Derivation of sample size formula for cluster randomized trials with binary responses using a general continuity correction factor and identification of optimal settings for small event rates

Majnu John[a,1], Madhu Mazumdar[a]

[a] *Division of Biostatistics and Epidemiology,*
Department of Public Health
Weill Cornell Medical College, 411 E. 69th St., New York, NY 10021.

Abstract

Trials for comparing interventions where cluster of subjects, rather than individuals, are randomized, are commonly called cluster randomized trials (CRTs). For comparison of binary outcomes in a CRT, although there are few published formulations for sample size computation, the most commonly used is the one developed by Donner, Birkett, and Buck (Am J Epidemiol 1981) probably due to its incorporation in the text book by Fleiss, Levin, and Paik (Wiley 2003). In this paper, we derive a new χ^2 approximation formula with a general continuity correction factor (c) and show that specially for the scenarios of small event rates (< 0.01), the new formulation recommends lower number of clusters than the Donner et al. formulation thereby providing better efficiency. All known formulations can be shown to be special cases at specific value of the general correction factor (e.g., Donner formulation is equivalent to the new formulation for $c = 1$). Statistical simulation is presented with data on comparative efficacy of the available methods identifying correction factors that are optimal for rare event rates. Table of sample size recommendation for variety of rare event rates along with code in"R" language for easy computation of sample size in other settings is also provided. Sample size calculations for a published CRT ("Pathways to Health study"

[1] Corresponding author: 411 E 69^{th} St., New York, NY 10021. e-mail: maj2023@med.cornell.edu, Phone: +01 212 746 4870, Fax: +01 212 746 8544

that evaluates the value of intervention for smoking cessation) are computed for various correction factors to illustrate that with an optimal choice of the correction factor, the study could have maintained the same power with a 20% less sample size.

Key words:
cluster randomized trials, binary responses, approximate sample size formula, continuity correction

1. Introduction

Cluster randomized trials (CRTs) are necessary in the evaluation of health care interventions because of practical and ethical reasons. The units for randomization ('clusters') for the evaluation of intervention in CRT are typically communities, clinics, hospital wards, or medical practices. This design has been used by investigators in the field of drug treated device for assessment of disease control [1] [2], in the evaluation of interventions intended to improve the delivery of health services and quality of care [3], in appraisal of health education activities [4] and prevention models for sexually transmitted diseases [5], to name a few. Donner and Klar (1994) provide an excellent review of the reasons for conducting CRTs, related design issues, and statistical methods for data analysis [6]. CRTs are usually less efficient than individually randomized trials (IRTs) because the responses of individuals within a cluster tend to be more similar than responses of individuals in different clusters [7]. This phenomenon is quantified by a statistics called intraclass correlation coefficient (ICC). Proper design with adequate sample size calculation is of utmost importance in trials utilizing CRT as they are logistically very demanding. For comparison of binary outcomes in a CRT, although there are few published formulations for sample size computation, the most commonly used is the one developed by Donner, Birkett, and Buck [8], probably due to its incorporation in the text book by Fleiss, Levin, and Paik [9]. However, there is room for improvement, especially for trials with binary endpoint with rare occurrence.

We begin with a motivating example from the field of surgery in section 2. Although leaving objects behind in the body cavity after surgery is rare (1 in 5,500 surgeries) [10], it has the potential to be lethal for patients and detrimental to hospitals and insurance companies. A new medical device for detecting these objects by scanning has been marketed but a CRT for evaluation of this device compared to manual sponge count is needed. In our attempt to provide power calculation for this study, we briefly reviewed the existing methods of sample size computation for binary endpoint for IRT (Section 3.1) and their extensions to CRT (Section 3.2). We were unsure about the applicability of the existing formulas for such low proportion of events. Therefore we derived a new sample size formula using a general correction factor (Section 3.3). All known formulations are shown to be special cases at specific value of the general correction factor. In section 4, we present a simulation study for comparison of all methods identifying correction factors that are optimal for rare event rates.

Sample size calculations for a published CRT ("Pathways to Health study" that evaluates the value of intervention for smoking cessation [11]) are computed for various correction factors to illustrate that with an optimal choice of the correction factor, the study could have maintained the same power with a 20% less sample size (Section 5). This points out the usefulness of our new formulation which allows choosing a different correction factor depending on the setting.

We present table of sample size recommendation for variety of rare event rates and also provide related code in "R" language for easy computation of sample size in other settings (Section 6). We return to our motivating example in this section and through computation of sample size for this trial, provide a guideline on how to approach designing of a CRT, specially in a setting where ICCs are not available. In section 7, we offer some discussion on issues related to varying cluster size, generation of correlated binary data, and the need for exact formulation of sample size.

2. Motivating Example: Planning of CRT for Evaluation of a Newly Developed Medical Device for Discovering Retained Surgical Sponges

Inadvertently leaving sponges inside patients who undergo surgery continues to occur despite manual counting of sponges by operating room personnel. Retained sponges may cause no adverse effects in patients and may remain undiscovered for decades. Alternatively, retained sponges may lead to serious sequelae, including sepsis, intestinal obstruction, fistulization, and death. Cima et al (2008) reviewed the incidence and characteristics of surgical retained foreign objects (RFOs) at a tertiary care institution during

4 years and found the incidence rate to be quite small (approximately 1 in $5,500$ operations $= 0.0002$) [10].

A new device recently approved by FDA (RF Surgical detection system) is currently poised for marketing and evaluation

(http://www.rfsurg.com/productoverview.htm). This system consists of two features: 1) sponges, gauzes, and towels with a small unobtrusive embedded chip (measuring at 3.5 mm by 11 mm) and 2) a Blair-Port wand with a 9 foot connection cord that has the capability of scanning a patient weighing up to 500 pounds. A CRT, in support of a comparative trial (Arm 1: Use of the new device for sponge detection versus Arm 2: Use of standard practice of manual sponge counting) with the binary endpoint of detection of a foreign object in body cavity after surgery, is being planned. Accuracy of the existing sample size formulations for extremely low incidence like the one in this study has been questioned in the field of IRT and adjustment to the sample size formulation has been suggested [21]. Therefore we reviewed the sample size formulations in CRT and concluded that some adjustment to the sample size formulation is needed for the special case of rare events in the context of CRT as well.

3. Methods

3.1. Review of Sample Size Formulations in IRT

Before discussing further the sample size formulas for CRTs, we offer a brief review of the most commonly used formulation of sample sizes for the IRTs. Sahai and Khurshid (1996) provides an excellent detailed review [8]. Consider the setting of a two armed clinical trial with dichotomous outcome (ie, event versus non-event). The null hypothesis under test is $H_0 : \pi_1 = \pi_2$ versus the alternative hypothesis of $H_1 : \pi_1 > \pi_2$ where π_i is the proportion of events in the *ith* population. The overall context is to estimate the sample size so that if in fact there is no difference between the two underlying proportions, then the chance is approximately α of falsely declaring the two proportions to differ, and if in fact the proportions are unequal, then the chance is approximately $1 - \beta$ of correctly declaring the two populations to differ, for $\alpha > 0, \beta < 1$. Throughout this manuscript, we assume equal sample sizes for the two groups.

“Exact” formulation for sample size, considered ‘gold standard’ in this setting is derived [11] but the approximate methods are most commonly utilized due to their simplicity in computation [9-10,12-14]. One of the commonly employed method is the "arcsine formula" [9],

$$n = \frac{z_{1-\alpha} + z_\beta}{2[(arcsine\sqrt{\pi_1} - arcsine\sqrt{\pi_2})^2]}, \tag{1}$$

where $\Phi(z_\gamma) = \gamma$ and Φ is the cumulative normal distribution function.

Another commonly used approximation is the "uncorrected χ^2 formula" (UC),

$$n = \frac{[z_{1-\alpha}\sqrt{(2\pi\mu)} + z_\beta\sqrt{\pi_1\mu_1 + \pi_2\mu_2}]^2}{(\pi_1 - \pi_2)^2}, \tag{2}$$

where $\mu_i = 1 - \pi_i, i = 1, 2$; $\pi = \frac{\pi_1+\pi_2}{2}$, $\mu = \frac{\mu_1+\mu_2}{2}$ [10].

It has been shown that the above two formulas, (1) and (2), give similar output but are considered to be serious underestimation of the sample size recommended by the "exact" method [11].

To rectify this, Kramer and Greenhouse (KG) developed a "corrected χ^2 method" given by

$$n = \frac{A\left[1 + \sqrt{1 + 8(\pi_1 - \pi_2)/A}\right]^2}{[4(\pi_1 - \pi_2)^2]}, \tag{3}$$

where $A = [z_{1-\alpha}\sqrt{(2\pi\mu)} + z_\beta\sqrt{\pi_1\mu_1 + \pi_2\mu_2}]^2$ [12]. Later, Casagrande, Pike and Smith (CPS) developed a χ^2 approximation sample size formula with a general correction factor, c', given by

$$n = \frac{A\left[1 + \sqrt{1 + 4(1 - 2c')(\pi_1 - \pi_2)/A}\right]^2}{[4(\pi_1 - \pi_2)^2]}[13]. \tag{4}$$

The KG and UC formulation were shown to be special cases of the CP formulation by setting $c' = -0.5$ and $c' = 0.5$ respectively. Casagrade and Pike also demonstrated that the sample size recommended by the UC and KG formulation are an underestimate and an overestimate of the sample size obtained through the exact formulation [13]. They proposed a formula based on $c' = 0$ and established that the sample size obtained via this new formula was satisfactorily close to that obtained by the exact method.

3.2. Review of Sample Size Formulations in CRT

In the two-armed clinical trial with CRT design, the most widely used sample size formula is the one given in the text book by Fleiss, Levin, and Paik (FLP) [14] which was originally developed by Donner, Birkett, and Buck (DBB) [15]. Assuming there are K clusters in each group, with clusters of equal size $\bar{n}$, the formulation is given by

$$K = \frac{[z_{\alpha/2}\sqrt{2\pi\mu f} + z_{\beta}\sqrt{f_1\pi_1\mu_1 + f_2\pi_2\mu_2}]^2}{\bar{n}(\pi_1 - \pi_2)^2}, \tag{5}$$

where f_1 and f_2 denote the variance inflation factors (VIFs) of the two sets of K clusters; $f = (f_1 + f_2)/2 = 1 + \rho(\bar{n} - 1)$ is the VIF under the null hypothesis of equality of proportions; and ρ being the intra-cluster co-efficient assumed to be the same within each groups [14-15].

Since in the IRTs, $f_1 = f_2 = 1$, it is clear that (5) is an extension of the uncorrected χ^2 formula to the CRT setting. To be exact, the FLP formula is actually an approximation of the DBB formula, by the fact that $\sqrt{f_1\pi_1\mu_1 + f_2\pi_2\mu_2} \approx \sqrt{2\pi\mu f}$.

Several other formulations of sample size computation are available for related but somewhat different settings. For example, Liu and Liang [16] proposed sample size formulation in the context of generalized linear models which used unified tools for correlated continuous and discrete responses. For the special case of the 'two-sample problem with binary responses', their general formula reduces to the DBB formula under the assumption of equal sample size in the two groups and an exchangeable correlation structure for the working correlation matrix. Fleiss *et al.* [14] extends FLP/DBB formula to the case where the 'exposure' varies across clusters. Hayes [18] present a formula which takes into account the between cluster variability, but this formula doesn't take into account the intra-cluster correlation. Manatunga *et al.* [19] and van Breukelen *et al.* [20] extends these sample size estimation methods to account for variability in cluster size.

3.3. Derivation of a new sample size formula for CRTs

In this section, we derive our χ^2 approximation formula for sample sizes with a general continuity correction factor. Let X_{ij} and Y_{ij} denote the outcomes of the i^{th} individual in the j^{th} cluster in the intervention and control groups respectively; $i = 1, \ldots n; j = 1, \ldots K$. We assume that $X_{ij} \sim$ Bernoulli(π_1) and $Y_{ij} \sim$ Bernoulli(π_2). Let us denote the intra-cluster correlations as ρ_1 and ρ_2 and the corresponding variance inflation factors (VIFs) in each group as f_1 and f_2.

$$\bar{X} = \frac{1}{nK} \sum_{i=1}^{n} \sum_{j=1}^{K} X_{ij}; \qquad\qquad \bar{Y} = \frac{1}{nK} \sum_{i=1}^{n} \sum_{j=1}^{K} Y_{ij}.$$

Define

$$d = \bar{X} - \bar{Y} \quad \text{and} \quad Z = \frac{d - \frac{1}{nK}}{\sqrt{\frac{2}{nK} \left(\frac{\bar{X}+\bar{Y}}{2} \right) \left[1 - \left(\frac{\bar{X}+\bar{Y}}{2} \right) \right] f}}.$$

Z is a test statistic ($-\chi^2$ test with Yates' correction–) commonly used to test the null hypothesis $H_0 : \pi_1 = \pi_2$. Let d^* be such that

$$z_{1-\alpha} = \frac{d^* - \frac{1}{nK}}{\sqrt{\frac{2}{nK} \left(\frac{\bar{X}+\bar{Y}}{2} \right) \left[1 - \left(\frac{\bar{X}+\bar{Y}}{2} \right) \right] f}}$$

under the null hypothesis. Here $f = \frac{f_1+f_2}{2}$ is the variance inflation factor calculated under the null hypothesis. Replacing $\left(\frac{\bar{X}+\bar{Y}}{2} \right)$ by its expectation, we get

$$z_{1-\alpha} \approx \frac{d^* - \frac{1}{nK}}{\sqrt{\frac{2}{nK} \bar{\pi} \bar{\mu} f}}.$$

Rearranging,

$$d^* \approx z_{1-\alpha} \sqrt{\frac{2\bar{\pi}\bar{\mu}f}{nK}} + \frac{1}{nK}. \tag{6}$$

Also

$$\Pr(d \geq d^*) \approx 1 - \Phi \left(\frac{d^* - (\pi_1 - \pi_2) - \frac{c}{nK}}{\sqrt{\frac{\pi_1 \mu_1 f_1 + \pi_2 \mu_2 f_2}{nK}}} \right)$$

where Φ is the cumulative normal distribution function and c is a correction factor to allow for the discreteness in the distribution of d. If this d^* is to lead to a test with power $1 - \beta$ then it should satisfy

$$-z_\beta \approx \frac{d^* - (\pi_1 - \pi_2) - \frac{c}{nK}}{\sqrt{\frac{\pi_1 \mu_1 f_1 + \pi_2 \mu_2 f_2}{nK}}}.$$

That is,

$$d^* \approx (\pi_1 - \pi_2) + \frac{c}{nK} - z_\beta \sqrt{\frac{\pi_1 \mu_1 f_1 + \pi_2 \mu_2 f_2}{nK}}. \tag{7}$$

Equating (6) and (7) and solving for K, we get,

$$K = \frac{A\left[1 + \sqrt{1 + \frac{4(\pi_1 - \pi_2)(1-c)}{A}}\right]^2}{4n(\pi_1 - \pi_2)^2}, \tag{8}$$

where

$$A = \left[z_{1-\alpha}\sqrt{2\bar{\pi}\bar{\mu}f} + z_\beta \sqrt{\pi_1 \mu_1 f_1 + \pi_2 \mu_2 f_2}\right]^2.$$

Note that when $c = 1$, the formula (8) is the same as the FLP formulation reproduced in equation (5) of this document. $c = 0$ and $c = -1$ are extensions of CGS and KG formulas respectively to the CRT setting. Also note that the sample size formula for the alternate hypothesis, $H_1 : \pi_1 < \pi_2$, may be obtained by switching π_1 and π_2 in formula (8)

4. Statistical Simulations for comparison of χ^2 approximations with different correction factors

We compare the performance of the sample size formula given by (8) with different choices of the continuity correction factor, c, via statistical simulations. The primary endpoint for comparison is the number of clusters recommended by different formulas maintaining the power closest to 80% to the one corresponding to exactly 80%.

4.1. Simulation Parameters

We compare the choices of $c = 1, 0, -1$ for the following settings of the parameters:

- (π_1, π_2) pairs of $(0.0002, 0.0001)$, $(0.01, 0.0005)$, $(0.1, 0.05)$, $(0.4, 0.25)$, $(0.60, 0.40)$;
- $\rho_1 = \rho_2 = 0.25$;
- $\bar{n} = 30$;
- $\alpha = 0.05$;

- $1 - \beta = 0.8$.

Because of symmetry, the sample sizes required for (π_1, π_2) is the same as that for $(1 - \pi_2, 1 - \pi_1)$. Therefore, we do not simulate cases where $min(\pi_1, \pi_2) \geq 0.5$.

4.2. Data Generation and parameter estimation Methods

We generated 500 correlated binary data within each cluster for the above parameters, using a Monte Carlo simulation method proposed by Lunn and Davies [17]. For each j, $X_{ij}'s$ were generated using the formula (given in [17]),

$$X_{ij} = (1 - U_{ij})V_{ij} + U_{ij}Z_j,$$

where V_{ij}, Z_j are independent with Bernoulli(π_1) distribution, and U_{ij} are independent with Bernoulli($\sqrt{\rho_1}$) distribution. Y_{ij}'s were obtained similarly with π_1 and ρ_1 replaced by π_2 and ρ_2. $\hat{\rho}$'s were estimated using the FLP formula. Empirical power was calculated by the proportion of times the χ^2 test statistic with Yates' correction exceeds the critical value.

4.3. Format of Tables Presenting Simulation Results

Simulation results are presented in Tables 1 with 8 columns. For each (π_1, π_2)-pair listed in column 1 and for each choice of $c(= 1, 0, -1)$ listed in column 2, the required number of clusters via the formula (8) is computed at 80% power and 0.05 significance level with specification of $\rho_1{=}\rho_2 = 0.25$. These number of clusters are shown at the top of each cell in column 3 (-the numbers without parenthesis-). Using the generated data, the mean estimated ρ_1, ρ_2, π_1, π_2 along with their standard deviations are presented in columns 4-7. Using the estimated ρ's, the number of clusters are estimated again and are presented in column 3 (-the numbers with parenthesis-). Empirical power is listed column 8.

Note the row labeled "*" corresponding to each (π_1, π_2)-pair. The top number in column n 3 corresponding to this row is obtained by trial and error and is the number of clusters required to achieve as close to the power of 80% as possible. The corresponding estimated ρ's and π's are presented in the same row for the next four columns. Using these estimates, the choice of c that provides the number of clusters closest to that corresponding to "*" is listed Below "*". This is the most "optimal" choice of c for each setting of (π_1, π_2)-pair.

4.4. Simulation Results and Explanation

Under the null hypothesis of equality of the rate of events, the nominal significance values were found, on average, to be around 0.05 (data not shown). The numbers in Table 1 may be better explained using a particular case of (π_1, π_2)-pair, say (0.01, 0.005), as an example (see the bolded numbers in table 1).

The required number of clusters obtained via (8) for $c = 1$, 0 and -1, with simulation parameters specified above are $1013, 1026$ and 1036 respectively. Respectively for $c = 1$, 0 and -1, the mean estimated ρ_1 are $0.244, 0.243$, and 0.244 and the mean estimated ρ_2's are $0.244, 0.243$ and 0.243. The corresponding empirical power is 83.8%, 85.6% and 85.6%, respectively. Since the ρ's are somewhat underestimated which could be an artifact of the data generation process, for true comparison of the number of clusters needed by different values of c, we need to compute these numbers based on the mean estimated ρ's as opposed to the true value of $\rho = 0.25$. Using the estimates, the corresponding number of clusters are 977, 990, and 1003 respectively for $c = 1$, 0, and -1. The last row ("*") gives the minimum number of clusters required to achieve 80% power which in this particular case is estimated to be 942.

The mean estimated ρ's in this case are comparable to those for $c = 1$, 0 and -1, and hence the performance of the formula (8) with correction factors $c = 1$, 0 and -1 may compared by assessing the deviances $35(= 977 - 942)$, $48(= 990 - 942)$, and $61(= 1003 - 942)$. The best choice of the correction factor is $c = 1$ corresponding to the least deviance compared to $c = 0$ and $c = -1$. Lastly, the number below "*" provides the choice of c that corresponds to the required number of clusters closest to that given by "*". In this particular case, a choice of $c = 3$ gives the number of clusters, 948, closest to that given by "*", 942. Therefore $c = 3$ is considered the most optimal choice of the correction factor for detecting a halving in proportion when $\pi_1 = 0.01$ and $\pi_2 = 0.005$.

4.5. Conclusions from Simulation Results

Similarly assessing the results from all 10 scenarios presented in Table 1, the following conclusions can be drawn:

- For (π_1, π_2) pairs with $max(\pi_1, \pi_2) < 0.01$, all three choices of $c = 0, -1, 1$ recommend higher sample size than optimal. Recommended value for number of clusters using $c = 3$ is more optimal.

- For all (π_1, π_2) pairs with $0.01 < max(\pi_1, \pi_2) \leq 0.4$, $c = 1$ is a better choice than $c = 0, -1$.
- For all (π_1, π_2) pairs with $0.4 < max(\pi_1, \pi_2) \leq 0.6$, all three correction factors specified (namely, $c = 1$, 0 and -1), provides underestimate for the required number of cluster sizes. More extreme values of c is needed to produce optimal sample size but the value varies for particular pairs.

An important cautionary note is that above conclusions are based on power simulations which used a large $\rho(= 0.25$. The optimal choices for the ranges of (π_1, π_2) may be different as seen in the illustrative example below.

Table 1: See Attached document

5. CRT for evaluation of nicotine gum and motivational interviewing for smoking cessation: An Illustrative Example

Although there has been significant decline in smoking prevalence among adults in the United States in the past few decades, it has not been the case in all the subpopulation of smokers. This is particularly true in smokers below the poverty level. An intervention study for smoking cessation in this subpopulation is of significance, especially since studies have shown high prevalence and motivation to quit among residents of low-income housing. Okuyemi *et al.*(2007) reported the results from a CRT that tested nicotine gum plus motivational interviewing (MI) for smoking cessation in 20 low-income housing developments (HDs), in which intervention participants (10 HDs) received educational materials addressing fruit and vegetable consumption, 8 weeks of 4 mg nicotine gum, and 5 MI sessions on quitting smoking, and comparison participants (10 HDs) received 5 MI sessions and educational materials only [11]. The sample size (-no. of clusters-) calculation was based on the assumptions that there would be 20 participants in each of the 20 HDs, a moderate intra-cluster correlation of 0.02, a 6-month quit rate of 6% (π_1) in the comparison arm, and a 18% (π_2) quit rate in the cessation arm. Power analysis based on the DBB formula showed that there would be 89% power to detect a significant difference between the two arms.

We suspect that the DBB formula overestimated the required number of clusters, and that formula (8) with a continuity correction factor of $c = 3$ will provide a better estimate of the required number of clusters. Indeed, a power simulation described in Table 2 provides support to our thinking. The number of clusters required in each arm to detect a difference of 12% at 89% power (with all other assumptions kept same) is respectively, 11, 9 and 8 cluster per arm, for $c = 1, 3, 4$ in (8). We generated 100,000 Monte Carlo two-samples of clusters, with 20 per cluster, for 4 different scenarios: 11, 10, 9, and 8 clusters in each arm (-note that, 10 clusters in each arm corresponds to the DBB formula that was used for the study design in [22]). Since the estimates of ρ_1 and ρ_2 are lower than the actual ρ_1 and ρ_2 used in Lunn and Davis [17] method, we used $\rho_1 = \rho_2 = 0.04$ (which is more conservative than 0.02 used in [17]) for our power simulations.

From Table 2, we observe that even with 9 clusters per arm (corresponding to $c = 3$), the empirical power is greater than 89% and with 8 clusters per arm (corresponding to $c = 4$), the empirical power is 88.1%. But the estimated ρ_2 is substantially higher than 0.02. So, with $\rho_1 = \rho_2 = 0.02$, there is reason to believe that even 8 clusters per arm would have been sufficient for 89% power. In other words, this example illustrates the point that formula (8) with $c = 3$ provides a more accurate estimate of the required number of clusters for this choice of (π_1, π_2). The reduction (that is, the improvement) in number of clusters is 10% $(20 - 18/20 = 0.1)$ which amounts to a substantial gain in terms of the cost of conducting the study. Incidentally, in this example, the sample size based on $c = 4$ provides the best estimate, with a 20% improvement.

Table 2: Sample size comparison for the illustrative example: Pathway to Health Study
Note: nominal power = 89%, $\alpha = 0.05$, Number of Monte Carlo samples = 100,000

Comparison of various sample size estimates for the Pathways to the Health Study						
Method used for sample size calculation	No. of clusters in each arm	Estimated Power	Mean $\hat{\rho_1}$ s.e	Mean $\hat{\rho_2}$ s.e	Mean $\hat{\pi_1}$ s.e	Mean $\hat{\pi_2}$ s.e
Formula (8) with $c = 1$; (FLP)	11	94.9%	0.024 0.0002	0.030 0.0001	0.060 <0.0001	0.180 0.0001
DBB Formula	10	92.9%	0.023 0.0002	0.029 0.0001	0.060 <0.0001	0.180 0.0001
Formula (8) with $c = 3$	9	90.8%	0.021 0.0002	0.028 0.0001	0.060 <0.0001	0.180 0.0001
Formula (8) with $c = 4$	8	88.1%	0.019 0.0002	0.027 0.0001	0.060 <0.0001	0.180 0.0001

6. Sample Size Table, Codes, and Guidelines for Designing CRT

We present Table 3 with some sample size tabulations for assistance with future trial design and show empirically the settings where $c = 3$ recommend lower sample size. This table provide the required number of clusters obtained via (8) using $c = 1$, 0, and -1 for various small values π's (< 0.01), $\rho = 0.01$, and average cluster size of 30, 5% level of significance, and 80% power. We also provide the number of clusters required based on $c = 3$. 'R' code for the sample size computation is provided below for easy implementation in any other setting.

6.1. Code for Sample Size Computation

```
ss.crt ← function(power, alpha, pi1, pi2, n, rho1, rho2, c){
## c is the continuity correction factor,
## n is the cluster size
## the function ss.crt returns the required number of clusters

VIF ← function(n, rho){
rslt ← 1 + ((n -1)*rho)
return(rslt) }


f.beta ← function(z, level){
rslt ← pnorm(z) - level}


f.z ← function(level){
rslt ← uniroot(f.beta, c(-10, 10), level = level)$root
return(rslt)}

A ← function(power, alpha, pi1, pi2, vif1, vif2){

pi ← (pi1 + pi2)/2
mu ← 1 - pi

mu1 ← 1 - pi1
mu2 ← 1 - pi2
vif ← (vif1 + vif2)/2

# beta ← 1 - power
zbeta ← f.z(power)
zalpha ← f.z(1-alpha)

rslt ← (zbeta*sqrt(pi1*mu1*vif1 + pi2*mu2*vif2)) + (zalpha*sqrt(2*pi*mu*vif))
rslt ← rslt^2
return(rslt)}

ss ← function(a, pi1, pi2, c){

term1 ← (4*(pi1 - pi2)*(1 - c))
term2 ← 1 + sqrt(1 + (term1/a))

numer ← (a*(term2)^2)
denom ← (4*(pi1-pi2)^2)
rslt ← numer/denom

return(rslt)  }

vif1 ← VIF(n, rho1)
vif2 ← VIF(n, rho2)
a ← A(power, alpha, pi1, pi2, vif1, vif2)

rslt ← ss(a, pi1, pi2, c)
rslt ← rslt/n

return(rslt) }
```

6.2. Connecting to the Motivating Example

Getting back to our motivating example, where we needed sample size computation for the (π_1, π_2)-pair of $(0.0002, 0.0001)$ the number of clusters needed are 7975, 8629, 9260, 6574 respectively for the value of the correction factors taking different values, with the last one being most optimal. This will mean that 6574 hospitals with 30 surgery each will be needed to detect a halving of the incidence rate of leaving foreign objects inside the patient during surgery. Since for practical purpose, it will be easier if we can recruit higher number of patients in each hospital (say 100 patients from each hospitals) and reduce our need for convincing a big number of hospitals to join the trial, we compute the number of clusters in that setting and find that xxx hospital will suffice.

7. Discussions

The effect of an intervention (therapeutic device or drug administration, lifestyle change, or health care delivery system change etc.) is often evaluated by a cluster randomized trial. This implies that organizational units such as hospitals, communities, or clinics are randomly allocated to treatment conditions and all persons sampled from such cluster receive the same treatment assigned to the cluster. Instead of randomizing clusters, one may randomize persons within each cluster which is statistically more efficient but not always more convenient. We provide a formulation for assessing number of clusters needed for trials like this and show few existing formulations to be special cases.

Assumption of equal cluster size has been made in our formulation. This is not a practical assumption. It has been shown that sampling 25% more clusters when the sample sizes within cluster are extremely variable compensates for all weaknesses arising from this increase in variability [24]. We expect this results to be applicable for our setting.

There has been various approaches for generating correlated binary data. Park (1996) provides a comprehensive review of related issues [25]. Lunn's procedure, although straight and simple to implement, gave slightly biased estimates of ρ for our simulation. Another algorithm presented in [25], although not as simple as Lunn and Davis's method to implement, but that still requires no complicated procedures might have yielded unbiased estimates of ρ's.

Relying on approximate formulas is not ideal in the setting of CRT unlike the IRT setting where there is a particular choice of the optimal correction factor for all situation. This fact emphasizes the need for deriving a formula based on exact methods for CRTs with dichotomous outcome.

8. Acknowledgements

The authors were supported by a pilot grant from the Weill Cornell Center for Education and Research on Therapeutics (CERT) grant (Agency of Health Research and Quality RFA-HS-05-14). Partial support for both the authors came from Clinical Translational Science Center (CTSC) grant (UL1-RR024996). Partial support for Dr. Mazumdar also came from Collaborative Program in Nutrition and Cancer Prevention (R25CA105012). The authors thank Alison M. Edwards for helping with the literature search on CRTs in medical journals and also for helping with creating the sample size tables.

References

[1] Lindblade KA, Eisele TP, Gimnig JE, Alaii JA, Odhiambo F, ter Kuile FO, *et al.* Sustainability of Reductions in Malaria Transmission and Infant Mortality in Western Kenya With Use of Insecticide-Treated Bed nets: 4 to 6 Years of Follow-up. JAMA 2004; 291: 2571-80.

[2] Kroeger A, Lenhart A, Ochoa M, Villegas E, Levy M, Alexander N, *et al.* Effective control of dengue vectors with curtains and water container covers treated with insecticide in Mexico and Venezuela: cluster randomized trials. BMJ 2006; 332: 1247-52.

[3] Grosskurth H, Mosha F, Todd J, Mwijarubi E, Klokke A, Senkoro K, *et al.* Impact of improved treatment of sexually transmitted diseases on HIV infection in rural Tanzania: randomized controlled trial. Lancet 1995; 346: 530-36.

[4] Gail MH, Byar DP, Pechacek TF, Corle DK. Aspects of statistical design for the community intervention trial for smoking cessation (COMMIT). Control Clin Trials 1992; 13: 6-21.

[5] Fontanet AL, Saba J, Chandelying V, Sakondhavat C, Bhiraleus P, Rugpao S, *et al.* Protection against sexually transmitted diseases by granting sex workers in Thailand the choice of using male or female condom: results from a randomized controlled trial. AIDS 1998; 12: 1851-59.

[6] Donner A, Klar N. Cluster randomization trials in epidemiology: Theory and application. J Statist Plann Inference 1994; 42: 37-56.

[7] Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. London: Arnold Publishers Limited; 2000.

[8] Donner A, Birkett N, Buck C. Randomization by cluster. Sample size requirements and analysis. Am J Epidemiol, 1981; 114: 906-14.

[9] Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions. 3rd ed. New York: John Wiley and Sons, Inc.; 2003.

[10] Cima RR, Kollengode A, Garnatz J, Storsveen A, Weisbrod C, Deschamps C: Incidence and characteristics of potential and actual retained foreign object events in surgical patients, J Am Coll Surg, 2008; 207(1):80-7.

[11] Okuyemi KS, James AS, Mayo MS, Nollen N, Catley D, Choi WS, *et al.* Pathways to Health: A cluster randomized trial of nicotine gum and motivational interviewing for smoking cessation in low-income housing. Health Educ Behav, 2007; 34: 43-54.

[12] Sahai H, Kurshid A. Formulae and table for the determination of sample size and power in clinical trials for testing differences in proportions for the two-sample design: a review. Stat Med 1996; 15: 1-21.

[13] Cochran WG, Cox GM. Experimental Designs. New York: John Wiley and Sons, Inc.; 1957.

[14] Haseman JK. Exact sample sizes for use with the Fisher-Irwin Test for 2 × 2 tables. Biometrics 1978; 34: 106-09.

[15] Kramer M, Greenhouse SW. Determination of sample size and selection of cases. In Cole JO, Gerard RW, editors. Psychopharmacology: Problems in Evaluation, Washington, D.C.: National Academy of Sceinces, National Research Council, 1959; Publication 583: 356-71.

[16] Casagrande JT, Pike MC, Smith PG. An improved approximate formula for calculating sample sizes for comparing two binomial distributions. Biometrics, 1978; 34: 483-86.

[17] Liu G, Liang K.-Y. Sample size calculation for studies with correlated observations. Biometrics, 1997; 53: 937-47.

[18] Lunn AD, Davies SJ. A note on generating correlated binary variables. Biometrika, 1998; 85: 487-90.

[19] Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. Int J Epidemiol, 1999; 28: 319-26.

[20] Manatunga AK, Hudgens MG, Chen S. Sample size estimation in cluster randomized studies with varying cluster size. Biometrical J, 2001; 43: 75-96.

[21] Lemeshow S, Hosmer D, Stewart JP: A comparison of sample size determination methods in the two group trial where the underlying disease is rare. Communication in Statistics - Simulation, Computation, 1981; B10(5), 437-449.

[22] Fleiss JL. Statistical Methods for Rates and Proportions. New York: John Wiley and Sons, Inc.; 1973.

[23] Donner A, Klar N.Pitfalls of and controversies in cluster randomization trials. Am J Public Health 2004;94(3):416-22.

[24] van Breukelen G JP, Candel M JJM, Berger M PF. Relative efficiency of unequal cluster sizes for variance component estimation in cluster randomized and multicenter trials. Stat Methods Med Res, 2008; 17: 439-58.

[25] Park CG, Park T, Shin DW. A simple method for generating correlated binary variates. Am. Statistician, 1996; 50: 306-10.