $\hat{a} = 1$ and $\hat{b} = 0.01$, the quantile $B^{-1}_{\hat{a},\hat{b}}(0.975) \approx 1$.

2.3 Zero ICC estimate

The two generic approaches discussed in Sections 2.1 and 2.2 use the standard errors of the variance components estimates in their own way to be able to approximate the distribution of the ICC estimate. However, the standard error of the variance component estimate $\hat{\sigma}_G$ does not exist or is equal to zero when this variance component estimate is zero. This is a consequence of our choice of REML estimation procedure. When the REML estimate leads to a zero estimate of the ICC, the proposed two approaches are not applicable any more.

The confidence limits $\hat{L}_+$ and $\hat{U}_+$ of the two approaches in equations (9) and (12) were actually developed for confidence intervals for the ICC conditionally on $\hat{\sigma}_G > 0$. Thus, these limits $\hat{L}_+$ and $\hat{U}_+$ should approximately satisfy $P(\hat{L}_+ \leq \text{ICC} \leq \hat{U}_+ | \hat{\sigma}_G > 0) = 1 - \alpha$. Although this conditional probability may be of interest on its own, as we will show, we are also interested in constructing confidence intervals for the values of ICC in all settings, including confidence limits for zero ICC estimates. This means that we need to construct an upper confidence limit $\hat{U}_0$ that would

approximately satisfy $P(\text{ICC} \leq \hat{U}_0 | \hat{\sigma}_G = 0) = 1 - \alpha$. The lower confidence limit $\hat{L}_0$ in this case is of course set equal to zero. If we now define the confidence limits $\hat{L}$ and $\hat{U}$ by $\hat{L} = I(\hat{\sigma}_G > 0)\hat{L}_+$ and $\hat{U} = I(\hat{\sigma}_G = 0)\hat{U}_0 + I(\hat{\sigma}_G > 0)\hat{U}_+$, with $I(A)$ the indicator variable equal to one when A is true and zero otherwise, we obtain the marginal confidence level

$$P(\hat{L} \leq \text{ICC} \leq \hat{U}) = P(\hat{L}_+ \leq \text{ICC} \leq \hat{U}_+ | \hat{\sigma}_G > 0) P(\hat{\sigma}_G > 0) + P(\text{ICC} \leq \hat{U}_0 | \hat{\sigma}_G = 0) P(\hat{\sigma}_G = 0) \approx 1 - \alpha \tag{13}$$

The choice for $\hat{U}_0$ is pragmatic. It is based on an approach borrowed from the balanced one-way random effects model. The same strategy is used for both generic approaches.

In balanced one-way random effects models, with m groups and n observations within groups, the probability of obtaining a non-positive estimate for σ_G based on the mean squares is equal to $P(\hat{\sigma}_G^2 \leq 0) = P(F_{m-1,m(n-1)} \leq \sigma_E^2/(n\sigma_G^2 + \sigma_E^2))$, with $F_{N,D}$ having a F-distribution.[17] The larger the value for σ_G the less likely the estimate $\hat{\sigma}_G$ is non-positive. The biggest value for σ_G^2 that is still likely to occur is obtained by choosing σ_G such that $P(\hat{\sigma}_G^2 \leq 0) = \alpha$, with $1 - \alpha$ the confidence level. This results into the equality $\sigma_G^2 = \sigma_E^2(1 - F^{-1}_{m-1,m(n-1)}(\alpha))/(nF^{-1}_{m-1,m(n-1)}(\alpha))$. Substituting this value in equation (8) and replacing the degrees of freedom $m - 1$ and $m(n - 1)$ by $\widehat{df}_G$ and $\widehat{df}_E$, respectively, the upper limit $\hat{U}_0$ becomes

$$\hat{U}_0 = \frac{1 - F^{-1}_{\widehat{df}_G, \widehat{df}_E}(\alpha)}{1 + (n-1)F^{-1}_{\widehat{df}_G, \widehat{df}_E}(\alpha)} \tag{14}$$

The degrees of freedom $\widehat{df}_G$ for the variance component estimate $\hat{\sigma}_G^2$ is unknown in two-way and higher-order variance component models when it consists of more than one variance component and it is difficult to determine from the observed data when the estimate is equal to zero.[34] Since $\hat{\sigma}_E$ would typically include the residual term of the variance components model, the degrees of freedom $\widehat{df}_E$ can always be determined using Satterthwaite's approach, as discussed in Section 2.1. We suggest the use of $\widehat{df}_G = 1$ when the ICC estimate is equal to zero, which makes $n - 1$ in equation (14) equal to $\widehat{df}_E/2$ in one-way random effects models. These choices for degrees of freedom and for $n - 1$ in equation (14) are proposed for any balanced or unbalanced variance components models.

Finally, it should be noted that the estimate of the degrees of freedom $\widehat{df}_G$ is less reliable when the estimate $\hat{\sigma}_G$ is close to zero, since small variations lead to unrealistic high numbers. This means that numerically an estimate of the ICC close to zero, but still positive, may be considered equivalent to a zero ICC estimate. In our approach, we have chosen a value of the ICC estimate to be equal to zero when it is smaller than 0.01. This seems realistic for many practical situations.

3 Motivating example (continued)

Recall that the motivating example investigates the agreement between oncologists on the volumes of the head and neck organs. There are five oncologists who observed the left submandibular gland of six patients at two time points. The REML estimates of the variance components are provided in Table 1. Following Section 2, the two variance components σ_G^2 and σ_E^2 are estimated by $\hat{\sigma}_G^2 = 1.409 + 2.946 = 4.355$ and $\hat{\sigma}_E^2 = 0.755 + 0.488 + 0.655 + 0.927 = 2.825$, respectively. The ICC is then determined: $\widehat{\text{ICC}} = 0.61$. To compute the confidence intervals using our generic methods we also need the variances and covariances of the variance components given in Table 1. We obtain these estimates using the MIXED procedure of SAS (version 9.2) and they are provided in Table 2.

Table 2. Estimates of variance–covariance of the estimates of variance components in three-way mixed effects model.

	$\hat{\tau}_{\hat{\sigma}_S^2}$	$\hat{\tau}_{\hat{\sigma}_O^2}$	$\hat{\tau}_{\hat{\sigma}_{TS}^2}$	$\hat{\tau}_{\hat{\sigma}_{TO}^2}$	$\hat{\tau}_{\hat{\sigma}_{SO}^2}$	$\hat{\tau}_{\hat{\sigma}_R^2}$
$\hat{\tau}_{\hat{\sigma}_S^2}$	4.845	0.005	−1.962	−0.001	−0.029	0.009
$\hat{\tau}_{\hat{\sigma}_O^2}$	0.005	0.759	−0.001	−0.104	−0.024	0.007
$\hat{\tau}_{\hat{\sigma}_{TS}^2}$	−1.962	−0.001	3.925	0.003	0.009	−0.017
$\hat{\tau}_{\hat{\sigma}_{TO}^2}$	−0.001	−0.104	0.003	0.209	0.007	−0.014
$\hat{\tau}_{\hat{\sigma}_{SO}^2}$	−0.029	−0.024	0.009	0.007	0.146	−0.043
$\hat{\tau}_{\hat{\sigma}_R^2}$	0.009	0.007	−0.017	−0.014	−0.043	0.086

Table 3. Approximate 95% confidence intervals for ICC in three-way mixed effects model.

Method	*LCL*	*UCL*	Width
F	0.165	0.857	0.692
Beta	0.311	0.863	0.552

For the *F*-approach, from Table 2 we can determine the standard errors of the variance components $\hat{\sigma}_G^2$ and $\hat{\sigma}_E^2$. We only used the standard errors and did not use the covariances among the individual variance component estimates. The standard errors are determined as $\hat{\tau}_G^2 = 4.845 + 3.925 = 8.77$ and $\hat{\tau}_E^2 = 0.750 + 0.146 + 0.209 + 0.086 = 1.191$. The numbers of degrees of freedom are now computed to be $\widehat{df}_G = 2(4.355)^2/8.77 = 4.325$ and $\widehat{df}_E = 2(2.825)^2/1.191 = 13.402$ for the variance component estimates $\hat{\sigma}_G^2$ and $\hat{\sigma}_E^2$, respectively. The confidence limits for the *F*-approach can now be calculated from equation (9) and the results are presented in Table 3. For the Beta-approach we determine the variance of the ICC estimate given in equation (5): $\hat{\tau}^2_{\widehat{\text{ICC}}} = 0.021$ where $\hat{\tau}_{\hat{\sigma}_G^2\hat{\sigma}_E^2} = -0.024$. Now using equation (11) the parameter estimates of the Beta distribution become $\hat{a} = 6.324$ and $\hat{b} = 4.102$. Then the confidence limits are obtained with equation (12), and they are provided in Table 3.

Apparently, the Beta-approach yields a narrower confidence interval. If both approaches provide 95% coverage, then the Beta-approach will be recommended. To investigate what the coverage probabilities are for our two generic approaches we carried out a simulation study.

4 Simulation designs and results

The simulation study consists of three parts. The first part investigates coverage probabilities of the two generic approaches for several settings of the three-way mixed effects model used in our agreement study (equation (1)). The second part compares our generic approaches to the approach developed for a three-way random effects model.[32] The third part compares our generic approaches with three alternative approaches developed for balanced and unbalanced one-way random effects model. In all studies, the number of simulations is 10000.

We investigated both, the marginal $P(\hat{L} \leq \text{ICC} \leq \hat{U})$ and conditional $P(\hat{L}_+ \leq \text{ICC} \leq \hat{U}_+ | \hat{\sigma}_G > 0)$ coverage probabilities for all settings. We report only marginal coverage probabilities, since the conditional coverage probabilities are very close to the marginal ones. For larger values of the ICC

Table 4. Variance component parameters in three-way mixed effects model.

	Design 1											
ICC	σ_S^2	σ_O^2	σ_{TS}^2	σ_{TO}^2	σ_{SO}^2	σ_R^2						
0.1	0.1	1.1	0.4	1.2	1.2	1.0						
0.2	0.4	0.8	1.0	1.9	1.9	1.0						
0.3	0.7	1.3	1.4	1.3	1.3	1.0						
0.4	0.4	0.3	1.8	1.0	1.0	1.0						
0.5	1.4	0.3	1.4	0.9	0.6	1.0						
0.6	1.5	0.8	3.0	0.5	0.7	1.0						
0.7	0.8	0.2	4.8	0.8	0.4	1.0						
0.8	4.0	0.29	1.5	0.04	0.04	1.0						
0.9	0.3	0.02	1.5	0.06	0.02	0.1						
	Design 2						Design 3					
	σ_S^2	σ_O^2	σ_{TS}^2	σ_{TO}^2	σ_{SO}^2	σ_R^2	σ_S^2	σ_O^2	σ_{TS}^2	σ_{TO}^2	σ_{SO}^2	σ_R^2
0.1	0.4	1.1	0.1	1.2	1.2	1.0	0.2	1.6	0.4	1.4	1.4	1.0
0.2	0.4	1.9	1.0	0.8	1.9	1.0	0.4	1.9	1.0	1.9	0.8	1.0
0.3	1.4	1.3	0.7	1.3	1.3	1.0	0.7	0.6	1.4	2.0	1.3	1.0
0.4	0.4	0.1	1.8	1.1	1.1	1.0	0.4	1.5	1.8	0.4	0.4	1.0
0.5	0.9	0.3	1.9	0.9	0.6	1.0	0.4	0.3	2.4	0.9	0.6	1.0
0.6	0.7	0.1	1.3	0.1	0.1	1.0	1.7	0.4	1.6	0.4	0.4	1.0
0.7	0.8	0.4	4.8	0.6	0.4	1.0	1.9	0.2	3.7	0.6	0.6	1.0
0.8	4.0	0.35	1.5	0.01	0.01	1.0	2.9	0.5	1.0	0.1	0.17	0.2
0.9	0.6	0.1	3.0	0.1	0.1	0.1	0.9	0.05	4.5	0.2	0.15	0.2

this is not surprising, since we hardly ever get an ICC estimate equal to zero. For smaller values, we did observe zero ICC estimates, but still the marginal and conditional coverage probabilities were close. This indicates that our upper confidence limit for zero ICC estimates seems to work reasonably well.

4.1 Three-way mixed effects model

The choices of variance components σ_S^2, σ_O^2, σ_{TS}^2, σ_{TO}^2, σ_{SO}^2, σ_R^2 for our three-way mixed effects model are given in Table 4. One of these settings are motivated by our example (ICC = 0.6 in design 1 of Table 4). The mean values of the fixed effect of time are irrelevant and we set them equal to zero and one for the first and second time points respectively. We consider several settings on the triplet (*I, J, K*) of sample sizes, where K is the number of observers evaluating each of J subjects at I time points: (2, 6, 3), (2, 6, 5), (2, 10, 5), (2, 10, 10), (2, 25, 5), and (2, 25, 12). These triplets are combined with each of the design settings in Table 4 to conduct the simulations.

We simulated the linear mixed effects model in equation (1) using the following formula $y_{ijk} = \mu_i + \sigma_S z_j + \sigma_O z_k + \sigma_{TS} z_{ij} + \sigma_{TO} z_{ik} + \sigma_{SO} z_{jk} + \sigma_R z_{ijk}$, with z_j, z_k, z_{ij}, z_{ik}, z_{jk}, and z_{ijk} mutually independently distributed from a standard normal distribution. The fixed and random effects were generated in three different data steps. The first data step generated the random effect of z_j, the random interaction effect of subjects with operators z_{jk}, the residuals z_{ijk}, and the means at the first and second time point respectively, μ_1 and μ_2. The second data step generated the random effect of observers z_k and the interaction effect of observers with time z_{ik}. The third and final data step

generated the random interaction effect between subjects and time z_{ij}. Then the data sets were merged to be able to determine the observations y_{ijk} of the three-way linear mixed effects model.

The marginal coverage probabilities (CP, %) for three-way mixed effects model under design 1 are provided in Table 5. We present the coverages for the two-sided 95% confidence intervals [LCL; UCL] and the one-sided lower 97.5% confidence intervals [LCL; 1]. The results for the other designs are not presented here since the coverages are very close to the ones shown in Table 5.

The Beta-approach results in slightly liberal coverage probabilities for the two-sided confidence intervals with small number of subjects (J). This is more prominent when the number of observers (K) increases with respect to the number of subjects. However, the coverage improves with an increase in the number of subjects. The F-approach results in conservative coverages in almost all settings. Overall, the Beta-approach performs the best. It gives quite accurate two-sided confidence intervals and provides relatively good symmetry over the whole range of ICC for the triplets with small sample sizes (≤ 10). The same is true for the mix of relatively larger sample sizes, e.g. (2, 25, 5),

Table 5. Marginal coverage probabilities for two-sided 95% confidence intervals and one-sided lower 97.5% confidence intervals in three-way mixed effects model.

				$CP_{[LCL;UCL]}$		$CP_{[LCL;1]}$						$CP_{[LCL;UCL]}$		$CP_{[LCL;1]}$	
I	J	K	ICC	F	Beta	F	Beta	I	J	K	ICC	F	Beta	F	Beta
2	6	3	0.1	0.982	0.973	0.990	0.984	2	10	10	0.1	0.968	0.949	0.997	0.992
			0.2	0.971	0.951	0.994	0.980				0.2	0.966	0.937	0.999	0.991
			0.3	0.975	0.942	0.995	0.975				0.3	0.977	0.947	0.999	0.989
			0.4	0.975	0.949	0.996	0.972				0.4	0.973	0.946	0.999	0.985
			0.5	0.979	0.949	0.998	0.972				0.5	0.972	0.939	0.999	0.985
			0.6	0.987	0.947	0.997	0.961				0.6	0.970	0.944	0.999	0.980
			0.7	0.984	0.944	0.996	0.958				0.7	0.975	0.950	0.999	0.980
			0.8	0.972	0.945	0.999	0.962				0.8	0.950	0.940	0.999	0.980
			0.9	0.987	0.946	0.997	0.952				0.9	0.977	0.955	0.999	0.975
2	6	5	0.1	0.979	0.983	0.998	0.992	2	25	5	0.1	0.977	0.952	0.987	0.977
			0.2	0.968	0.949	0.998	0.990				0.2	0.983	0.948	0.989	0.975
			0.3	0.970	0.942	0.999	0.988				0.3	0.983	0.946	0.985	0.965
			0.4	0.967	0.948	0.999	0.984				0.4	0.990	0.953	0.995	0.971
			0.5	0.966	0.939	0.999	0.983				0.5	0.987	0.951	0.990	0.968
			0.6	0.974	0.943	0.999	0.973				0.6	0.988	0.937	0.990	0.947
			0.7	0.975	0.949	0.999	0.974				0.7	0.990	0.948	0.993	0.957
			0.8	0.952	0.938	0.999	0.973				0.8	0.980	0.954	0.993	0.968
			0.9	0.976	0.951	0.999	0.965				0.9	0.992	0.944	0.994	0.949
2	10	5	0.1	0.979	0.970	0.993	0.987	2	25	12	0.1	0.980	0.949	0.996	0.985
			0.2	0.977	0.948	0.996	0.986				0.2	0.988	0.951	0.997	0.983
			0.3	0.984	0.945	0.996	0.980				0.3	0.989	0.948	0.997	0.980
			0.4	0.978	0.943	0.998	0.980				0.4	0.988	0.952	0.998	0.981
			0.5	0.980	0.945	0.998	0.980				0.5	0.983	0.945	0.995	0.977
			0.6	0.986	0.947	0.998	0.968				0.6	0.987	0.946	0.997	0.970
			0.7	0.986	0.947	0.997	0.968				0.7	0.989	0.952	0.997	0.972
			0.8	0.968	0.945	0.998	0.971				0.8	0.963	0.946	0.995	0.974
			0.9	0.986	0.950	0.998	0.961				0.9	0.989	0.950	0.998	0.968

(2, 25, 12). It should be noted that we observed with the Beta-approach estimates of a and b that were either negative or provided U-shaped densities. However, this hardly ever occurred for ICC values larger than 0.1. For ICC = 0.1, we did encounter parameter estimates $\hat{a}$ and $\hat{b}$ below one but still positive for maximally 3.25% of the simulated data sets. This maximum only occurred for design 2 in Table 4 for triplet (2,6,3). The same design also demonstrated one (or both) estimate(s) below zero for a maximum proportion of 2.1% of simulated data sets. All other settings (e.g. triplets and designs) with an ICC = 0.1 provided substantially lower proportions.

4.2 Three-way random effects model

The recently developed approach[32] is similar to our F-approach, but it uses Satterthwaite's approximation on the linear combination of mean squares. According to the authours,[32] their approach performs better than two other approaches which use bootstrap confidence intervals. The choices of variance components σ_S^2, σ_T^2, σ_O^2, σ_{TS}^2, σ_{SO}^2, σ_{TO}^2, σ_R^2 for the three-way random effects model are given in Table 6. The parameters for design A are identical to earlier study,[32] but we also explored other settings of the variance components, which are listed under design B. Different settings on the triplet (I, J, K) of sample sizes are considered, where K is again the number of observers evaluating each of J subjects at I time points: (2, 6, 3), (2, 20, 3), (2, 30, 3), (2, 60, 3), (2, 6, 5), (2, 10, 5) and (2, 20, 5). Our set of triplets include the settings (2, 30, 3), (2, 60, 3) which were investigated by others.[32] The simulated data were generated in a similar way as the linear mixed effects model used in section 4.

For the purpose of comparison, we present the marginal coverage probabilities for the two-sided 95% confidence intervals only for the triplets investigated earlier.[32] The results for both designs are shown in Table 7. For design A, we do not see any obvious winner between the three methods which would consistently give the nominal coverage. The F-approach is conservative for the triplet (2, 30, 3) and it is liberal for the triplet (2, 60, 3). However, for design A the F-approach is most frequently closest to the nominal coverage. For design B, the Beta-approach is most closest to the nominal value. Though, it is in some settings liberal and in other settings conservative. The other triplets (data not shown) did not provide a clear winner either. Those settings did not give more extreme coverage probabilities, than the ones observed in Table 7. For design B, the ANOVA approach did not provide good coverages and they were frequently liberal. The fact that the Beta-approach results

Table 6. Variance component parameters in three-way random effects model.

	Design A							Design B						
ICC	σ_S^2	σ_T^2	σ_O^2	σ_{TS}^2	σ_{SO}^2	σ_{TO}^2	σ_R^2	σ_S^2	σ_T^2	σ_O^2	σ_{TS}^2	σ_{SO}^2	σ_{TO}^2	σ_R^2
0.1	0.67	1.0	1.0	1.0	1.0	1.0	1.0	0.17	0.1	0.2	0.4	0.5	0.2	0.1
0.2	1.5	1.0	1.0	1.0	1.0	1.0	1.0	0.38	0.1	0.2	0.4	0.5	0.2	0.1
0.3	2.57	1.0	1.0	1.0	1.0	1.0	1.0	0.64	0.1	0.2	0.4	0.5	0.2	0.1
0.4	4.0	1.0	1.0	1.0	1.0	1.0	1.0	1.00	0.1	0.2	0.4	0.5	0.2	0.1
0.5	6.0	1.0	1.0	1.0	1.0	1.0	1.0	1.50	0.1	0.2	0.4	0.5	0.2	0.1
0.6	9.0	1.0	1.0	1.0	1.0	1.0	1.0	2.25	0.1	0.2	0.4	0.5	0.2	0.1
0.7	14.0	1.0	1.0	1.0	1.0	1.0	1.0	3.50	0.1	0.2	0.4	0.5	0.2	0.1
0.8	24.0	1.0	1.0	1.0	1.0	1.0	1.0	6.00	0.1	0.2	0.4	0.5	0.2	0.1
0.9	54.0	1.0	1.0	1.0	1.0	1.0	1.0	13.50	0.1	0.2	0.4	0.5	0.2	0.1

Table 7. Marginal coverage probabilities for two-sided 95% confidence intervals in three-way random effects model.

				Design A			Design B		
I	J	K	ICC	F	Beta	ANOVA	F	Beta	ANOVA
2	30	3	0.1	0.955	0.971	0.956	0.970	0.973	0.921
			0.2	0.960	0.956	0.946	0.969	0.935	0.895
			0.3	0.962	0.946	0.940	0.981	0.933	0.900
			0.4	0.960	0.945	0.938	0.981	0.938	0.904
			0.5	0.959	0.941	0.936	0.985	0.953	0.916
			0.6	0.960	0.934	0.932	0.984	0.964	0.914
			0.7	0.963	0.936	0.935	0.987	0.973	0.919
			0.8	0.958	0.928	0.933	0.989	0.975	0.922
			0.9	0.956	0.914	0.924	0.988	0.977	0.928
2	60	3	0.1	0.952	0.966	0.947	0.972	0.964	0.888
			0.2	0.952	0.945	0.938	0.980	0.918	0.880
			0.3	0.944	0.939	0.932	0.984	0.934	0.898
			0.4	0.936	0.928	0.924	0.982	0.950	0.904
			0.5	0.939	0.927	0.923	0.985	0.966	0.917
			0.6	0.932	0.912	0.910	0.984	0.970	0.917
			0.7	0.931	0.910	0.911	0.983	0.973	0.918
			0.8	0.928	0.902	0.908	0.985	0.974	0.923
			0.9	0.919	0.888	0.896	0.984	0.973	0.922

in better coverages for design B demonstrates that Beta is particularly accurate for small variance components (see Table 6). For smaller variance components, the Beta-approach starts to achieve perfect results when the number of subjects starts to exceed the number of observers 3–4 times. The only setting for which *F*-approach shows more accurate results than the Beta-approach is when the number of subjects are of the same order as the number of oncologists (e.g. (2,6,5)). Note that for larger number of subjects (e.g. 30) the occurrence of negative estimates for parameters a and b or estimates that give a *U*-shaped Beta-distribution are negligible.

4.3 One-way random effects model

For balanced and unbalanced one-way random effects model, we compare our generic approaches to three alternative approaches. These alternatives are Searle's exact method for balanced design and an adjusted version of Searle for unbalanced design,[17] Fisher's z-transformation introduced for equal group sizes[4] and extended for unequal group sizes,[37] and Smith's approach.[23] Fisher's z-transformation $0.5[\ln(1+(n_0-1)\widehat{\text{ICC}}) - \ln(1-\widehat{\text{ICC}})]$, with n_0 a weighted average of the within group sample sizes, was already applied[21,22] to unbalanced one-way random effects models, although it is not straightforward how to generalize such transformation to higher-order variance components models. Smith's approach was also studied for unbalanced one-way random effects models.[21,22]

We consider the one-way random effects model with balanced and unbalanced designs, similar to the one investigated earlier.[21] The residual variance component is set equal to $\sigma_R^2 = 1$. The variance component for groups is selected equal to $\sigma_G^2 = 0.112, 0.25, 0.43, 0.668, 1.0, 1.5, 2.35, 4.0, 9.0$. These choices lead to the values of $\text{ICC} = 0.1, \ldots, 0.9$ (0.1), respectively. Settings on the pairs (J, K) of sample sizes are considered, where J is the number of groups evaluated at K repeats: (5, 6), (10, 6),

Table 8. Marginal coverage probabilities for two-sided 95% confidence intervals in one-way balanced and unbalanced random effects model.

J	K	P	ICC	F	Beta	Searle	Fisher	Smith
25	3	0	0.1	0.930	0.940	0.967	0.997	0.973
			0.2	0.922	0.953	0.955	0.992	0.931
			0.3	0.940	0.963	0.948	0.992	0.927
			0.4	0.942	0.970	0.953	0.993	0.934
			0.5	0.933	0.957	0.950	0.993	0.929
			0.6	0.935	0.952	0.949	0.992	0.932
			0.7	0.935	0.949	0.951	0.993	0.936
			0.8	0.939	0.948	0.951	0.994	0.939
			0.9	0.943	0.945	0.949	0.992	0.939
5	6	0	0.1	0.977	0.966	0.948	0.971	0.822
			0.2	0.947	0.962	0.951	0.973	0.836
			0.3	0.930	0.960	0.948	0.974	0.841
			0.4	0.918	0.963	0.950	0.974	0.846
			0.5	0.918	0.961	0.950	0.978	0.856
			0.6	0.897	0.971	0.953	0.974	0.868
			0.7	0.887	0.945	0.952	0.973	0.883
			0.8	0.877	0.938	0.952	0.975	0.896
			0.9	0.876	0.934	0.950	0.972	0.913
5	6	0.1	0.1	0.976	0.962	0.949	0.973	0.826
			0.2	0.945	0.955	0.950	0.972	0.834
			0.3	0.928	0.954	0.952	0.974	0.839
			0.4	0.918	0.958	0.950	0.974	0.842
			0.5	0.919	0.958	0.945	0.971	0.847
			0.6	0.903	0.962	0.949	0.974	0.863
			0.7	0.889	0.951	0.946	0.971	0.876
			0.8	0.882	0.941	0.952	0.975	0.894
			0.9	0.885	0.935	0.948	0.973	0.911
5	6	0.2	0.1	0.980	0.963	0.950	0.975	0.834
			0.2	0.950	0.953	0.949	0.974	0.836
			0.3	0.935	0.948	0.951	0.977	0.842
			0.4	0.927	0.951	0.949	0.977	0.846
			0.5	0.922	0.951	0.948	0.976	0.851
			0.6	0.912	0.956	0.949	0.974	0.862
			0.7	0.895	0.951	0.947	0.974	0.875
			0.8	0.894	0.950	0.951	0.975	0.896
			0.9	0.886	0.941	0.950	0.977	0.913

SAS codes for simulations and data analysis are available upon request from the first author.

(5, 10), (5, 25), (12, 25), (25, 3), (50, 3). This set contains the pairs (25, 3), (50, 3) investigated by others.[21] Unbalanced data in the simulation study are implemented by introducing missing data in the full data set by random allocation, as earlier described.[34] Thus, the selected mechanism for missingness is "completely missing at random".[38] We investigated the proportions of missing values equal to $P = 0$, 0.1, 0.2. All combinations of settings on variance components, pairs of sample sizes, and proportions of missingness are studied.

The marginal coverage probabilities for the two-sided 95% confidence intervals for the set of representative pairs are presented in Table 8. In family designs (25, 3), (50, 3), Smith's approach was considered[21] as the most consistent. In such designs the number of subjects within groups is substantially smaller than the number of groups. Smith performs very poorly in our settings. Fisher shows conservative coverage probabilities over the whole range. The Beta-approach provides the most accurate coverages among the approximate methods for both balanced and unbalanced designs. Searle's approach is overall the best for one-way random effects models, but the Beta-approach seems to compete with this approach for the settings in Table 8. Interestingly, as unbalancedness increases the Beta-approach more often wins over approximated method of Searle. In other triplets (data not shown), the Beta-approach gives liberal coverages ($\approx 90\%$), when the number of groups is small ($J = 5$) and the number of repeats is large ($K = 25$). In all other settings, the Beta-approach behaves quite accurately.

5 Discussion

Lack of general approaches on constructing confidence intervals on ICCs in agreement studies motivated this work. In this paper, we have proposed two generic closed-form approaches for constructing confidence intervals on ICCs of the form

$$\sum_{q=1}^{Q} \sigma_q^2 \Big/ \left(\sum_{q=1}^{Q} \sigma_q^2 + \sum_{p=Q+1}^{P} \sigma_p^2 \right)$$

Both approaches take into consideration the skewness of the distribution of the ICC, but they model it differently. We examined these approaches primarily on three-way mixed and random effects models, and on the one-way random effects model. The generic F-approach is often conservative, but it works better when more variance components are involved. This implies that the generic F-approach is not the most suitable method for the one-way random effects model. The Beta-approach demonstrates coverages which are (very) close to the nominal value and these results are consistent across the investigated settings which are typical for agreement studies.

In comparison with the ANOVA method developed for three-way random effects model,[32] the Beta-approach outperformed ANOVA for settings with smaller variance components with almost all investigated sample size triplets. For settings with larger variance components, the F-approach outperformed the ANOVA method, although these two were close to each other. However, none of them showed particularly good coverages. For balanced and unbalanced one-way random effects models, Searle's methods outperformed all other methods, including our two generic approaches, but the Beta-approach is quite competitive. Most interestingly, the Beta-approach achieves almost the same accuracy as the approximate method of Searle for unbalanced designs.

Limitations of the proposed approaches are that they are intended for particular forms of the ICCs. For agreement studies, these are often the appropriate form, but further advancements on our generic approaches for other forms are required. Another limitation is that we did not simulate all possible relevant settings for the use of our approaches. This means that we cannot yet claim that our generic approaches, in particular the Beta-approach, is universally good. Possibly some theoretical work needs to be done to prove that the Beta-approach is a suitable approximation to the ICC estimate in all settings. Additionally, our pragmatic approach on confidence intervals for zero ICCs was based on the one-way ANOVA and worked well in our simulations, though it does require more theoretical effort to demonstrate that it is suitable for other variance components models.

The main strength of our generic approaches is that they can be applied to any variance components model. In particular they are applicable to unbalanced designs, for which it is more complicated to construct generic approaches based on moment estimates. Furthermore, closed-form methods are typically beneficial for sample size calculations when a certain confidence length on ICCs is required. Finally, our closed-form method worked especially well for clustered or dependent data (agreement studies) in settings with only limited numbers of clusters. In these settings it is more difficult to construct confidence intervals due to lack of asymptotic approaches.

Alternative generic approaches to our closed-form methods are bootstrap and other resampling methods. But it is not straightforward how to implement these approaches for data from higher-order linear models due to the complexity of bootstrapping variance components, the complexity of multiple clusters, and the diversity of bootstrap methods. The complexity of bootstrap methods on variance components has been discussed in the literature,[39] indicating that depending on the application each variance component should be simulated implicitly or explicitly. Other researchers[40] discuss different ways of selecting bootstrap samples from (one-way) clustered data: the randomized cluster bootstrap, cluster bootstrap, two-stage bootstrap, reverse two-stage bootstrap, random effects bootstrap, and residual bootstrap. These methods all do take into account sampling the clusters, but they may differ in sampling observations within clusters or may differ in sampling order (first clusters and then observations or first observations and then clusters). For higher-order variance component models clusters are formed in different ways and even more possibilities for sampling would become available.[41] Confidence intervals on variance components for models similar to ours (three-way ANOVA models) has been studied[41] earlier using bootstrapping at different cluster levels of the data. None of the bootstrap methods demonstrated good results on all variance components. Only when a bias-adjustment for the estimation of the variance components was implemented,[42] all level bootstrap methods seem to behave equally well. However, the bias adjustment procedures were only developed for balanced data and it would become more complicated to do the same for unbalanced designs. Furthermore, it is not evident that these results can directly be translated into good coverages on confidence intervals for ICCs. This point was elegantly indicated by some researchers,[43] who studied confidence intervals on ICCs for one-way ANOVA models. It was demonstrated that bootstrapping the ICC should adopt an appropriate transformation to make the standard error of the ICC (almost) independent of the variance components of the linear model.[43] It was concluded that standard methods of bootstrapping lead to markedly less than nominal coverages when there are 30 or fewer clusters and that good results are obtained consistently only for 50 or more clusters. The bootstrap *t*-method with the transformed ICC provided good results, also for as low as 10 clusters, but this method provided wider confidence intervals than the closed-form approaches for normal balanced data. Thus, more research on bootstrapping is needed and it would be of interest to compare our generic approaches to bootstrap or other resampling procedures.

In conclusion, the *F*-approach provides reasonable accuracy of the confidence interval on ICCs across a diverse range of settings. Though, the Beta-approach is more accurate and therefore is recommended for agreement studies, particularly for the mix of small sample sizes.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Brouwer CL, Steenbakkers RJHM, van den Heuvel E, et al. 3D variation in delineation of head and neck organs at risk. *Radiat Oncol* 2012; **7**: 32.
2. Galton F. Family likeness in stature. *Proc R Soc* 1886; **40**: 42–73.
3. Pearson KVII. Mathematical contributions to the theory of evolution – III. Regression, heredity, and panmixia. *Phil Trans R Soc: Ser A* 1896; **187**: 253–318.
4. Fisher RA. *Statistical methods for research workers (*rev.*)*. Edinburgh: Oliver and Boyd, 1925.
5. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measure* 1960; **20**: 37–46.
6. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971; **76**: 378–382.
7. Fleiss JL and Cuzick J. The reliability of dichotomous judgments: unequal numbers of judges per subject. *Appl Psychol Measure* 1979; **3**: 537–542.
8. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep* 1966; **19**: 3–11.
9. Fleiss JL and Shrout PE. Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika* 1978; **43**: 259–262.
10. McGraw KO and Wong S. Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1996; **1**: 30–46.
11. Haber M and Barnhart HX. Coefficients of agreement for fixed observers. *Stat Methods Med Res* 2006; **15**: 255–271.
12. Barnhart HX, Haber MJ and Lin LI. An overview on assessing agreement with continuous measurements. *J Biopharm Stat* 2007; **17**: 529–569.
13. Burdick RK and Graybill FA. *Confidence intervals on variance components*. New York: Marcel Dekker, 1992.
14. Webb NM, Shavelson RJ and Haertel EH. Reliability coefficients and generalizability theory. *Handbook of statistics* 2006; **26**: 81–124.
15. Brennan RL. Generalizability theory and classical test theory. *Appl Measure Educ* 2010; **24**: 1–21.
16. Narayanan A, Greco M and Campbell JL. Generalisability in unbalanced, uncrossed and fully nested studies. *Med Educ* 2010; **44**: 367–378.
17. Searle SR, Casella G and McCulloch CE. *Variance components*. New Jersey: John Wiley & Sons, 2006.
18. McCulloch CE, Searle SR and Neuhaus JM. *Generalized, linear, and mixed models*. New York: Wiley, 2001.
19. Searle SR. *Linear models*. New York: John Wiley & Sons, 1971.
20. Wald A. A note on the analysis of variance with unequal class frequencies. *Ann Math Stat* 1940; **11**: 96–100.
21. Donner A and Wells G. A comparison of confidence interval methods for the intraclass correlation coefficient. *Biometrics* 1986; **42**: 401–412.
22. Ukoumunne OC. A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Stat Med* 2002; **21**: 3757–3774.
23. Smith CA. On the estimation of intraclass correlation. *Ann Hum Genet* 1957; **21**: 363–373.
24. Zou KH and McDermott MP. Higher-moment approaches to approximate interval estimation for a certain intraclass correlation coefficient. *Stat Med* 1999; **18**: 2051–2061.
25. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometr Bull* 1946; **2**: 110–114.
26. Graybill FA and Wang CM. Confidence intervals for proportions of variability in two-factor nested variance component models. *J Am Stat Assoc* 1979; **74**: 368–374.
27. Graybill FA and Wang CM. Confidence intervals on nonnegative linear combinations of variances. *J Am Stat Assoc* 1980; **75**: 869–873.
28. Ting N, Burdick RK, Graybill FA, et al. Confidence intervals on linear combinations of variance components that are unrestricted in sign. *J Stat Comput Simul* 1990; **35**: 135–143.
29. Gui R, Graybill FA, Burdick RK, et al. Confidence intervals on ratios of linear combinations for non-disjoint sets of expected mean squares. *J Stat Plan Inf* 1995; **48**: 215–227.
30. Cappelleri JC and Ting N. A modified large-sample approach to approximate interval estimation for a particular intraclass correlation coefficient. *Stat Med* 2003; **22**: 1861–1877.
31. Ting N, Burdick RK and Graybill FA. Confidence intervals on ratios of positive linear combinations of variance components. *Stat Probab Lett* 1991; **11**: 523–528.
32. Wang L, Keen KJ and Holland B. Estimation of reliability in a three-factor model. *Stat Med* 2011; **30**: 1254–1265.
33. Swiger LA, Harvey WR, Everson DO, et al. The variance of intraclass correlation involving groups with one observation. *Biometrics* 1964; **20**: 818–826.
34. Van den Heuvel ER. A comparison of estimation methods on the coverage probability of Satterthwaite confidence intervals for assay precision with unbalanced data. *Commun Stat – Simul Comput* 2010; **39**: 777–794.
35. Johnson LN, Kotz S and Balakrishnan N. *Continuous univariate distributions*. vol. 2. New York: John Wiley & Sons, 1995.
36. Knight K. *Matehmatical statistics*. Boca Raton: Chapman & Hall/CRC, 2000.
37. Zhivotovskii LA. Estimation of the intraclass correlation coefficient (translated). *Genetika* 1978; **15**: 1235–1242.
38. Little RJA and Rubin DB. *Statistical analysis with missing data*. New Jersey: John Wiley & Sons, 2002.
39. Hinkley DV. Bootstrap methods. *J R Stat Soc: Ser B* 1988; **50**: 321–337.
40. Field CA and Welsh AH. Bootstrapping clustered data. *J R Stat Soc: Ser B* 2007; **69**: 369–390.
41. Tong Y and Brennan RL. Bootstrap estimates of standard errors in generalizability theory. *Educ Psychol Measure* 2007; **67**: 804–817.
42. Brennan RL. Unbiased estimates of variance components with bootstrap procedures. *Educ Psychol Measure* 2007; **67**: 784–803.
43. Ukoumunne OC, Davison AC, Gulliford MC, et al. Non-parametric bootstrap confidence intervals for the intraclass correlation coefficient. *Stat Med* 2003; **22**: 3805–3821.