Supplementary Material
Sample size considerations in the design of cluster randomized trials of combination HIV prevention

Rui Wang, Ravi Goyal, Quanhong Lei, M. Essex, Victor De Gruttola

Although the sample size formula we used can be derived from models in which cluster-level random effects are assumed to be independent across clusters, as are individual outcomes within clusters (e.g., an exchangeable correlation structure), we find that deviations from this assumption do not affect the validity of the sample size formula.

To illustrate the assertions above, we consider the setting where we have c clusters and sample m subjects within each cluster. Let Y_{ik} denote the outcome for the kth individual in the ith cluster, $i = 1, \ldots, c$ and $k = 1, \ldots, m$. Let μ_i denote the cluster means for continuous outcomes or the probability of successes in the ith cluster for binary outcomes, and μ_i, $i = 1, \cdots, c$, are randomly sampled from a probability distribution with mean μ and variance σ^2_{BC}. Here we use μ to denote the overall mean or the overall probability of successes and σ^2_{BC} to denote the variance of μ_i. Within the ith cluster, $Y_{i1}, \cdots, Y_{im} \mid \mu_i$ follows a multivariate distribution with mean μ_i, variance-covariance matrix $V = (\sigma_{k,\ell})$:
For binary outcomes, $\sigma_{k,k} = \mu_i(1-\mu_i)$, for $k = 1, \ldots, m$. For continuous outcomes, with additional normality assumptions about the distribution of μ_i and the multivariate distribution for the outcomes within the same cluster, the above model can be written as the following random effects model with correlated error terms:

$$Y_{ik} = \mu + \alpha_i + \epsilon_{ik}, \tag{1}$$

where $\alpha_i \sim N(0, \sigma^2_{BC})$, $(\epsilon_{i1}, \ldots, \epsilon_{im})^T \sim MVN(\underline{0}, V)$, where $\underline{0}$ is a vector of 0 of length m, and $V = (\sigma_{k,\ell})$. Let $\sigma^2_{WC} = \sigma_{k,k} = Var(Y_{ik} \mid \mu_i)$ denote the within-cluster variance. Let $\sigma^2 = Var(Y_{ik})$ denote the total variance. For continuous outcomes, $\sigma^2 = \sigma^2_{BC} + \sigma^2_{WC}$; for binary outcome, $\sigma^2 = \mu(1-\mu)$.

Let $\rho_C = \frac{\sigma^2_{BC}}{\sigma^2}$ and $\rho_{k,\ell} = \frac{\sigma_{k,\ell}}{\sigma^2}$, the unconditional correlation matrix for the m subjects within the same community is:

$$\left\{ \begin{array}{cccccccc} 1 & \rho_{1,2}+\rho_C & \cdots & \rho_{1,\ell}+\rho_C & \cdots & \cdots & \rho_{1,m}+\rho_C \\ \rho_{2,1}+\rho_C & 1 & \cdots & \rho_{2,\ell}+\rho_C & \cdots & \cdots & \rho_{2,m}+\rho_C \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{m-1,1}+\rho_C & \cdots & \cdots & \cdots & \cdots & 1 & \rho_{m-1,m}+\rho_C \\ \rho_{m,1}+\rho_C & \cdots & \cdots & \cdots & \cdots & \rho_{m,m-1}+\rho_C & 1 \end{array} \right\} \quad (2)$$

If $\sigma_{k,\ell} = 0$ for all $1 \leq k \neq \ell \leq m$, that is, outcomes within the same cluster are independent conditional on μ_i, the unconditional correlation matrix is exchangeable, and the intraclass correlation ρ represents the correlation between any two subjects within the same cluster. In general, the intraclass correlation ρ, defined as $\frac{E\{(Y_{jk}-\mu)(Y_{j\ell}-\mu)\}}{E(Y_{ik}-\mu)^2}$, where the expectation in the numerator is over all distinct pairs of individuals ($k \neq \ell$) taken from the same cluster and over all clusters and the expectation in the denominator is taken over all individuals and all clusters, can be expressed as

$$\frac{\sigma^2_{BC} + \{m(m-1)\}^{-1} \sum_{1\leq k\neq \ell \leq m} \sigma_{k,\ell}}{\sigma^2}. \quad (3)$$

Let $\sigma^2_B = E\{(Y_{jk}-\mu)(Y_{j\ell}-\mu)\} = \sigma^2_{BC} + \{m(m-1)\}^{-1} \sum_{1\leq k\neq \ell \leq m} \sigma_{k,\ell}$, the between-cluster variability corresponding to within-cluster correlation ρ in this setting. If $\sigma_{k,\ell} = 0$ for all $1 \leq k \neq \ell \leq m$, then $\sigma^2_B = \sigma^2_{BC}$; otherwise, they in general differ. The corresponding coefficient of variation

$$k = \frac{\sigma_B}{\mu} = \frac{\sqrt{\sigma^2_{BC} + \{m(m-1)\}^{-1} \sum_{1\leq k\neq \ell \leq m} \sigma_{k,\ell}}}{\mu}. \quad (4)$$

It can be shown that the design effect ($DEFF$) in this case becomes:

$$DEFF = 1 + (m-1)\rho_C + \frac{\sum_{1\leq k\neq \ell \leq m} \rho_{k,\ell}}{m} = 1 + (m-1)\rho. \quad (5)$$

To estimate between-cluster variability σ^2_B, coefficient of variance k, intraclass correlation ρ, and the design effect $DEFF$, it is sufficient to use summary measures from each cluster.

This is not surprising since the observed variance of cluster summary measures provides an unbiased estimate of the true between-cluster variability, irrespective of the actual within-cluster correlation structure. For continuous outcome Y_{ik}, let $\overline{Y}_{i.}$ denote the individual cluster means, $\overline{Y}_{..}$ denote the overall mean, and s^2 denote the empirical variance of cluster means $\overline{Y}_{i.}$, we can estimate σ_B^2, k, ρ, and $DEFF$ as follows:

$$\widehat{\sigma}_B^2 = s^2 - \frac{\widehat{\sigma}_{WC}^2}{m}, \text{ where } \widehat{\sigma}_{WC}^2 = \sum_{i,k} \frac{(Y_{ik} - \overline{Y}_{i.})^2}{c(m-1)} \tag{6}$$

$$\hat{\rho} = \frac{\widehat{\sigma}_B^2}{\hat{\sigma}^2}, \text{ where } \hat{\sigma}^2 = \sum_{i,k} \frac{(Y_{ik} - \overline{Y}_{..})^2}{mc - 1} \tag{7}$$

$$\hat{k} = \frac{\widehat{\sigma}_B}{\overline{Y}_{..}}, \tag{8}$$

and

$$\widehat{DEFF} = 1 + (m-1)\hat{\rho}. \tag{9}$$

Similar formulas exist for binary outcome Y_{ik}, where

$$\widehat{\sigma}_B^2 = s^2 - \frac{\overline{Y}_{..}(1 - \overline{Y}_{..})}{m}, \tag{10}$$

$$\hat{\rho} = \frac{\widehat{\sigma}_B^2}{\hat{\sigma}^2}, \text{ where } \hat{\sigma}^2 = \overline{Y}_{..}(1 - \overline{Y}_{..}) \tag{11}$$

and $\hat{k}$ and $\widehat{DEFF}$ remain the same as in (8) and (9).

For a simulation study showing the validity of using cluster-level summary measures to estimate k, we consider 30 communities and with 20 individuals from each enrolled in the study. The correlation matrix is given in equation (2). It is 'arbitrary' in the sense that conditional on random effects α_i, the variance-covariance matrix V for the error terms ϵ_{ij} is a random correlation matrix generated using methods described in [1] and implemented in an R package *clusterGeneration*. V is fixed for repeated simulations. For each experiment, we generate outcome data Y_{ik} for 600 subjects based on model (1). We let $\sigma_{WC} = 1$ and

Table 1: Actual and estimated intraclass correlation ρ, coefficient of variation k, and design effect $DEFF$ corresponding to an arbitrary random within-cluster correlation matrix. $\widehat{E}$ denotes average of estimates from 1000 simulated studies.

σ_{BC}	ρ	$\widehat{E}(\hat{\rho})$	k	$\widehat{E}(\hat{k})$	$DEFF$	$\widehat{E}(\widehat{DEFF})$
0.0	0.030	0.029	0.172	0.158	1.563	1.553
0.1	0.039	0.039	0.199	0.189	1.745	1.743
0.2	0.067	0.067	0.264	0.259	2.272	2.276
0.3	0.110	0.110	0.346	0.344	3.085	3.083
0.4	0.163	0.162	0.435	0.434	4.106	4.084
0.5	0.224	0.224	0.529	0.529	5.250	5.248
0.6	0.286	0.284	0.624	0.626	6.443	6.397
0.7	0.349	0.347	0.721	0.729	7.626	7.602
0.8	0.408	0.405	0.818	0.827	8.758	8.702
0.9	0.464	0.462	0.916	0.940	9.814	9.766
1.0	0.515	0.514	1.015	1.045	10.781	10.758

$\mu = 1$. Table 1 presents the actual and estimated k, ρ, and $DEFF$. Columns with heading ρ, k, and $DEFF$ represent the true values and are calculated using formulas (3), (4), and (5) respectively. For each value of σ_{BC}, and for each simulated experiment, columns $\hat{\rho}$, $\hat{k}$, and $\widehat{DEFF}$ are calculated based on formulas (7), (8), and (9) respectively; and $\hat{E}(\hat{\rho})$, $\hat{E}(\hat{k})$, and $\hat{E}(\widehat{DEFF})$ are calculated as the sample average from 1000 simulated experiments. As σ_{BC} increases from 0 to 1, the design effect increases from 1.6 and 10.8. The design effect estimated based on $\hat{\rho}$ is unbiased - the sample averages of the estimated values are within a relative 2% of the true values.

References

1. Joe H. Generating Random Correlation Matrices Based on Partial Correlations. *Journal of Multivariate Analysis* 2006; 97, 21772189.

Sample size considerations in the design of cluster randomized trials of combination HIV prevention

Manuscript ID:	CT-13-0219.R2
Submitting Author:	Wang, Rui Save Wang, Rui (proxy) *primary affiliation* 221 Longwood Ave. Room 255 Boston Massachusetts 02115 United States
Authors & Institutions:	Wang, Rui proxy — Brigham and Women's Hospital - Sleep Medicine, Boston, Massachusetts, United States Goyal, Ravi proxy — Harvard School of Public Health - Biostatistics, Boston, Massachusetts, United States Lei, Quanhong proxy — Harvard School of Public Health - Biostatistics, Boston, Massachusetts, United States Essex, M. proxy — Harvard School of Public Health - Immunology and Infectious Diseases, Boston, Massachusetts, United States DeGruttola, Victor proxy — Harvard School of Public Health - Biostatistics, Boston, Massachusetts, United States
Contact Author (populates the ##PROLE_AUTHOR_..## e-mail tags):	Wang, Rui Save Current Contact Author: Wang, Rui (proxy)
Running Head:	Design of cluster randomized trials
Keywords:	cluster randomized trials, network models, design effect, HIV prevention