Association of intracluster correlation measures with outcome prevalence for binary outcomes in cluster randomised trials

Ariane Mbekwe Yepnang, Agnès Caille, Sandra M Eldridge, Bruno Giraudeau

Journal Title
XX(X):1–20

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Intracluster correlation measures in cluster randomized trials with binary outcomes

Ariane M. Mbekwe Yepnang[1], Agnès Caille[1,2], Sandra M. Eldridge[3] and Bruno Giraudeau[1,2]

Abstract

In cluster randomized trials, a measure of intracluster correlation such as the intraclass correlation coefficient (ICC) should be reported for each primary outcome. Providing intracluster correlation estimates may help in calculating sample size of future cluster randomized trials and also in interpreting the results of the trial from which they are derived. For a binary outcome, the ICC is known to be associated with its prevalence, which raises at least two issues. First, it questions the use of ICC estimates obtained on a binary outcome in a trial for sample size calculations in a subsequent trial in which the same binary outcome is expected to have a different prevalence. Second, it challenges the interpretation of ICC estimates because they do not solely depend on clustering level. Other intracluster correlation measures have been proposed for clustered binary data settings including the variance partition coefficient, the median odds ratio and the tetrachoric correlation coefficient. Under certain assumptions, the theoretical maximum possible value for an ICC associated with a binary outcome can be derived and we propose to consider the relative deviation of an ICC estimate to this maximum value, as another measure of the intracluster correlation. We conducted a simulation study to explore the dependence of these intracluster correlation measures on outcome prevalence and found that all these measures are associated with prevalence. Even if some were slightly less dependent than the ICC in some scenarios, in general, none differs from the ICC regarding the dependence on prevalence.

[1]Université de Tours, Université de Nantes, INSERM, SPHERE U1246, Tours, France
[2]INSERM CIC1415, CHRU de Tours, Tours, France
[3]Centre for Primary Care and Public Health, Queen Mary University of London, London, UK

Corresponding author:
Ariane M. Mbekwe Yepnang, Bd Tonnellé 37044 Tours cedex 9, France
Email: ariane.mbekweyepnang@etu.univ-tours.fr

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

Keywords
Intraclass correlation coefficient, binary outcome, prevalence, cluster, variance partition coefficient, median odds ratio, tetrachoric correlation coefficient

1 Introduction

The Consolidated Standards for Reporting of Trials (CONSORT) statement extension for cluster randomized trials recommends reporting a measure of intracluster correlation, such as the intraclass correlation coefficient (ICC), for each primary outcome.[1] This was previously recommended by Donner and Klar to help in sample size calculation of future cluster randomized trials.[2] In addition, providing intracluster correlation estimates, which we may also call clustering estimates, may help in interpreting the results of the trial. Indeed, interventions may affect the level of clustering, which is important to be known for a complete interpretation of the trial's results. For example, if clustering is lower in the intervention arm as compared to the control one, this means that there is a better homogeneity in outcomes among clusters of the intervention arm, as compared to those of the control arm, which may result from a standardization in practices due to the intervention itself. Yet, when the outcome is binary, the ICC is known to be associated with the prevalence of the outcome.[3] This association challenges the interpretation of the ICC because ICC values no longer just depend on clustering level. This association can also be problematic for future sample size calculations. Indeed, if the study to be planned ("future study") is expected to have outcome prevalences different from those associated to the study for which we obtained ICC estimates ("past study"), this challenges sample size calculation. If prevalences associated to the "future study" are closer to 50% than those associated to the "past study", ICC estimates are expected to be higher. Therefore calculating sample size using ICC estimates from the "past study" may lead to an under-powered "future study". Conversely, if prevalences associated to the "future study" are further from 50% than those of the "past study", the sample size calculation may lead to an over-powered "future study".

To overcome this drawback of the ICC, Mbekwe et al. investigated whether the R coefficient is independent of the outcome prevalence.[4] R is defined as a ratio for which the numerator is the conditional probability that a member of a cluster has the outcome given that another member of the cluster also has the outcome, and the denominator is the outcome prevalence.[5] The R coefficient seemed to be an alternative to the ICC in that Crespi et al. asserted that R may be less influenced by the outcome prevalence than the ICC. Unfortunately, R depends on prevalence and cannot be considered a better alternative to the ICC.[4]

Other measures of intracluster correlation for binary outcomes have been proposed and, to our knowledge, none has investigated whether they depend or not on the outcome prevalence. These are the variance partition coefficient (VPC),[6] the median odds ratio (MOR)[7] and the tetrachoric correlation coefficient (TCC).[8] Otherwise, for a given prevalence, Eldridge[9] derived the theoretical maximum possible value for an ICC. Therefore for an ICC estimate and its associated outcome prevalence, we considered the relative difference between this ICC estimate and the theoretical maximum possible value associated to the observed prevalence. This relative difference was also considered as an intracluster correlation measure. In this paper, we aim at investigating whether these intracluster correlation measures are independent from the outcome prevalence.

In section 2, we define the selected measures. In section 3, we report a simulation study conducted to explore the dependence of these measures on outcome prevalence. Section 4 is dedicated to comparing the different measures. We report an example in section 5, continue with a discussion in section 6 and conclude in section 7.

2 Definitions

We consider one arm composed of k clusters of size $n_i (i = 1, 2, ..., k)$; X_{ij} the binary outcome of the jth, $j = 1, 2, ..., n_i$ individual in the ith cluster with $X_{ij} = 1$ for success and $X_{ij} = 0$ for failure, $X_i = \sum_{j=1}^{n_i} X_{ij}$ the total number of successes in the ith cluster, $p_i = X_i/n_i$ the proportion of success in cluster i and $N = \sum_{i=1}^{k} n_i$ the total number of individuals. We assume that the success probability p is the same for all individuals (i.e., $\mathrm{P}(X_{ij} = 1) = p$). We consider the following model

$$p_i = \mathrm{g}(\mu + \gamma_i) \tag{1}$$

where $\gamma_i \sim \mathcal{N}(0, \sigma_\gamma^2)$, σ_γ^2 is the cluster-level variance and g^{-1} can be, for instance, the logit or the probit function.

2.1 The variance partition coefficient (VPC)

In the context of multilevel regression models, the VPC represents the proportion of the total variance found at the highest-level source of variation.[6] For example, if we have patients nested between hospitals, the lowest-level units are patients and the highest-level units are hospitals. The VPC represents the proportion of the total variance between hospitals. The VPC is equivalent to the ICC for a random intercept model fitted to a continuous outcome. Otherwise, for a binary outcome, Goldstein et al. proposed the four approaches below to estimate the VPC when we consider model (1):

- The first approach consists of using a first order Taylor expansion, which leads to the following approximation

$$\mathrm{VPC}_1 = \frac{\frac{\sigma_\gamma^2 p^2}{(1+e^\mu)^2}}{\frac{\sigma_\gamma^2 p^2}{(1+e^\mu)^2} + p(1-p)} \tag{2}$$

  when assuming that g^{-1} is the logit function in model (1). VPC_1 can be estimated by using estimates of μ, p and σ_γ^2. μ and σ_γ^2 are estimated from the fitted mixed-effects logistic regression model and p is estimated as the observed overall prevalence.

- The second approach consists of using simulations, which leads to the following approximation

$$\mathrm{VPC}_2 = \frac{v_2}{v_2 + v_1} \tag{3}$$

  where v_1 and v_2 are obtained following the steps below:

  - Simulate B values for $\widehat{\gamma}$, denoted $\widehat{\gamma}^{(b)}$, $b = 1, 2, ..., B$, from $\mathcal{N}(0, \widehat{\sigma}_\gamma^2)$ with $\widehat{\sigma}_\gamma^2$ an estimate of σ_γ^2 from the fitted model (1).

– Estimate μ from the fitted model (1) as $\widehat{\mu}$.

– Estimate $\widehat{p}^{(b)}$ as $\frac{e^{\widehat{\mu}+\widehat{\gamma}^{(b)}}}{1+e^{\widehat{\mu}+\widehat{\gamma}^{(b)}}}$.

– Estimate v_1 as the mean of the $\widehat{p}^{(b)}(1-\widehat{p}^{(b)}), b = 1, 2, ..., B$ and the level two variance v_2 as the variance of the $\widehat{p}^{(b)}, b = 1, 2, ..., B$.

Goldstein et al. recommended simulating about 5000 values for $\widehat{\gamma}$.[6]

- The third approach consists of treating X_{ij} as a normally distributed variable and calculating the VPC as we do with a continuous outcome. We estimated VPC by using the analysis of variance ICC estimator as recommended by Donner and Koval for low to moderate ICC values (< 0.5).[10] Then VPC_3 is equal to[11]

$$\text{VPC}_3 = \frac{\text{MSB} - \text{MSW}}{\text{MSB} + (n_0 - 1)\text{MSW}} \tag{4}$$

where $\text{MSB} = \frac{1}{k-1}\sum_{i=1}^{k}\left[n_i\left(p_i - \frac{\sum_{i=1}^{k} X_i}{N}\right)^2\right]$, $\text{MSW} = \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - p_i)^2$ and $n_0 = \frac{1}{k-1}\left(N - \frac{\sum_{i=1}^{k} n_i^2}{N}\right)$.

- The last approach consists of treating X_{ij} as arising from an underlying continuous variable. The following formula allows for calculating VPC_4

$$\text{VPC}_4 = \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \frac{\pi^2}{3}} \tag{5}$$

assuming that the underlying variable follows a standard logistic distribution [i.e., g^{-1} is the logit function in model (1)].

2.2 The median odds ratio (MOR)

The MOR has been proposed in the context of social epidemiology to quantify the relative importance of different sources of variation. It is based on the mixed-effects logistic regression model.[7] The aim was to find a function of the relevant random effects parameters that has a nice interpretation in terms of an odds ratio. The MOR, which is a measure of heterogeneity, quantifies the variation between clusters. Considering two individuals from two distinct clusters, we may consider the odds ratio between the individual of higher probability of having the outcome of interest and the individual of lower probability. Odds ratios are then estimated for any pair of individuals and the MOR is the median value of theses odds ratios. The lower bound of MOR is 1, which corresponds to a situation where there is no variation between clusters, whereas if MOR is large, there is considerable between-cluster variation. The MOR is defined by

$$\text{MOR} = \exp\left[\sqrt{2 \times \sigma_\gamma^2} \times \ \Phi^{-1}(0.75)\right] \tag{6}$$

where Φ is the cumulative distribution function of the standard normal distribution.

MOR can be estimated by using an estimate of σ_γ^2 obtained from the fitted mixed-effects model (1) and considering a logit link function.

2.3 The tetrachoric correlation coefficient (TCC)

2.3.1 Historical definition of the TCC[8]

Let us consider two binary variables V_1 and V_2 measured for a given subject. V_1 equals 1 for success and 0 for failure; idem for V_2. We suppose that the observed variables V_1 and V_2 are manifestations of underlying continuous variables U_1 and U_2, which have standard normal distributions. The 2×2 contingency table for V_1 and V_2 is shown by

		V_2		
		Success	Failure	Total
	Success	a	b	$a+b$
V_1	Failure	c	d	$c+d$
	Total	$a+c$	$b+d$	n

Let be:

- $p_{_{V_1}} = \frac{a+b}{n}$ the proportion of success on variable V_1,
- $p_{_{V_2}} = \frac{a+c}{n}$ the proportion of success on variable V_2,
- h_1 and h_2, real numbers for which prob$(U_1 > h_1) = p_{_{V_1}}$ and prob$(U_2 > h_2) = p_{_{V_2}}$, and
- r the correlation between U_1 and U_2. r is the tetrachoric correlation coefficient.

The bivariate density function of (U_1, U_2) is $f_{U_1,U_2}(x,y) = \frac{1}{2\pi\sqrt{1-r^2}} \exp\left[-\frac{(x^2+y^2-2rxy)}{2(1-r^2)}\right]$.
Thus,

$$a = n * \text{prob}(U_1 > h_1, U_2 > h_2) = \frac{n}{2\pi\sqrt{1-r^2}} \int_{h_1}^{+\infty} \int_{h_2}^{+\infty} \exp\left[-\frac{(x^2+y^2-2rxy)}{2(1-r^2)}\right] dxdy \quad (7)$$

Using the Leibnitz's theorem, Pearson[8] rewrote equation (7) as

$$\frac{ad-bc}{n^2 H_1 H_2} = \sum_{s=1}^{\infty} c_s \frac{r^s}{s!} \quad (8)$$

where $c_1 = 1$, $c_2 = h_1 h_2$, $c_3 = (h_1^2 - 1)(h_2^2 - 1)$, $c_4 = h_1 h_2 (h_1^2 - 3)(h_2^2 - 3)$ and the other terms can be computed as $c_{s+1} = s(2s - 1 - h_1^2 - h_2^2)c_{s-1} - s(s-1)(s-2)^2 c_{s-3} + h_1 h_2 [c_s + s(s-1)c_{s-2}]$ with $h_1 = \Phi^{-1}\left[\frac{\frac{(a+c)-(b+d)}{n}-1}{2}\right]$, $h_2 = \Phi^{-1}\left[\frac{\frac{(a+b)-(c+d)}{n}-1}{2}\right]$, $H_1 = \frac{1}{\sqrt{2\pi}}\exp(\frac{-h_1^2}{2})$ and $H_2 = \frac{1}{\sqrt{2\pi}}\exp(\frac{-h_2^2}{2})$. The r coefficient is the solution for equation (8).

2.3.2 The TCC in the context of cluster randomized trials

Here, we assume that X_{ij} is from a latent normal continuous outcome $Y_{ij} \sim \mathcal{N}(\mu_{_Y}, \sigma_{_Y}^2)$ so that $X_{ij} = 1$ if $Y_{ij} > \mu_{_Y} + h_{_Y}\sigma_{_Y}$ and $X_{ij} = 0$ otherwise, where $h_{_Y}$ is a constant such that the outcome prevalence $p = 1 - \Phi(h_{_Y})$. In this context, from works of Kirk[12] and Kraemer[13], Donner and Eliasziw[14] reported that

$$\rho_{\mathrm{X}} = \frac{1}{2\pi p(1-p)} \int_0^{\rho_{\mathrm{Y}}} \frac{1}{\sqrt{1-x^2}} \exp\left(\frac{-h_{\mathrm{Y}}^2}{1+x}\right) dx, \tag{9}$$

where ρ_{X} is the ICC associated with the binary outcome X_{ij} and ρ_{Y} is the ICC associated with the underlying continuous outcome Y_{ij}. ρ_{Y} corresponds to the tetrachoric correlation coefficient (i.e., the correlation coefficient associated with the latent variable underlying the binary outcome of interest). Although equation (9) was provided in the case of fixed cluster sizes ($n_i = 2, i = 1, 2, ..., k$), Caille et al. showed that this formula also holds for cluster sizes > 2, fixed or variable.[15]

For a given ρ_{X}, ρ_{Y} is the solution of equation (9) and can be approximated by using the trapezoidal rule defined by Davis and Rabinowitz.[16]

2.4 The relative deviation of the ICC estimate to its theoretical maximum possible value

Eldridge derived the theoretical maximum possible ICC value for a binary outcome.[9] Under the assumption that true cluster prevalences $p_i, i = 1, ..., k$ follow a beta distribution Beta(α, β), then the mean and the variance of $p_i, i = 1, ..., k$ are respectively equal to

$$(S_1) \begin{cases} \mathrm{E}(p_i) = \frac{\alpha}{\alpha+\beta} \\ \mathrm{V}(p_i) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{cases}$$

Otherwise, considering that $p = \mathrm{E}(p_i)$ and that $\rho_{\mathrm{X}} = \frac{\mathrm{V}(p_i)}{p(1-p)}$ [17], we have

$$(S_2) \begin{cases} \mathrm{E}(p_i) = p \\ \mathrm{V}(p_i) = \rho_{\mathrm{X}} p(1-p) \end{cases}$$

(S_1) and (S_2) lead to

$$(S_3) \begin{cases} \alpha = \frac{1-\rho_{\mathrm{X}}}{\rho_{\mathrm{X}}} p \\ \beta = \frac{1-\rho_{\mathrm{X}}}{\rho_{\mathrm{X}}} (1-p) \end{cases}$$

If we assume that cluster prevalences follow a unimodal distribution, α and β are greater than or equal to 1.[18]

- $\alpha = \frac{1-\rho_{\mathrm{X}}}{\rho_{\mathrm{X}}} p$ and $\alpha \geq 1 \Rightarrow \rho_{\mathrm{X}} \leq \frac{1}{1+\frac{1}{p}}$.
- $\beta = \frac{1-\rho_{\mathrm{X}}}{\rho_{\mathrm{X}}} (1-p)$ and $\beta \geq 1 \Rightarrow \rho_{\mathrm{X}} \leq \frac{1}{1+\frac{1}{1-p}}$.

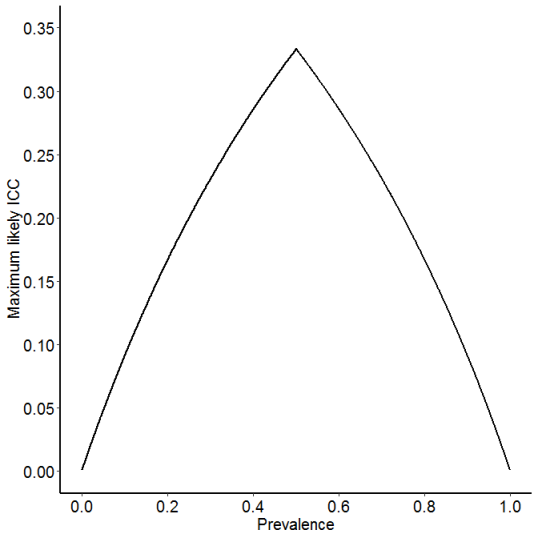

Figure 1. Theoretical maximum possible values of the intraclass correlation coefficient (ICC) for each prevalence value.

We obtain the following definition for $\rho_{\max}$, the theoretical maximum possible value of ICC.

$$\rho_{\max}(p) = \begin{cases} \frac{1}{1+\frac{1}{p}} \text{ if } p < 0.5 \\ \frac{1}{1+\frac{1}{1-p}} \text{ if } p > 0.5 \\ \frac{1}{3} \text{ if } p = 0.5 \end{cases} \tag{10}$$

For each prevalence value, under the assumptions made, $\rho_{\max}$, is shown in Figure 1.

For a given prevalence p, we define the relative deviation of the ICC estimate to its theoretical maximum possible value as

$$R_d(p) = \frac{\rho_{\max}(p) - \rho_{\mathrm{X}}(p)}{\rho_{\max}(p)} \times 100 \tag{11}$$

where $\rho_{\max}(p)$ is the theoretical maximum possible ICC value associated with prevalence p and $\rho_{\mathrm{X}}(p)$ is the ICC associated with the binary variable X. To estimate R_d, one first needs to estimate the prevalence and the ICC. Then, from the estimated prevalence, $\rho_{\max}$ can be estimated using function (10). R_d varies between 0% (if $\rho_{\mathrm{X}} = \rho_{\max}$) and 100% (if $\rho_{\mathrm{X}} = 0$).

3 Simulation study

We conducted a simulation study to investigate the shape of the relationship between the measures defined in the previous section and the prevalence. The principle was as follows. We generated correlated binary data W_{ij} with pre-specified outcome prevalence p_{W} and intraclass correlation ρ_{W}. Because we wanted to obtain datasets with the same level of clustering whatever the outcome prevalence, we associated W_{ij}

with a latent normal continuous outcome $Z_{ij} \sim \mathcal{N}(\mu_{\text{Z}}, \sigma^2_{\text{Z}})$ so that $W_{ij} = 1$ if $Z_{ij} > \mu_{\text{Z}} + h_{\text{Z}}\sigma_{\text{Z}}$ and $W_{ij} = 0$ if $Z_{ij} \leq \mu_{\text{Z}} + h_{\text{Z}}\sigma_{\text{Z}}$ where h_{Z} is a constant such as $p_{\text{W}} = 1 - \Phi(h_{\text{Z}})$.

ρ_{Z}, the ICC associated with the underlying latent variable Z_{ij}, was specified at first. Then, we varied p_{W} and for each pair (ρ_{Z}, p_{W}), we calculated ρ_{W} by using (9). We proceeded in this way to have the common underlying level of clustering equal to ρ_{Z} for each value of p_{W}. Thus, for each pair $(p_{\text{W}}, \rho_{\text{W}})$, we defined W_{ij} as[19]

$$W_{ij} = (1 - S_{ij})T_{ij} + S_{ij}R_i, \tag{12}$$

where $S_{ij} \sim \text{Binom}(1, \sqrt{\rho_{\text{W}}}), T_{ij} \sim \text{Binom}(1, p_{\text{W}})$ and $R_i \sim \text{Binom}(1, p_{\text{W}})$.

3.1 Simulation plan

Steps of the data generation for each pair $(p_{\text{W}}, \rho_{\text{W}})$ were as follows:

a. Generation of a dataset

1. Simulate n_{w_i}, $i = 1, ..., k$ variable cluster sizes from a negative binomial distribution with mean m and variance v; m then corresponds to mean cluster sizes.
2. For each individual, simulate $S_{ij}, i = 1, ..., k, \;\; j = 1, ..., n_i$ under a binomial distribution with parameters 1 and $\sqrt{\rho_{\text{W}}}$.
3. For each individual, simulate $T_{ij}, i = 1, ..., k, \;\; j = 1, ..., n_i$ under a binomial distribution with parameters 1 and p_{W}.
4. For each cluster, simulate $R_i, i = 1, ..., k$, under a binomial distribution with parameters 1 and p_{W}.
5. Calculate W_{ij} according to equation (12).

b. Analysis

We varied p_{W} between 0.01 and 0.99. For each p_{W}, statistical analyses were conducted to estimate the different measures of intracluster correlation on the same datasets. Steps are as follows:

1. Estimate p_{W} as $\widehat{p}_{\text{W}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} W_{ij}/N_{\text{W}}$, with N_{W} the total number of individuals.
2. Estimate μ and σ^2_γ using model (1) with g^{-1} being the logit function.
3. Estimate VPC_1 as $\widehat{\text{VPC}}_1$ using equation (2).
4. Estimate VPC_2 as $\widehat{\text{VPC}}_2$ using equation (3).
5. Estimate VPC_3 as $\widehat{\text{VPC}}_3$ using the analysis of variance estimator of ICC for continuous outcome defined in equation (4). All negative values of $\widehat{\text{VPC}}_3$ were truncated to 0.
6. Estimate VPC_4 as $\widehat{\text{VPC}}_4$ using equation (5).
7. Estimate MOR as $\widehat{MOR}$ using equation (6).
8. Estimate r as $\widehat{r}$ using equation (8) with data structured as detailed in Appendix 1. The stop criterion was 10^{-5}.

9. Estimate ρ_{W} as $\widehat{\rho}_{\text{W}}$ using the analysis of variance estimator of ICC for a binary outcome. $\widehat{\rho}_{\text{W}}$ is then equal to $\widehat{\text{VPC}}_3$. All negative values of $\widehat{\rho}_{\text{W}}$ were truncated to 0.
10. Estimate ρ_{Y} as $\widehat{\rho}_{\text{Y}}$ using equation (9) with p estimated as $\widehat{p}_{\text{W}}$ and ρ_{X} estimated as $\widehat{\rho}_{\text{W}}$.
11. Estimate R_d as $\widehat{R}_d$ using equation (11) with ρ_{X} estimated as $\widehat{\rho}_{\text{W}}$.

c. Performance measures

For each scenario, we generated 10000 datasets and for each p_{W} value, we summarized results by computing $\overline{\widehat{p}_{\text{W}}}$, $\overline{\widehat{\text{VPC}}_1}$, $\overline{\widehat{\text{VPC}}_2}$, $\overline{\widehat{\text{VPC}}_3}$, $\overline{\widehat{\text{VPC}}_4}$, $\overline{\widehat{\text{MOR}}}$, $\bar{\widehat{r}}$, $\overline{\widehat{\rho}_{\text{W}}}$, $\overline{\widehat{\rho}_{\text{Y}}}$ and $\overline{\widehat{R}_d}$ the empirical means of the estimated $\widehat{p}_{\text{W}}$, $\widehat{\text{VPC}}_1$, $\widehat{\text{VPC}}_2$, $\widehat{\text{VPC}}_3$, $\widehat{\text{VPC}}_4$ $\widehat{\text{MOR}}$, $\widehat{r}$, $\widehat{\rho}_{\text{W}}$, $\widehat{\rho}_{\text{Y}}$ and $\widehat{R}_d$, respectively.

Three initial values of ρ_{Z} (0.01, 0.05, 0.3) were considered in simulations, which are realistic values observed in cluster randomized trials, and for each value, cluster numbers of $k = 10$, 20 and 50. We considered variable cluster sizes of mean $m = 25$ and variance $v = 225$, which corresponds to a situation in which the coefficient of variation of cluster size $\sqrt{v}/m$ equals a value (0.6), which appears to be a plausible one.[20]

All programming involved using R software, v4.0.2.

3.2 Simulation results

Figure 2 displays plots of variance partition coefficient expected means $\overline{\widehat{\text{VPC}}_1}$, $\overline{\widehat{\text{VPC}}_2}$, $\overline{\widehat{\text{VPC}}_3}$ and $\overline{\widehat{\text{VPC}}_4}$ as a function of the prevalence expected mean $\overline{\widehat{p}_{\text{W}}}$, for three values of ρ_{Z} and three numbers of clusters. $\overline{\widehat{\text{VPC}}_2}$ and $\overline{\widehat{\text{VPC}}_3}$ tended to increase when prevalence varied from 0 to 0.5 and decreased when prevalence varied from 0.5 to 1. The same phenomenon was observed for $\overline{\widehat{\text{VPC}}_1}$, except that there is no longer symmetry with the 0.5 prevalence value when $\rho_{\text{Z}} = 0.3$, which is a mathematical consequence of the definition of VPC_1. The curve of $\overline{\widehat{\text{VPC}}_4}$ when $\rho_{\text{Z}} = 0.3$ differed from those associated with other ρ_{Z} values. This may be related to the between-cluster variance estimates. The proportion of null estimates are important for extreme prevalence, even when $\rho_{\text{Z}} = 0.3$, but in a weaker proportion as when ρ_{Z} is smaller (0.05 or 0.01) (Table 1).

On Figure 3, $\overline{\widehat{\text{MOR}}}$ appeared to be higher for extreme than for less extreme prevalence values. This convexity of the curves of $\overline{\widehat{\text{MOR}}}$ for low values of ρ_{Z} (0.01 or 0.05) and concavity for $\rho_{\text{Z}} = 0.3$ when removing extreme prevalence values (cf e.Figure 2) are similar to those observed for $\overline{\widehat{\text{VPC}}_4}$, which is a consequence of both measures depending only on the between-cluster variance. For $\rho_{\text{Z}} = 0.3$ and $k = 50$, $\overline{\widehat{\text{MOR}}}$ seemed to be nearly constant but is simply due to the fact that the scale was chosen to be the same as for $k = 10$ or $k = 20$ and $\rho_{\text{Z}} = 0.3$. However, using another scale allows to see that the curve remains concave (cf e.Figure 2). MOR, as VPC_4 only depend on the between-cluster variance, and therefore, as expected, variations in $\overline{\widehat{\text{MOR}}}$ as prevalence changes are similar to those observed for $\overline{\widehat{\text{VPC}}_4}$.

Figure 4 displays the TCC expected means as well as ICC expected means obtained by using the analysis of variance approach. As expected, maximum ICC expected means were observed around the 0.5 prevalence value, and minimum values were observed when the prevalence approached 0 or 1. For

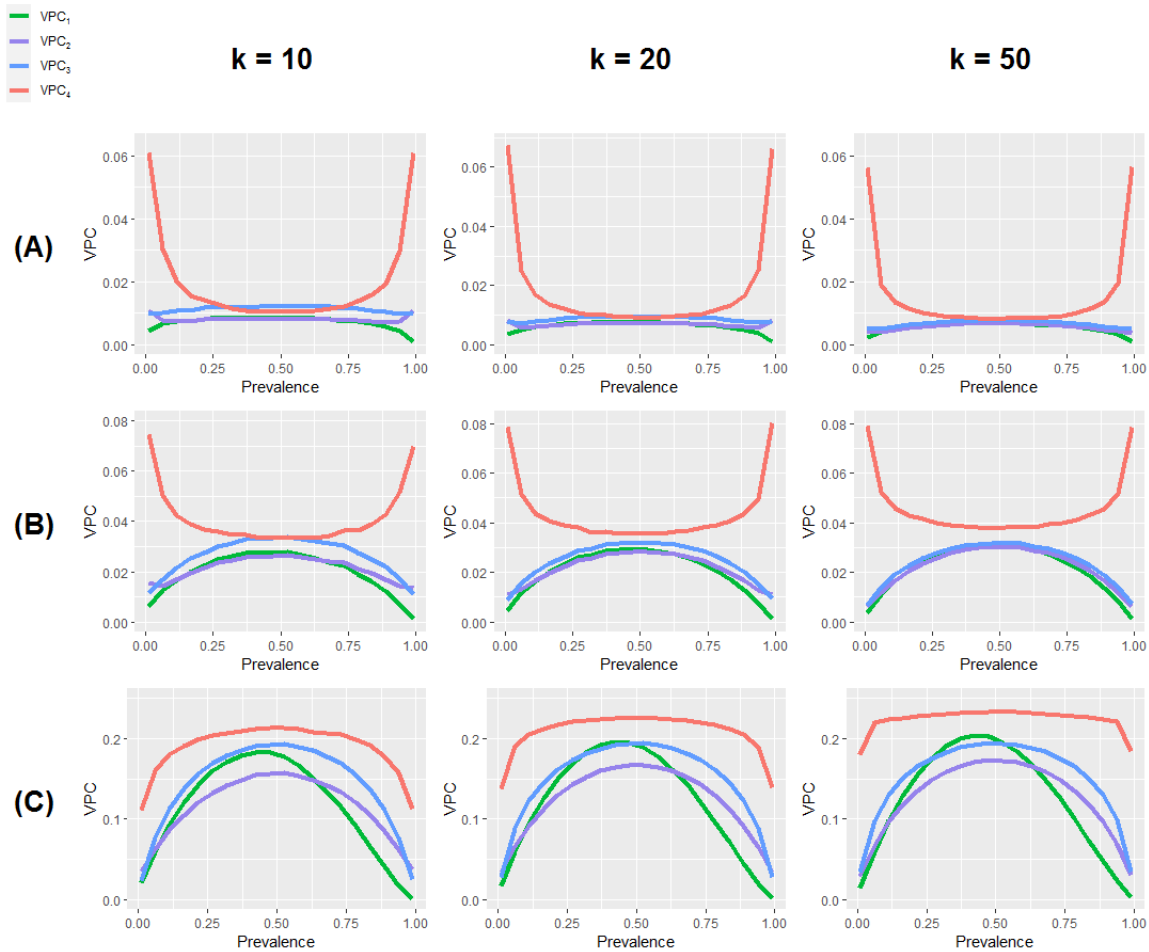

Figure 2. Variance partition coefficient expected mean as a function of outcome prevalence expected mean. These means were computed from 10000 simulated datasets. Three situations were considered for the underlying continuous outcome clustering level [ICC for continuous outcome equal to 0.01 **(A)**, 0.05 **(B)** or 0.3 **(C)**]. We considered cluster numbers of 10, 20 and 50, and cluster sizes were variable, with mean 25 and variance 225.

$\rho_Z = 0.05$, using Kirk's formula leads to intracluster correlation measures very stable over different prevalence values, except for extreme prevalences, and expected means are very close to the theoretical continuous ICC value. However, this is no longer true for other ρ_Z values. Otherwise using the analysis of variance approach to estimate an ICC leads to concave curves. Moreover, such an approach leads to underestimation of the theoretical continuous ICC (except when $\rho_Z = 0.01$ and $k = 10$), with a bias that increases with increasing ρ_Z. We plotted the ICC expected mean, when using an analysis of variance

p_{W}	$\rho_{\mathrm{Z}} = 0.01$			$\rho_{\mathrm{Z}} = 0.05$			$\rho_{\mathrm{Z}} = 0.3$		
	$k = 10$	$k = 20$	$k = 50$	$k = 10$	$k = 20$	$k = 50$	$k = 10$	$k = 20$	$k = 50$
0.01	60.68	53.5	48	57.98	51.73	41.64	56.21	46.4	32.61
0.06	59.49	52.82	44.55	49.25	37.47	20.24	35.35	20.05	4.03
0.11	58.06	50.63	39.47	42.75	26.99	9.85	22.38	8.18	0.57
0.16	57.79	48.38	34.72	35.61	20.03	5.11	13.49	2.89	0.06
0.22	55.07	45.04	32.52	30.31	15.24	2.37	7.79	1.16	0.01
0.27	54.24	44.46	30.24	26.62	10.89	1.35	4.32	0.44	0
0.32	54.28	44.03	29.1	23.49	9.03	0.63	2.72	0.08	0
0.37	53.18	42.32	28.44	21.68	7.35	0.4	1.72	0.02	0
0.42	53.05	42.9	27.14	20.58	6.53	0.25	0.75	0.01	0
0.47	52.38	41.78	27.39	20.06	6.2	0.27	0.56	0	0
0.53	52.54	42.58	27.18	19.84	6.08	0.27	0.48	0	0
0.58	53.12	42.25	27.21	20.81	6.88	0.28	0.81	0.02	0
0.63	53.54	42.42	28.08	22.01	7.65	0.51	1.55	0.07	0
0.68	53.94	43.22	29.49	23.89	8.91	0.64	2.55	0.14	0
0.73	54.39	45.76	31.34	26.9	11.41	1.24	4.64	0.35	0
0.78	55.18	46.03	32.33	30.78	14.93	2.81	7.68	1.27	0
0.84	56.87	48.53	35.53	35.07	19.56	4.94	13.71	3.14	0.08
0.89	58.36	50.08	39.1	40.96	26.62	10.07	22.63	7.6	0.46
0.94	59.23	51.44	43.85	48.8	37.01	19.8	36.89	20.01	4.05
0.99	61.64	54.85	47.63	59.71	51.89	42.65	57.22	47.49	31.91

Table 1. Proportion of cases over the 10000 runs where the estimated between-cluster variance $\widehat{\sigma}^2_\gamma$ was equal to zero. This proportion was computed for each prevalence value and for each of the 9 considered scenarios. These proportions are shown in percentage (%).

approach, which comes down to be equivalent to the approach used for $\widehat{\mathrm{VPC}}_3$. As a consequence, the green curves in Figure 4 are equivalent to the blue ones in Figure 2. The TCC estimated using the original approach overestimates the theoretical continuous ICC for very low ρ_{Z} values and underestimates it as soon as ρ_{Z} = 0.05. Finally, as for VPC_4 and MOR, the TCC expected mean curves were convex for low intracluster correlation values and concave for high values.

In Figure 5, red dots represent the proportion of datasets for which the estimated ICC value was actually lower than or equal to the theoretical maximum value. The lower the ρ_{Z}, the higher the proportion. In fact, it varies from 20% to 100% with the lowest proportions obtained for $\rho_{\mathrm{Z}} = 0.3$. This finding is due to the fact that the maximum theoretical possible ICC value (Figure 1) is always $\leq$ 1/3, whatever the prevalence. Thus, if $\rho_{\mathrm{Z}} = 0.3$ and the prevalence is equal to 0.5, the corresponding theoretical maximum value for ρ_{W}, computed using equation(9), is 0.19. Indeed there are many cases in which $\widehat{\rho}_{\mathrm{W}}$ is $>$ 1/3, or at least greater than this 0.19 value. For each ρ_{Z} and prevalence, we computed $\overline{\widehat{p}_{\mathrm{W}}}$ and $\overline{\widehat{R}_d}$ only when the estimated ICC was lower than or equal to its associated theoretical maximum value. Plots of $\overline{\widehat{R}_d}$ according to $\overline{\widehat{p}_{\mathrm{W}}}$ are then affected by this proportion of "eligible" datasets. As a consequence, especially when prevalences are extreme but also when ρ_{Z} = 0.3, results must be interpreted cautiously. However,

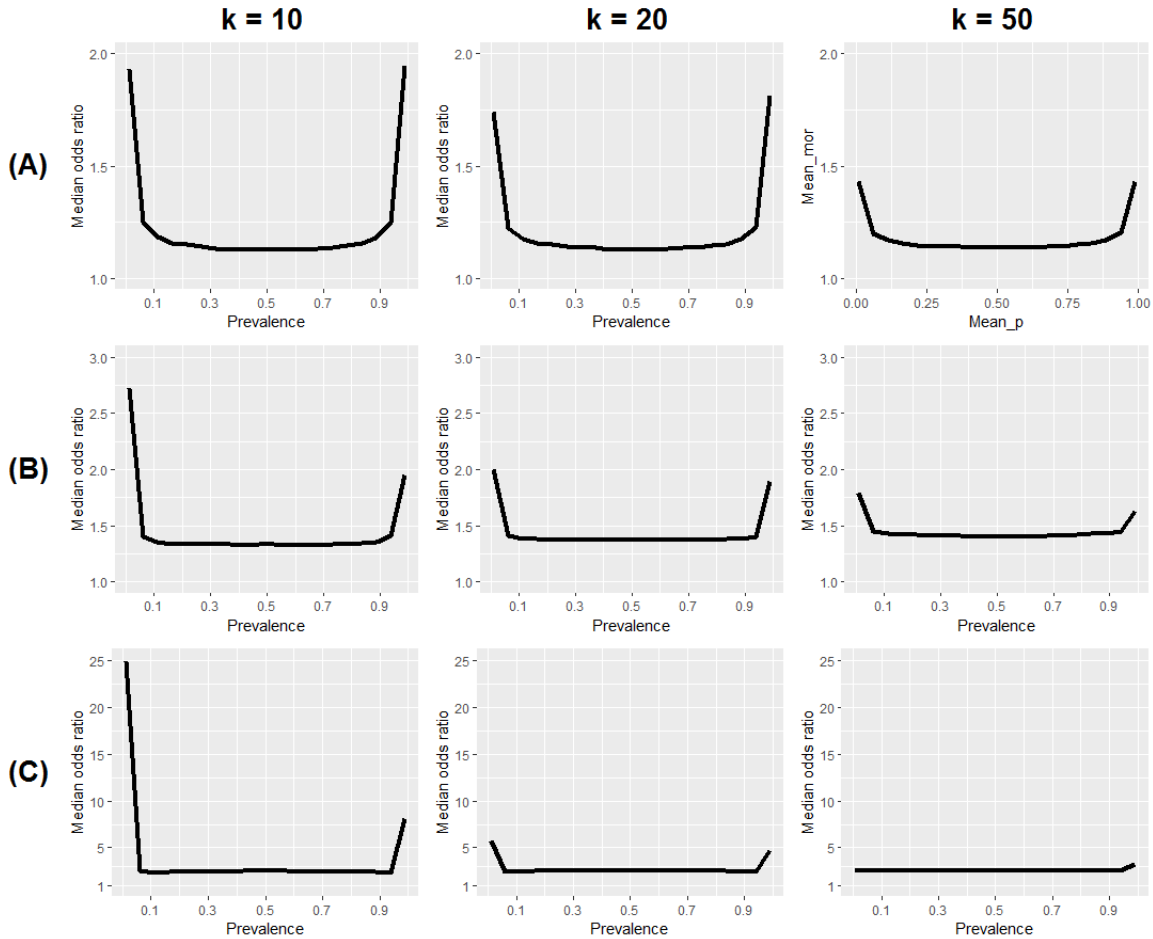

Figure 3. Median odds ratio expected mean as a function of outcome prevalence expected mean. These means were computed from 10000 simulated datasets. Three situations were considered for the underlying continuous outcome clustering level [ICC for continuous outcome equal to 0.01 **(A)**, 0.05 **(B)** or 0.3 **(C)**]. We considered cluster numbers of 10, 20 and 50, and cluster sizes were variable, with mean 25 and variance 225.

overall, the relative deviation of the ICC estimate to its theoretical maximum possible value expected mean varies as the outcome prevalence expected mean varies.

Generally, for a fixed ρ_Z, the number of clusters has no impact on the relationship between the different measure expected means and the outcome prevalence expected means. We also observed a symmetry of the different curves around the 0.5 prevalence value except for the curves associated with $\widehat{\text{VPC}}_1$ and $\widehat{\text{MOR}}$ in some scenarios.

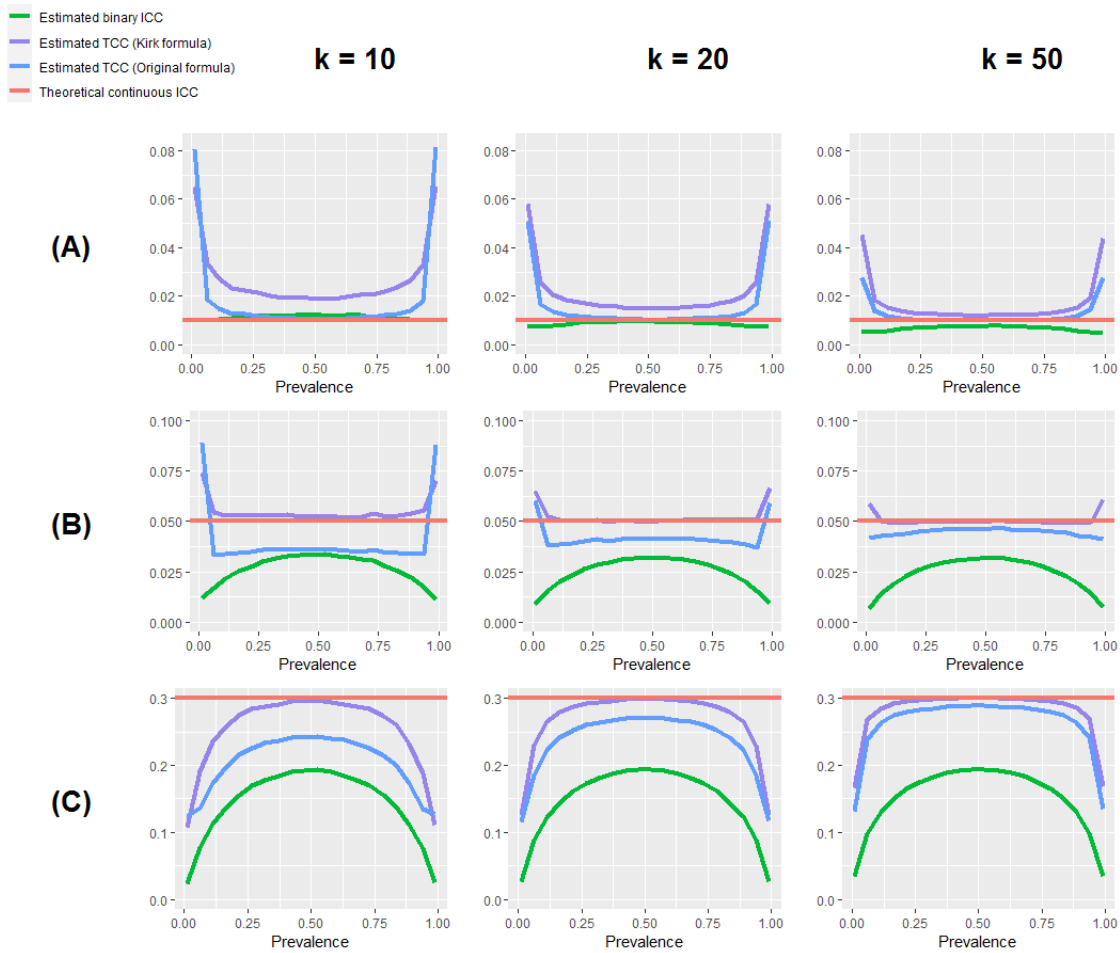

Figure 4. Tetrachoric correlation coefficient (TCC) and intraclass correlation coefficient (ICC) expected means as functions of outcome prevalence expected mean. These means were computed from 10000 simulated datasets. Three situations were considered for the underlying continuous outcome clustering level [ICC for continuous outcome equal to 0.01 **(A)**, 0.05 **(B)** or 0.3 **(C)**]. We considered cluster numbers of 10, 20 and 50, and cluster sizes were variable, with mean 25 and variance 225.

Even if the dependence on prevalence of these measures seems to be lower when removing extreme prevalence values, it still remains (supplementary files).

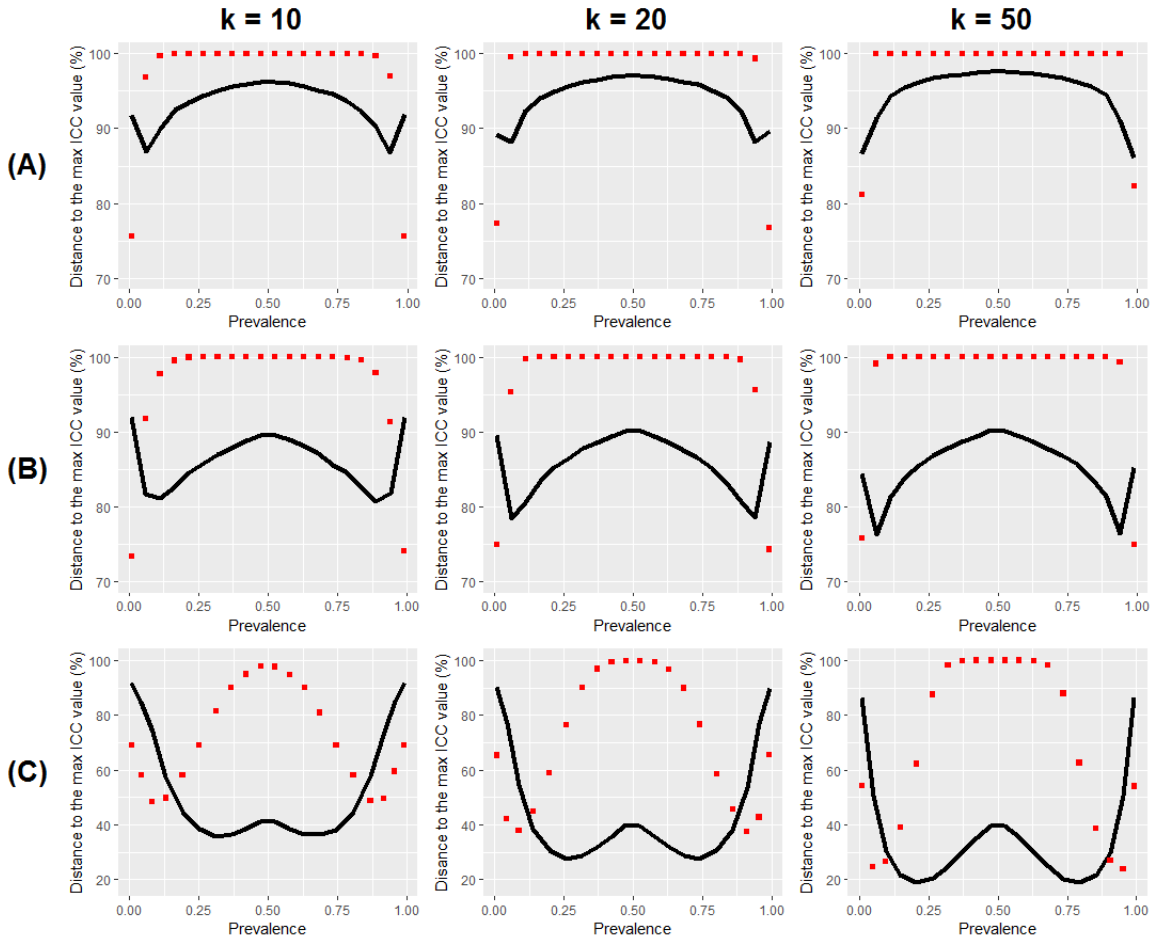

Figure 5. Relative deviation of the ICC estimate to its theoretical maximum possible value expected mean as a function of outcome prevalence expected mean (black curve) and proportion of estimated ICC values less than or equal to the theoretical maximum possible value expected mean as a function of the outcome prevalence expected mean (red dots). We simulated 10000 datasets and computed the mean over those for which the estimated ICC was less than or equal to the maximum possible ICC value. Three situations were considered for the underlying continuous outcome clustering level [ICC for continuous outcome equal to 0.01 **(A)**, 0.05 **(B)** or 0.3 **(C)**]. We considered cluster numbers of 10, 20 and 50, and cluster sizes were variable, with mean 25 and variance 225.

4 Comparison of intracluster correlation measures

Figures 2 to 5 allowed us to visualize the relationship between VPC_1, VPC_2, VPC_3, VPC_4, MOR, r, ρ_{Y} and R_d on one hand and prevalence on the other. However, they do not permit a direct comparison

Initial values	ρ_W/VPC_3	VPC_1	VPC_2	VPC_4	MOR	r	ρ_Y	R_d
$\rho_Z = 0.01$ and $k = 10$	1	1	0.42	1	1	1	1	1
$\rho_Z = 0.01$ and $k = 20$	1	1	0.61	1	1	1	1	1
$\rho_Z = 0.01$ and $k = 50$	1	1	1	1	1	1	1	1
$\rho_Z = 0.05$ and $k = 10$	1	1	1	0.8	0.61	0.5	0.49	0.61
$\rho_Z = 0.05$ and $k = 20$	1	1	1	1	0.73	0.5	0.61	0.61
$\rho_Z = 0.05$ and $k = 50$	1	1	1	1	1	1	0.31	1
$\rho_Z = 0.3$ and $k = 10$	1	1	1	1	0.5	1	1	1
$\rho_Z = 0.3$ and $k = 20$	1	1	1	1	0.61	1	1	0.61
$\rho_Z = 0.3$ and $k = 50$	1	1	1	1	0.61	1	1	0.41

Table 2. Maximal information coefficients between the different intracluster correlation measure expected means and the outcome prevalence expected mean.

between the studied measures of intracluster correlation, notably because scales are not the same from one Figure to another. To compare them, we studied the relationship between these measures and prevalence. The Pearson, Spearman and Kendall correlations were not suitable because as we could observe, the relationship between the different measures expected means and the outcome prevalence expected means were neither linear nor monotonic. We then opted for the maximal information coefficient (MIC) proposed by Reshef et al.[21] The MIC has been proposed as a measure of dependence for the relationship between two variables. It is particularly useful when we do not know the kind of the relationship to search for. It deals with parabolic and non monotonic relationships, as we previously observed. MIC takes values between 0 (independence) and 1 (strong dependence).

We then computed the MIC (for details, see Appendix 2) between the different intracluster correlation measure expected means and the outcome prevalence expected means. Results are shown in Table 2.

We noticed a strong dependence on prevalence of the estimated binary ICC, $\overline{\widehat{\rho_W}}$, in each scenario, (i.e., whatever the value of ρ_Z and that of k. For $\rho_Z = 0.01$, $\overline{\widehat{\text{VPC}}_2}$ was a bit less dependent on prevalence than the other measure expected means). For $\rho_Z = 0.05$, $\overline{\widehat{\text{MOR}}}$, $\overline{\widehat{r}}$, $\overline{\widehat{\rho_Y}}$ and $\overline{\widehat{R}_d}$ appeared to be less dependent on prevalence than $\overline{\widehat{\rho_W}}$, $\overline{\widehat{\text{VPC}}_1}$, $\overline{\widehat{\text{VPC}}_2}$ and $\overline{\widehat{\text{VPC}}_4}$. For $\rho_Z = 0.3$, $\overline{\widehat{\text{MOR}}}$ and $\overline{\widehat{R}_d}$ had the least dependence on prevalence. On the whole, $\overline{\widehat{\text{VPC}}_1}$ was as dependent on prevalence as $\overline{\widehat{\rho_W}}$ is. Overall, MIC values are quite high, whatever the situation considered and the intracluster correlation measure, which does not play in favor of independence between these intracluster correlation measures and outcome prevalence.

5 Example

We illustrated our simulation study by using an example from the Pithagore study[22]. The study is a cluster randomised trial conducted to test whether a multifaceted intervention aimed at increasing the translation into practice of a protocol for early management of postpartum haemorrhage, would reduce the incidence of severe postpartum haemorrhage. Fifty four maternity units were randomized to the intervention arm (a combination of outreach visits to discuss the protocol in each local context, reminders, and peer reviews of severe incidents) and 52 to the control arm (no intervention). We considered in this example a secondary outcome of the study, namely the administration of sulprostone due to uterine atony

	Control arm		Intervention arm	
	Estimate	95% CI	Estimate	95% CI
ρ_W/VPC$_3$	0.159	[0.075, 0.234]	0.085	[0.040, 0.133]
VPC$_1$	0.204	[0.090, 0.287]	0.098	[0.028, 0.173]
VPC$_2$	0.154	[0.078, 0.208]	0.082	[0.027, 0.139]
VPC$_4$	0.211	[0.102, 0.286]	0.109	[0.034, 0.184]
MOR	2.444	[1.793, 2.987]	1 .832	[1.381, 2.775]
r	0.211	[0.071, 0.396]	0.080	[0.022, 0.140
ρ_Y	0.248	[0.117, 0.362]	0.133	[0.063, 0.207]
R_d	49.38%	[23.87%, 75.73%]	73.08%	[58.18%, 87.21%]

Table 3. Estimated intracluster correlation measures for the administration of sulprostone in severe postpartum haemorrhage. The estimated prevalence was 0.459 in the control arm and 0.540 in the intervention arm. 95% CI, 95% confidence interval.

or retained placenta. There were 790 individuals in the control arm and 769 in the intervention arm. The estimated prevalence for the administration of sulprostone was 0.459 in the control arm versus 0.540 in the intervention arm.

The estimated intracluster correlation measures, as for the ICC, differed between the control and intervention arms (Table 3).

The ICCs were estimated as 0.159 in the control arm and 0.085 in the intervention arm. This difference in ICC estimates may reflect a standardization of practices in the intervention arm, which translates by a lower variability in maternity rates of administration of sulprostone.

The variance partition coefficient estimates varied between 0.154 and 0.211 in the control arm and 0.085 and 0.109 in the intervention arm. Therefore, 15.4% to 21.1% of the total variation in the administration of sulprostone was explained by the variation between maternity units in the control arm, versus 8.5% to 10.9% in the intervention arm.

The median estimate of the odds ratios between women with a higher probability to receive sulprostone and women with a lower probability was 2.444 in the control arm versus 1.832 in the intervention arm. This indicates a moderate correlation within maternity units.

The TCCs, with estimates of 0.211 and 0.248 in the control arm, and 0.080 and 0.133 in the intervention arm, also indicate a moderate correlation within maternity units of the intervention arm. The relative deviation of the ICC estimate to its theoretical maximum possible value was 49.38% in the control arm and 73.08% in the intervention arm.

6 Discussion

The aim of this work was to evaluate the dependence of different measures of intracluster correlation on prevalence in the context of cluster randomized trials with binary outcomes. The simulation results showed that all the proposed measures were associated with prevalence, although some measures seem slightly less associated than others, as shown by the estimated maximal information coefficients.

VPC$_1$, VPC$_2$, VPC$_4$ and MOR have been computed using an estimate of the between-cluster variance, which can be obtained by fitting a mixed-effects logistic regression model. This variance was often

estimated as zero, notably for low continuous ICC values, but also when prevalences were extreme. We decided to keep these between-cluster variance null estimates because we wanted to produce realistic results. Indeed, in practice, cluster-effects variance can be estimated at zero on some data even if the true value is not. Among the proposed VPCs, VPC_4 is, in practice, the most used.[23] However, it did not perform well in this work, probably because the approach used to simulate correlated binary data implicitly supposed a latent normal variable rather than a logistic one, as is supposed with VPC_4 .

Merlo et al. presented an example of health care utilisation in Sweden to show how to calculate and interpret several measures of variance, including the MOR, that are appropriate for investigating contextual phenomena of a binary nature.[24] They showed that the MOR provided more interpretable information than the ICC on the relevance of the residential area for understanding the individual propensity of consulting private physicians. They also concluded, using hypothetical data, that the MOR may be used for comparisons between studies with different prevalence. This latter result agrees with our simulation results in that the MOR was less dependent on prevalence than the ICC in some scenarios.

Regarding the TCC, estimated either using the historical approach or the Kirk's formula, even if in some cases they appeared to be less dependent on prevalence than the ICC, in the other cases they were not better. Concerning the original formulation of the TCC, we can explain this finding by the fact that the cluster randomized trial setting differs from the one for which Pearson defined the TCC.[8] Indeed, it was initially defined as the correlation between two continuous variables, rather than the intracluster correlation. However, to set it appropriate in our setting, we computed all possible pairs in each cluster. This transformation may have contributed to modify the TCC as initially defined, thus distorting its potential good properties in terms of dependence on prevalence. Concerning Kirk's adaptation of the TCC, we can explain the obtained results by the fact that we need first to estimate the ICC for binary data (with the analysis of variance approach) and then used the formula (9). Yet, it is well-known that the ICC estimators are biased.[25].

The relative deviation of the ICC estimate to its theoretical maximum possible value, for its part, is questionable because it is based on some hypotheses such as a beta distribution for cluster prevalences, or unimodality, which, in practice, are not always satisfied. Nevertheless, ICC estimates were nearly always lower than the theoretical maximum value, for an intracluster correlation ≤ 0.05 and prevalence between 0.1 and 0.9, which cover most of the situations in CRTs.

7 Conclusion

The dependence of ICC on prevalence is a challenging issue. We presented here some measures that seemed to be interesting to overcome this drawback of the ICC. Some approaches have advantages over others but in a very limited way. None of the studied measures copes satisfactorily with the dependence on prevalence.

References

1. Campbell MK, Elbourne DR and Altman DG. Consort statement: extension to cluster randomised trials. *Bmj* 2004; 328(7441): 702–708.
2. Donner A and Klar N. *Design and analysis of cluster randomization trials in health research.* London: Arnold, 2000.

3. Gulliford M, Adams G, Ukoumunne O et al. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *Journal of clinical epidemiology* 2005; 58(3): 246–251.
4. Mbekwe Yepnang AM, Caille A, Eldridge SM and Giraudeau B. Is the R coefficient of interest in cluster randomized trials with a binary outcome? *Statistical Methods in Medical Research* 2020; 29 (9): 2470–2480.
5. Crespi CM, Wong WK and Wu S. A new dependence parameter approach to improve the design of cluster randomized trials with binary outcomes. *Clinical Trials* 2011; 8(6): 687–698.
6. Goldstein H, Browne W and Rasbash J. Partitioning variation in multilevel models. *Understanding statistics: statistical issues in psychology, education, and the social sciences* 2002; 1(4): 223–231.
7. Larsen K, Petersen JH, Budtz-Jrgensen E and Endahl L. Interpreting parameters in the logistic regression model with random effects. *Biometrics* 2000; 56(3): 909–914.
8. Pearson K. I. Mathematical contributions to the theory of evolution.VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character* 1900; 195(262-273): 1–47.
9. Eldridge SM. Assessing, understanding and improving the efficiency of cluster randomised trials in primary care. Queen Mary, University of London; 2005.
10. Donner A and Koval JJ. The estimation of intraclass correlation in the analysis of family data. *Biometrics* 1980: 19– 25.
11. Donner A. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review/Revue Internationale de Statistique* 1986: 67–82.
12. Kirk DB. On the numerical approximation of the bivariate normal (tetrachoric) correlation coefficient. *Psychometrika* 1973; 38(2): 259–268.
13. Kraemer HC. Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika* 1979; 44(4): 461–472.
14. Donner A and Eliasziw M. Statistical implications of the choice between a dichotomous or continuous trait in studies of interobserver agreement. *Biometrics* 1994; 550–555.
15. Caille A, Leyrat C and Giraudeau B. Dichotomizing a continuous outcome in cluster randomized trials: impact on power. *Statistics in medicine* 2012; 31(24): 2822–2832.
16. Davis PJ and Rabinowitz P. *Methods of numerical integration.* Academic Press; 1984.
17. Commenges D, Jacqmin H. The intraclass correlation coefficient: distribution-free definition and test. *Biometrics* 1994: 517–526.
18. Krishnamoorthy K. *Handbook of Statistical Distributions with Applications*. Chapman and Hall/CRC; 2006.
19. Lunn AD and Davies SJ. A note on generating correlated binary variables. *Biometrika* 1998; 85(2): 487–490.
20. Eldridge SM, Ashby D and Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *International journal of epidemiology* 2006; 35(5): 1292–1300.
21. Reshef DN, Reshef YA, Finucane HK et al. Detecting novel associations in large data sets. *science*. 2011; 334(6062): 1518–1524.
22. Deneux-Tharaux C, Dupont C, Colin C et al. Multifaceted intervention to decrease the rate of severe postpartum haemorrhage: the PITHAGORE6 cluster-randomised controlled trial. *BJOG: An International Journal of Obstetrics & Gynaecology* 2010; 117(10): 1278–1287.
23. Austin PC, Merlo J. Intermediate and advanced topics in multilevel logistic regression analysis. *Statistics in medicine* 2017; 36(20): 3257–3277.
24. Merlo J, Chaix B, Ohlsson H et al. A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of*

Epidemiology & Community Health 2006; 60(4): 290–297.
25. Ridout MS, Demetrio CG and Firth D. Estimating intraclass correlation for binary data. *Biometrics* 1999; 55(1): 137–148.
26. Kinney JB and Atwal GS. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences* 2014; 111(9): 33543359.

Appendix 1: r estimation

Considering the simulated binary variable W_{ij}, the Pearson ICC estimator consists in computing the Pearson correlation coefficient over all possible pairs of observations that can be constructed within clusters[11]. Thus, in each cluster i, we constructed $n_{w_i}(n_{w_i}-1)$ pairs of observation. After constructing all pairs, we had $\sum_{i=1}^{k} n_{w_i}(n_{w_i}-1)$ pairs of observations. These pairs are equivalent to 2 binary variables of length $\sum_{i=1}^{k} n_{w_i}(n_{w_i}-1)$ from 2 normal continuous variables, which allowed us to compute the r coefficient as defined in equation (8).

Appendix 2: MIC calculation

Let D be a finite set of ordered pairs. The x-values of D can be partitioned into x bins and the y-values into y bins, empty bins being allowed. Such a partition leads to an x-by-y grid. For a given grid G, we denoted by $D|G$ the distribution of the points of D on the xy cells of G. $D|G$ is defined so that the probability mass in each cell is the fraction of points in D that falls in that cell. For a fixed D and for a fixed $\{x, y\}$, different grids G can be constructed, leading to different distributions $D|G$.

The maximal information coefficient (MIC) is then defined by the following formula[21]

$$\text{MIC}(D) = \max_{xy<E(n_{\text{D}})} \left\{ \frac{\max[I(D|G)]}{\log(\min\{x,y\})} \right\} \tag{13}$$

with $I(D|G)$ being equal to[26]

$$I(D|G) = \int_{-\infty}^{+\infty} f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy$$

where $f(x,y)$ is the joint probability distribution of x and y, $f(x)$ and $f(y)$ are the marginal distributions of $f(x,y)$ and $E(n_{\text{D}})$ is a function of n_{D}, the number of pairs. Reshef et al. set $E(n_{\text{D}})$ to $n_{\text{D}}^{0.6}$. The maximum on the numerator of equation (13) is taken over all grids G with x columns and y rows. $I(D|G)$, named the mutual information of $D|G$, represents the amount of information that the value of one variable reveals about the value of the other. The principle behind the MIC is to compute the highest normalized mutual information achieved by any x-by-y grid, then to compute the maximum over the highest values obtained when varying x and y.