Is the R Coefficient of Interest in Cluster Randomized Trials with a Binary Outcome?

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

Is the R coefficient of interest in cluster randomized trials with a binary outcome?

Journal Title
XX(X):1–14

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

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

Abstract

In cluster randomized trials, the intraclass correlation coefficient (ICC) is classically used to measure clustering. When the outcome is binary, the ICC is known to be associated with the prevalence of the outcome. This association challenges its interpretation and can be problematic for sample size calculation. To overcome these situations, Crespi et al. extended a coefficient named R, initially proposed by Rosner for ophthalmologic data, to cluster randomized trials. Crespi et al. asserted that R may be less influenced by the outcome prevalence than is the ICC, although the authors provided only empirical data to support their assertion. They also asserted that "the traditional ICC approach to sample size determination tends to overpower studies under many scenarios, calling for more clusters than truly required", although they did not consider empirical power. The aim of this study was to investigate whether R could indeed be considered independent of the outcome prevalence. We also considered whether sample size calculation should be better based on the R coefficient or the ICC. Considering the particular case of 2 individuals per cluster, we theoretically demonstrated that R is not symmetrical around the 0.5 prevalence value. This in itself demonstrates the dependence of R on prevalence. We also conducted a simulation study to explore the case of both fixed and variable cluster sizes greater than 2. This simulation study demonstrated that R decreases when prevalence increases from 0 to 1. Both the analytical and simulation results demonstrate that R depends on the outcome prevalence. In terms of sample size calculation, we showed that an approach based on the ICC is preferable to an approach based on the R coefficient because with the former, the empirical power is closer to the nominal one. Hence, the R coefficient

[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]*

does not outperform the ICC for binary outcomes because it does not offer any advantage over the ICC.

Keywords
Intraclass correlation coefficient, R coefficient, binary outcome, prevalence, cluster

1 Introduction

Cluster randomized trials are increasingly being used in health research. In such a setting, clusters of individuals are randomly allocated to different arms[1]. Clusters may be families, schools, worksites, medical practices, towns or other social units. In such trials, outcomes for individuals from the same cluster are more similar than are outcomes for individuals from different clusters.

The intraclass correlation coefficient (ICC) is classically used to measure this resemblance. It can be defined as the proportion of total variance due to between-cluster variation or the correlation between any 2 members of the same cluster[1,2]. An ICC equal to 0 indicates independence among individuals of a cluster, whereas an ICC equal to 1 indicates that individuals from a given cluster have identical outcomes.

The Consolidated Standards for Reporting of Trials (CONSORT) extension for cluster randomized trials recommends reporting a measure of intracluster correlation, such as the ICC, for each primary outcome[3]. This has actually two aims. The first is that it may help interpret the results of the trial. Indeed, the assessed intervention may affect the level of clustering, and this result is important for a complete interpretation of trial result. When the outcome is binary, the ICC is known to be associated with the prevalence of the outcome[4]. As the prevalence increases from 0 to 0.5, the ICC increases. Because of this association, ICC values are expected to differ when prevalences differ, even when the clustering level remains identical. This association with prevalence challenges the interpretation of the ICC because ICC values do not just depend on clustering level.

The second reason for providing clustering estimates is that such values are of help for sample size calculation of future studies. Yet, the association between the ICC and the outcome prevalence can be problematic in sample size calculation if the study to be planned is expected to have prevalences different from those from which we derived ICC estimates.

To overcome these situations, Crespi et al.[5] extended a coefficient named R, initially proposed by Rosner[6] for ophthalmologic data, to cluster randomized trials. 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. Crespi et al. asserted that R may be less influenced by the outcome prevalence than the ICC. To support this assertion, the authors provided an illustration with an example, stating that mathematical proof was not possible. Moreover, they proposed sample size formulas using R coefficients or ICCs and used these formulas to calculate required sample size in diverse situations. They concluded that the "the traditional ICC approach to sample size determination tends to overpower studies under many scenarios, calling for more clusters than truly required". However, this latter conclusion was based on sample size calculations without any consideration of empirical power.

The aim of this study was to investigate whether R is indeed independent of the outcome prevalence. We also investigated which sample size calculation, based on the R coefficient or the ICC, provides the empirical power closest to the nominal power.

We define R in section 2 and provide its estimator in section 3. In section 4.1, we explore theoretically the relation between R and the outcome prevalence in the special case of clusters of size 2. In section 4.2, we report a simulation study to explore the situation of cluster sizes greater than 2, both fixed and variable. An illustration using real data is provided in section 5. In section 6.1, we compare sample size calculation using R or the ICC in another simulation study and in section 6.2, we illustrate the asymmetry of R in sample size calculation. We conclude with a short discussion in section 7.

2 Definitions

In this section up to and including section 4, we will consider one arm composed of k clusters of size $n_i (i = 1, 2, ..., k)$. Let X_{ij} be the outcome of the jth, $j = 1, 2, ..., n_i$ individual in the ith cluster, with X_{ij} a binary variable whose possible values are 1 for success and 0 for failure. $X_i = \sum_{j=1}^{n_i} X_{ij}$ is the number of successes in cluster i and $N = \sum_{i=1}^{k} n_i$ is the total number of individuals. We also assume that the success probability p is the same for all individuals (i.e., $\mathrm{P}(X_{ij} = 1) = p$).

2.1 R as defined by Rosner

Rosner[6] worked on methods for analysing ophthalmologic data. In this special case, the cluster unit is the individual, with 2 observations (eyes) per individual. Rosner defined R as:

$$\mathrm{P}(X_{ij} = 1 | X_{ij'} = 1) = Rp, \quad j, j' = 1, 2 \text{ and } j \neq j'. \tag{1}$$

R is a measure of dependence between 2 eyes of the same person. If $R = 1$, the outcome of the 2 eyes from a given individual are independent, but if $Rp = 1$, the outcome of the 2 eyes from a given individual are identical, whatever the individual. The lower bound of R is 1 (provided the 2 observations are positively correlated) and the upper bound is $\frac{1}{p}$.

2.2 Crespi's extension of R

Crespi et al. extended the formula from Rosner to the case of clusters of fixed size (m) potentially greater than 2[5]:

$$\mathrm{P}(X_{ij} = 1 | X_{ij'} = 1) = Rp, \quad j, j' = 1, 2, ..., m \text{ and } j \neq j'. \tag{2}$$

R can be seen as a quantification of how much or less likely a member of a cluster is to be successful given that another member of the cluster is successful.

3 Estimating R

3.1 The Rosner R estimator: $\widehat{R}_r$

In the special case of clusters of fixed size $m = 2$, Rosner[6] showed that the maximum likelihood estimator of R is:

$$\widehat{R}_r = \frac{4k k_2}{(k_1 + 2k_2)^2}, \tag{3}$$

where k_0 is the number of clusters with no success, k_1 the number of clusters with 1 success, k_2 the number of clusters with 2 successes and thus, $k = k_0 + k_1 + k_2$.

3.2 The Crespi R estimator: $\widehat{R}_c$

Under the common correlation model, the ICC has been defined as[7]:

$$\rho = \frac{\mathrm{P}(X_{ij} = 1 | X_{ij'} = 1) - p}{1 - p}, \quad j \neq j'. \tag{4}$$

Therefore, from equation (2) we derive that:

$$R = 1 + \frac{\rho(1-p)}{p}. \tag{5}$$

Crespi et al. suggested that R be estimated by using $\widehat{\rho}$ and $\widehat{p}$ estimates of both ρ and p, respectively.

Many methods have been proposed to estimate the ICC for binary outcomes[8]. Simulation results reported by Ridout et al. showed that the ANOVA estimator, the Fleiss-Cuzick estimator and some of the moment estimators performed well in terms of bias, standard deviation and mean square error.

When using the Fleiss-Cuzick estimator and considering clusters of size 2, the R estimator using the approach proposed by Crespi et al. is equivalent to its maximum likelihood estimator (3) as proposed by Rosner. Therefore, we used the Fleiss-Cuzick estimator defined as:

$$\widehat{\rho}_{\mathrm{FC}} = 1 - \frac{\sum_{i=1}^{k} X_i(n_i - X_i)/n_i}{(N-k)\widehat{p}(1-\widehat{p})}. \tag{6}$$

4 Relation between R and outcome prevalence

4.1 Exploration of the symmetry of the R estimator around a prevalence of 0.5

In this section, we consider a dataset e containing k_e clusters, each with two observations of a binary outcome. The estimated prevalence of success is $\widehat{p}_e$. If we are interested in measuring the clustering for success, the associated R estimate is (cf. (3)):

$$\widehat{R}_{c,\widehat{p}_e} = \frac{4k_e k_{2e}}{(k_{1e} + 2k_{2e})^2}.$$

Let us now consider that we are interested in measuring the clustering for failure rather than success; thus, all 0 values are replaced by 1 and vice versa. The associated estimate prevalence of failure is $1 - \widehat{p}_e$.

It can be shown that $\widehat{R}_{c,1-\widehat{p}_e} = \widehat{R}_{c,\widehat{p}_e}$ if and only if $\widehat{p}_e = 1/2$ or $\widehat{R}_{c,1-\widehat{p}_e} = \widehat{R}_{c,\widehat{p}_e} = 1$ (Appendix), which means that apart from these situations, $\widehat{R}_{c,1-\widehat{p}_e} \neq \widehat{R}_{c,\widehat{p}_e}$ and there is no symmetry. Conversely, if one focuses on the ICC estimator, defined as $\widehat{\rho}_{\mathrm{FC}} = 1 - \frac{k_{1e}}{2k_e\widehat{p}_e(1-\widehat{p}_e)}$ in case clusters are of fixed size of 2, symmetry around the 0.5 prevalence value is evident.

For example, if we consider the situation of 100 clusters with $k_{2e} = 6$, $k_{1e} = 18$ and $k_{0e} = 76$, the estimated success prevalence is 0.15 and $\widehat{R}_{c,\widehat{p}_e} = 2.64$. After permutating between 1s and 0s values, $k_{2e} = 76$, $k_{1e} = 18$ and $k_{0e} = 6$, the estimated failure prevalence is 0.85 and $\widehat{R}_{c,1-\widehat{p}_e} = 1.05$. Conversely, in both cases, the ICC estimate is 0.29. This asymmetry around a prevalence of 0.5 of the R coefficient is sufficient to conclude that it depends on prevalence.

Moreover, this asymmetry is counterintuitive because we would expect that the degree of intracluster resemblance to be the same whether we consider the resemblance in success or failure for a given dataset.

4.2 Simulation study

In the previous section, we showed that R is associated with outcome prevalence, for clusters of size 2. We then investigated the shape of the relation between R and prevalence. We considered the most general case of cluster sizes greater than 2, both fixed and variable. To this end, we conducted a simulation study according to the following principle. We generated correlated binary data with pre-specified outcome prevalence p and intraclass correlation ρ_{bin}, where ρ_{bin} is the ICC associated with the binary outcome X_{ij}. Because we wanted to obtain datasets with the same level of clustering whatever the outcome prevalence, we associated X_{ij} to a latent normal continuous outcome $Y_{ij} \sim \mathcal{N}(\mu, \sigma^2)$ in such a way that $X_{ij} = 1$ if $Y_{ij} > \mu + h\sigma$ and $X_{ij} = 0$ if $Y_{ij} \leq \mu + h\sigma$, where h is a constant such as $p = 1 - \Phi(h)$, Φ being the cumulative distribution function of the standard normal distribution. In this context, from works of Kirk[9] and Kraemer[10], Donner and Eliasziw[11] reported that

$$\rho_{\text{bin}} = \frac{1}{2\pi p(1-p)} \int_0^{\rho_{\text{cont}}} \frac{1}{\sqrt{1-x^2}} \exp\left(\frac{-h^2}{1+x}\right) dx, \tag{7}$$

where ρ_{bin} is the ICC associated with the binary outcome X_{ij} and ρ_{cont} is the ICC associated with the underlying continuous outcome Y_{ij}. ρ_{cont} corresponds to the tetrachoric correlation coefficient (i.e., the correlation coefficient associated with the latent variable underlying the binary outcome of interest).

We first specified ρ_{cont}, then we varied p and calculated ρ_{bin} for each value of p using (7). Doing so allowed for having the common underlying level of clustering equal to ρ_{cont} for each value of p. Thus, for each pair (p, ρ_{bin}), we defined X_{ij} as[12]:

$$X_{ij} = (1 - U_{ij})V_{ij} + U_{ij}Z_i, \tag{8}$$

where $U_{ij} \sim \text{Binom}(1, \sqrt{\rho_{\text{bin}}})$, $V_{ij} \sim \text{Binom}(1, p)$ and $Z_i \sim \text{Binom}(1, p)$.

4.2.1 Simulation plan

Steps of the data generation for each pair (p, ρ_{bin}) were as follows:

1. For the following cluster sizes:
   - **a.** Variable cluster sizes: simulate n_i cluster sizes, $i = 1, ..., k$, from a negative binomial distribution with mean m and variance v; m then corresponds to mean cluster sizes.
   - **b.** Fixed cluster sizes: set $n_i = m, i = 1, ..., k$.
2. For each cluster, simulate $Z_i, i = 1, ..., k$, under a binomial distribution with parameters 1 and p.
3. For each individual, simulate $V_{ij}, i = 1, ..., k, \quad j = 1, ..., n_i$ under a binomial distribution with parameters 1 and p.
4. For each individual, simulate $U_{ij}, i = 1, ..., k, \quad j = 1, ..., n_i$ under a binomial distribution with parameters 1 and $\sqrt{\rho_{\text{bin}}}$.
5. Calculate X_{ij} according to equation (8).

We varied p between 0.01 and 0.99. Statistical analyses were conducted with the three following steps:

1. Estimate p as $\widehat{p} = (\sum_{i=1}^{k} X_i)/N$.
2. Estimate ρ_{FC} as $\widehat{\rho}_{\text{FC}}$ using the Fleiss-Cuzick estimator.
3. Calculate $\widehat{R}_c$ as $\widehat{R}_c = 1 + \widehat{\rho}_{\text{FC}} \frac{1-\widehat{p}}{\widehat{p}}$.

All negative values of $\widehat{\rho}_{\text{FC}}$ were truncated to 0, then associated $\widehat{R}_c$ values were equal to 1.

We generated 50000 datasets for each scenario and for each value of p, we summarized results by computing $\overline{\widehat{p}}$, $\overline{\widehat{\rho}_{\text{FC}}}$ and $\overline{\widehat{R}_c}$, the empirical means of the estimated $\widehat{p}$, $\widehat{\rho}_{\text{FC}}$ and $\widehat{R}_c$, respectively.

Simulations were run considering three initial values of ρ_{cont} (0.01, 0.05, 0.3) over the generated datasets, which are realistic values observed in cluster randomized trials, and for each value, cluster numbers of k =10, 20 and 50. We considered the case of fixed cluster sizes of 25. We also generated 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 realistic one[13]. This led to 18 scenarios considered.

4.2.2 Simulation results

Figure 1 displays three plots ρ_{bin}, $\overline{\widehat{\rho}_{\text{FC}}}$ and $\overline{\widehat{R}_c}$ as a function of $\overline{\widehat{p}}$ and for three values of ρ_{cont}. As expected, maximum values of the estimated binary ICC were reached around the 0.5 prevalence value, and minimum values were reached when the prevalence approached 0 or 1. The Fleiss-Cuzick method underestimated the ICC, except for the case of small values of ρ_{bin} (around 0.005) probably because negative estimates were truncated at 0. Furthermore, when $\rho_{\text{cont}} = 0.01$ and for extreme prevalence

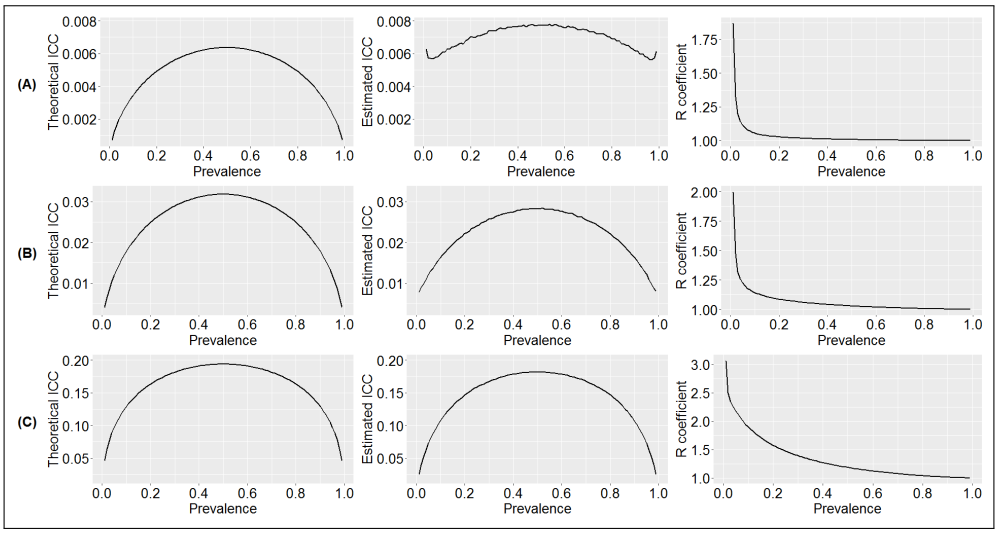

Figure 1. Theoretical intraclass correlation coefficient (ICC) ρ_{bin}, mean of estimated ICCs $\overline{\widehat{\rho}_{\text{FC}}}$ and mean R coefficient estimates $\overline{\widehat{R}_c}$ for binary outcomes as a function of estimated outcome prevalence. These means were computed from 50000 simulated datasets. Three situations were considered for the underlying continuous outcome clustering level [ρ_{cont} = 0.01 **(A)**, 0.05 **(B)** or 0.3 **(C)**]. Cluster sizes were variable, with mean 25 and variance 225, and we considered 20 clusters.

values, $\overline{\widehat{\rho}_{\text{FC}}}$ appeared to be higher than for less extreme values. This was an artefact due to the proportion of truncated estimates. Indeed, for $p = 0.01$ or $p = 0.99$, about 65% of negative ICC estimates were truncated to 0, whereas for $p = 0.1$ or $p = 0.9$, about 52% of negative ICC estimates were truncated to 0. The R coefficient decreased with increasing prevalence value. When the prevalence tended to 0, R tended to infinity, and when the prevalence tended to 1, R tended to 1. Results were qualitatively the same whatever the underlying continuous variable ICC or the number of clusters and whether cluster sizes were fixed or variable (supplementary files). Both $\overline{\widehat{\rho}_{\text{FC}}}$ and $\overline{\widehat{R}_c}$ depend on prevalence, but R is not symmetric, which invites a preference of the ICC over the R coefficient.

5 Example

To empirically illustrate the relation between R and the success prevalence, we used data from the Health Services Research Unit in Aberdeen (https://www.abdn.ac.uk/hsru/what-we-do/tools/index.php#panel177). These data provide estimates of ICCs from changing professional practice studies. Clusters were hospitals, hospital units, hospital directorates, general practices, physicians or pharmacies.

We used 145 ICCs from binary outcomes and associated prevalence values. ICCs ranged from 0 to 0.659 (median 0.057, interquartile range [IQR] 0.012–0.105). Prevalence values ranged from 0.032 to 0.995 (median 0.452, IQR 0.209–0.819). We estimated R by using equation (5). R values ranged from 1 to 4.217 (median 1.084, IQR 1.006–1.243).

We plotted R on the logarithm scale as a function of prevalence (Figure 2). The strength of the association between R and prevalence was estimated by the Spearman correlation coefficient. The estimated correlation coefficient was -0.721 (95% CI -0.833 – -0.580, p-value < 0.001).

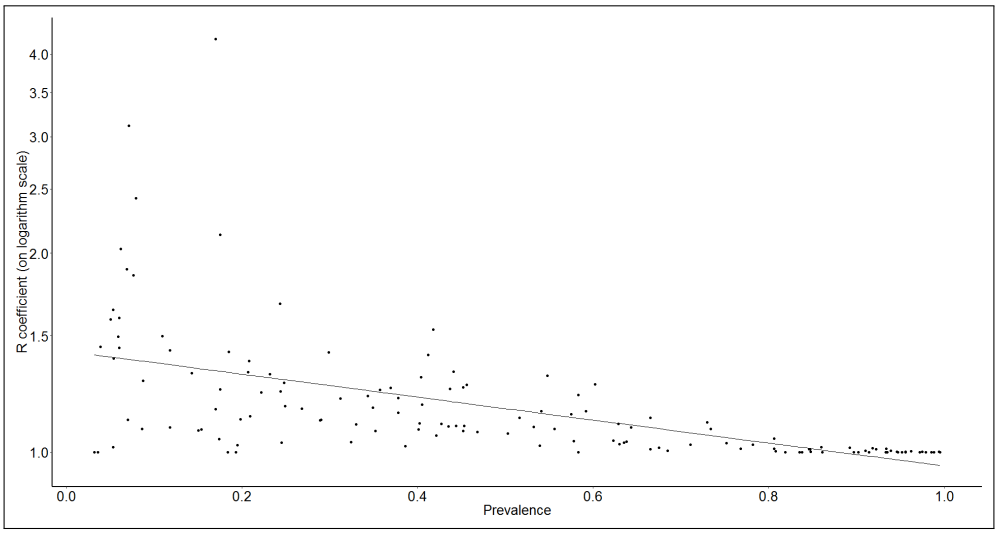

Figure 2. Association between R and prevalence by using data from the Health Technology Assessment review. The Spearman correlation coefficient was estimated at -0.721 (95% CI -0.833 – -0.580).

6 Sample size considerations using R

6.1 Empirical power in sample size calculation when using R

Crespi et al.[5] proposed sample size formulas with the R coefficient or ICC and used them to calculate required sample sizes in diverse situations. The authors considered three approaches: 1) one based on two R coefficients, that is, one for each arm (R-based approach); 2) one based on two ICCs (ICC A approach) and 3) one based on an ICC assumed to be common to the two arms (ICC B approach). They also considered two situations: 1) the prevalence levels of the study to be planned differ from those of the study previously conducted and from which R coefficients and ICCs have been estimated and 2) the prevalence levels are identical. The required number of clusters differed according to the approach used for sample size calculation. Focusing on the first situation (i.e, change in prevalence level between the previously conducted study and the planned one), the authors concluded that: 1) "when moving from high to moderate or from moderate to low prevalence, the ICC approaches can grossly overpower the study, calling for many more clusters than required to achieve desired power" and that 2) "when moving to a higher prevalence setting, ICC A underpowers the study". However, the approach used by Crespi et al. is debatable. Indeed, they derived three required sample sizes using the three previously cited approaches. Then, using the sample size formula based on two R coefficients, they derived power associated with the three sample sizes previously calculated. As a consequence, they observed a power close to 80% for the approach based on two R coefficients (slight deviations from 80% are due to rounding of the number of clusters). For the two other approaches, power was greater than 80% because the required number of clusters was higher when using an approach based on ICCs than the approach based on the R coefficient. However, doing so does not demonstrate anything, because Crespi et al. did not check whether the empirical power actually equals the nominal power. Therefore, we investigated whether Crespi et al.'s assertions were correct, estimating the empirical power associated with each situation they considered. Indeed, claiming that approach A is overpowered only because it requires a larger sample size than approach B is not correct. This is true only if the empirical power associated with approach B equals the nominal one, which was not verified by Crespi et al. Therefore, we performed a simulation study to assess which approach is preferable.

6.1.1 Simulation plan

1. Let us consider that a two-arm cluster randomized trial has already been previously conducted and prevalences were p_1 and p_2 in arm 1 and 2, respectively, with associated R_1 and R_2 coefficients. Given that p_1, p_2, R_1 and R_2, ρ_1 and ρ_2 can be derived by using equation (5). In practice, p_1, p_2, R_1, R_2, ρ_1 and ρ_2 can be replaced by their associated estimates, and this holds true for the upcoming issues.

2. We plan to conduct a new study with expected prevalences p'_1 and p'_2. For this, we calculate k_R, k_{ICCA} and k_{ICCB}, the required number of clusters, by using the following formulas:

   - R-based approach

$$k_R = \lceil \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 (p'_1(1 - p'_1 + (m-1)(R_1 - 1)p'_1) + p'_2(1 - p'_2 + (m-1)(R_2 - 1)p'_2))}{m(p'_1 - p'_2)^2} \rceil \quad (9)$$

- ICC A approach

$$k_{ICCA} = \lceil \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 (p'_1(1-p'_1)(1+(m-1)\rho_1) + p'_2(1-p'_2)(1+(m-1)\rho_2))}{m(p'_1 - p'_2)^2} \rceil \quad (10)$$

- ICC B approach

$$k_{ICCB} = \lceil \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 (p'_1(1-p'_1) + p'_2(1-p'_2))(1+(m-1)\rho_{comb})}{m(p'_1 - p'_2)^2} \rceil \quad (11)$$

with m the fixed cluster sizes, α and β the type I and type II error, respectively; $z_{1-\alpha/2}$ and $z_{1-\beta}$ the quantiles of the standard normal distribution corresponding to probability values of $1-\alpha/2$ and $1-\beta$, respectively; and $\lceil a \rceil$ the ceiling function giving the smallest integer $\geq a$.

Of note, for the latter formula we computed ρ_{comb} as $(\rho_1 + \rho_2)/2$ as already advised by Donald and Donner[14] rather than using Crespi et al.'s approach. Indeed, their approach is valid only with no intervention effect (i.e, $p_1 = p_2$). Otherwise, it may lead to a ρ_{comb} value that may even be greater than both ρ_1 and ρ_2. This latter phenomenon is known and has been illustrated by Giraudeau[15] for continuous outcomes.

3. We estimated the empirical power associated with each sample size. For this, we used the same simulation plan as that we described in section 4.2. Considering p_1 and ρ_1, we derived, using formula (7), an estimate of $\rho_{\text{cont},1}$, the ICC associated with the latent variable underlying the binary outcome in arm 1. Then, considering p'_1 (the prevalence we hypothesized to observe in arm 1 in the future study) and $\widehat{\rho}_{\text{cont},1}$, we derived, again using formula (7), ρ'_1, the expected ICC for binary data we expect to observe in arm 1 if prevalence is p'_1. We then simulated binary correlated data as in section 4.2 with prevalence equal to p'_1 and ICC equal to ρ'_1. The same approach was used for the second arm. Each dataset was analyzed by using an adjusted chi-square test[2]. Empirical power was estimated as the proportion of the 5000 generated datasets for which a significant (p-value < 0.05) result was observed.

Simulations were run considering the same scenarios as Crespi et al., when clusters are of size $m = 20$. We considered the situation in which prevalences in the future study are expected to be different from those of the previously conducted study (from $(p_1, p_2) = (0.7, 0.5)$ to $(p'_1, p'_2) = (0.5, 0.3)$, from $(p_1, p_2) = (0.5, 0.3)$ to $(p'_1, p'_2) = (0.3, 0.1)$, from $(p_1, p_2) = (0.3, 0.1)$ to $(p'_1, p'_2) = (0.5, 0.3)$ and from $(p_1, p_2) = (0.5, 0.3)$ to $(p'_1, p'_2) = (0.7, 0.5)$) and the situation in which they are expected to be equal ($(p'_1, p'_2) = (p_1, p_2) = (0.7, 0.5)$, $(p'_1, p'_2) = (p_1, p_2) = (0.5, 0.3)$ and $(p'_1, p'_2) = (p_1, p_2) = (0.3, 0.1)$). R_1 and R_2 were chosen to be equal ($(R_1, R_2) \in \{1.04, 1.08, 1.12, 1.16, 1.20\}^2$ and $R_1 = R_2$) or different ($R_1 \in \{1.02, 1.04, 1.06, 1.08, 1.10\}$ and $R_2 = 1.2$, $R_2 \in \{1.02, 1.04, 1.06, 1.08, 1.10\}$ and $R_1 = 1.2$). From these prevalence and R values, ρ_1 values ranged from 0.009 to 0.467 and ρ_2 values from 0.002 to 0.2. α and β were chosen to be equal to 0.05 and 0.2, respectively.

All programming was implemented by using R software, v3.6.1. Code is available on `https://github.com/Mbekwe/Simulation-study-with-R-software.git`.

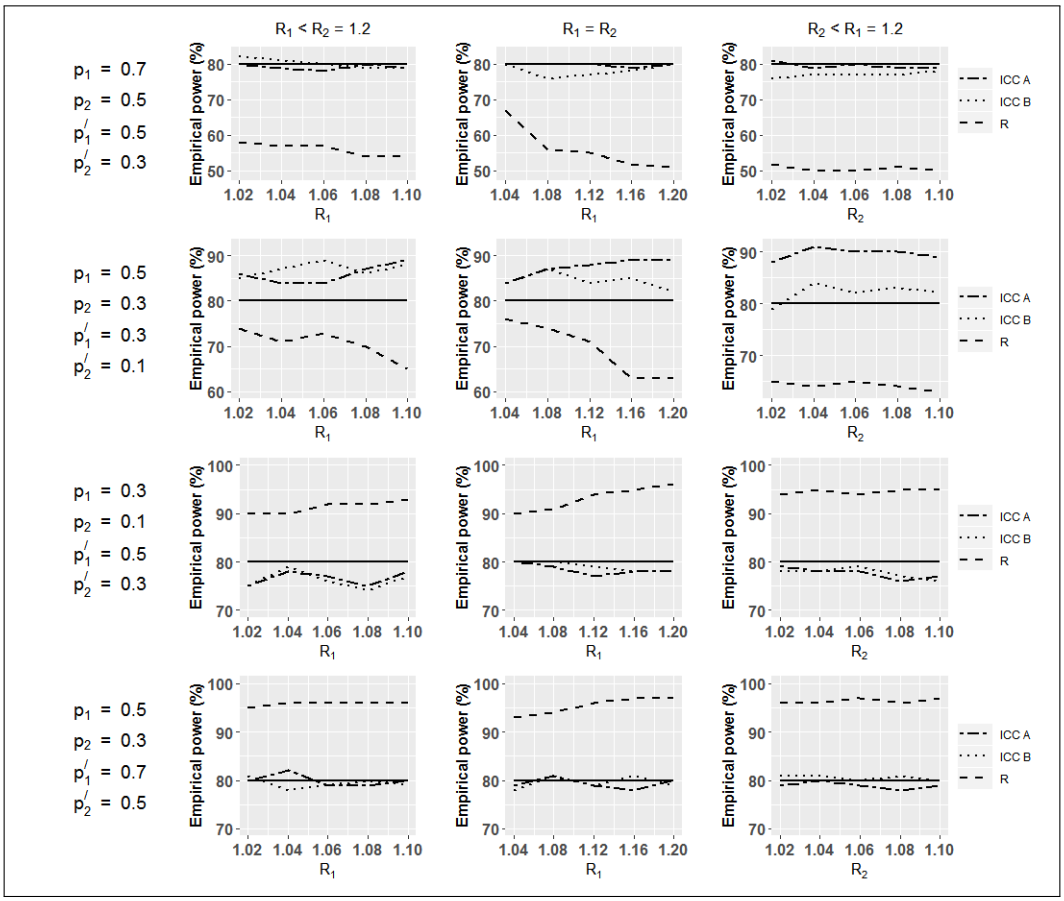

Figure 3. Comparison of three approaches (R-based, ICC A and ICC B) in empirical power estimation. These three approaches were used to compute sample size. 5000 datasets were simulated for each combination of p_1, p_2, p_1', p_2', R_1 and R_2, and the empirical power was computed as the proportion of datasets for which a significant difference is observed. Prevalences between the previous and future study were chosen to be different.

6.1.2 Simulation results

When prevalences in the previously conducted study were high and those in the future one were expected to be moderate (first row of Figure 3), the sample size computed by using the R-based approach did not reach the theoretical power of 80%. The empirical power was always largely lower than 80%. Conversely, when using ICC-based approaches, the empirical power was close to 80%, even closer when we considered two ICCs rather than a common one. When we moved from moderate to low prevalence (second row of Figure 3), the R-based approach still led to fewer clusters than necessary. When using ICC-based approaches, this led to more clusters than necessary but with an empirical power closer to 80% than with the R-based approach. When we moved from low to moderate prevalence (third row of

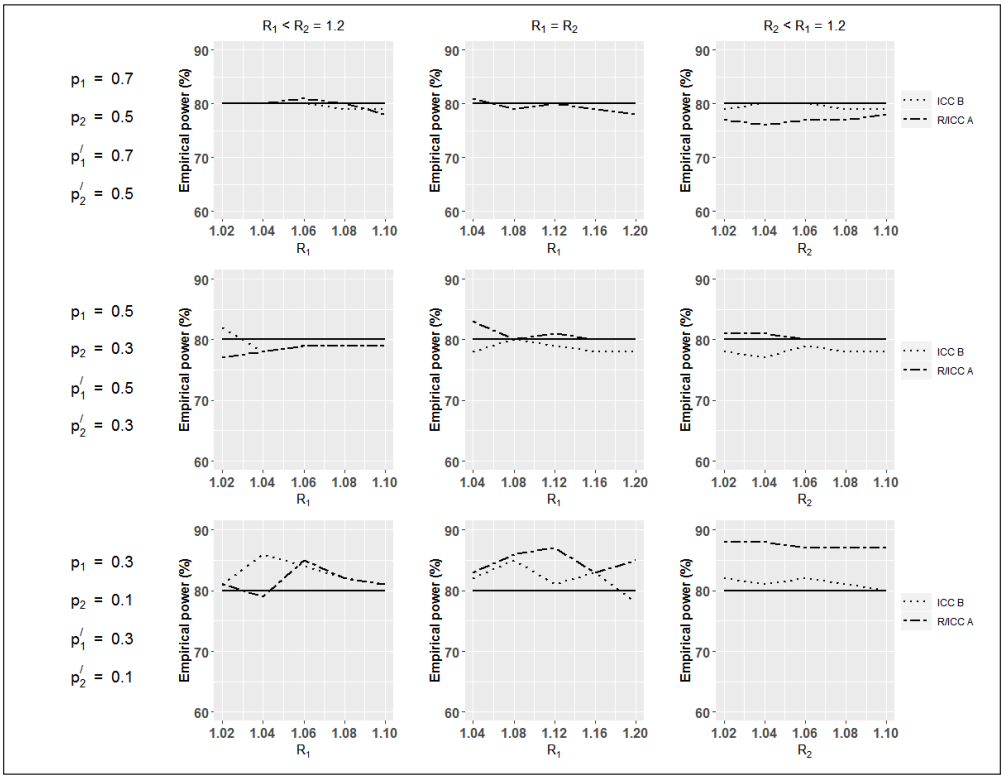

Figure 4. Comparison of three approaches (R-based, ICC A and ICC B) in empirical power estimation. These three approaches were used to compute sample size. 5000 datasets were simulated for each combination of p_1, p_2, p_1', p_2', R_1 and R_2, and the empirical power was computed as the proportion of datasets for which a significant difference has been observed. Prevalences between previous and future study were chosen to be equal.

Figure 3) or from moderate to high prevalence (fourth row of Figure 3), the sample size computed using the R-based approach was greater than necessary: the empirical power was always greater than 80%. Conversely, using ICC-based approaches still led to empirical power close to 80%. Use of the R-based approach was under-powered (when moving from high to moderate prevalence or from moderate to low prevalence) or over-powered (when moving from low to moderate prevalence or from moderate to high prevalence). In all cases, using the R-based approach was worse than using ICC-based approaches. In the no prevalence change setting (Figure 4), the R-based approach and ICC A approach were identical and their empirical power was similar to that of the ICC B approach. Thus, sample size calculation using an ICC approach performs better than using the R approach.

6.2 Asymmetry when using R in sample size calculation

Another drawback of the R approach is its asymmetry in sample size calculation. To illustrate this point, let us consider the situation presented in section 4.1. A previously conducted study in which successes

were considered had a prevalence of 0.15, an ICC of 0.29 and a R coefficient of 2.64. Suppose we want to detect an increase of 10 percentage points in the proportion of success in a future study. To detect an increase in success rate from 0.15 to 0.25, with a power of 80% at the 5% level, 179 clusters for 358 individuals would be needed for each arm. Let us now focus on failures rather than successes. The previously conducted study had a failure rate of 0.85, an ICC of 0.29 and a R coefficient of 1.05. If we now want to detect a decrease from 0.85 to 0.75 in the failure rate (equivalent to the increase in success), with a power of 80% at the 5% level, 149 clusters for 298 individuals would be needed for each arm. Therefore, using the R coefficient would lead to two different required sample sizes, although intuitively, they should be equal.

7 Discussion

In this paper, we have described the R coefficient and explored its association with the outcome prevalence by using mathematical developments for fixed cluster sizes of 2 and using simulations for the most general cases of fixed and variable cluster sizes greater than 2.

We show that the R coefficient decreases with increasing prevalence, so R depends on the outcome prevalence.

Furthermore, R is not symmetrical around a prevalence of 0.5. Thus, R performs even worse than the ICC to measure clustering in the sense that for a given dataset, the R value is not the same when we are interested in success or failure for the same variable. Consequently, R is not an appropriate coefficient if one wants an index independent of the outcome prevalence. Sample size calculation using an approach based on the ICC appears to be preferable because the empirical power is closer to the nominal one versus an approach based on the R coefficient. Moreover, when using R, the sample size differs depending on whether we focus on success or failure. Even if both R and ICCs have limits, our results encourage the use of the ICC over the R coefficient for binary outcomes. Further work is needed to explore or develop other measures, notably the tetrachoric correlation coefficient, which can be used to quantify clustering without being influenced by the outcome prevalence, thus allowing a direct comparison of clustering for outcomes with different prevalences.

Declaration of conflicting interests

The Authors declare that there is no conflict of interest.

Funding

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

Supplemental material

Supplemental material for this article is available online.

References

1. Eldridge S and Kerry S. *A practical guide to cluster randomised trials in health services research*, volume 120. John Wiley & Sons, 2012.

2. Donner A and Klar N. *Design and analysis of cluster randomization trials in health research.* London: Arnold, 2000.
3. Campbell MK, Elbourne DR and Altman DG. Consort statement: extension to cluster randomised trials. *Bmj* 2004; 328(7441): 702–708.
4. 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.
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. Rosner B. Statistical methods in ophthalmology: an adjustment for the intraclass correlation between eyes. *Biometrics* 1982; 38(1): 105–114.
7. Eldridge SM, Ukoumunne OC and Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. *International Statistical Review* 2009; 77(3): 378–394.
8. Ridout MS, Demetrio CG and Firth D. Estimating intraclass correlation for binary data. *Biometrics* 1999; 55(1): 137–148.
9. Kirk DB. On the numerical approximation of the bivariate normal (tetrachoric) correlation coefficient. *Psychometrika* 1973; 38(2): 259–268.
10. Kraemer HC. Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika* 1979; 44(4): 461–472.
11. 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.
12. Lunn AD and Davies SJ. A note on generating correlated binary variables. *Biometrika* 1998; 85(2): 487–490.
13. 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.
14. Donald A and Donner A. Adjustments to the Mantel–Haenszel chi-square statistic and odds ratio variance estimator when the data are clustered. *Statistics in Medicine* 1987; 6(4): 491–499.
15. Giraudeau B. Model mis-specification and overestimation of the intraclass correlation coefficient in cluster randomized trials. *Statistics in medicine* 2006; 25(6): 957–964.

Appendix: Symmetry of the R coefficient

The R estimator is defined as:

$$\widehat{R}_{c,\widehat{p}} = \frac{4kk_2}{(k_1+2k_2)^2},$$

where the $\widehat{p}$ index refers to a focus on success, whose prevalence is $\widehat{p}$.
Considering the symmetrical situation when we are interested in failure, we have:

$$\widehat{R}_{c,1-\widehat{p}} = \frac{4kk_0}{(k_1+2k_0)^2}.$$

A symmetry of R around a 0.5 prevalence value would lead to:

$$\begin{aligned}
\widehat{R}_{c,\widehat{p}} = \widehat{R}_{c,1-\widehat{p}} \quad &\Leftrightarrow \quad \frac{4kk_2}{(k_1+2k_2)^2} = \frac{4kk_0}{(k_1+2k_0)^2} \\
&\Leftrightarrow \quad (k_0-k_2)(k_1^2-4k_0k_2) = 0 \\
&\Leftrightarrow \quad k_0 = k_2 \ \text{ or } \ k_1^2 = 4k_0k_2
\end{aligned}$$

$$k_0 = k_2 \Leftrightarrow \widehat{p} = 1/2$$

$$k_1^2 = 4k_0k_2 \Leftrightarrow \widehat{R}_c = \frac{4k_0k_2 + 4k_1k_2 + 4k_2^2}{(k_1 + 2k_2)^2} = \frac{k_1^2 + 4k_1k_2 + 4k_2^2}{(k_1 + 2k_2)^2} = 1.$$