Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2016, Vol. 25(2) 505–519

DOI: 10.1177/0962280212460443
smm.sagepub.com
SAGE

Design effects for sample size computation in three-level designs

Tina D Cunningham[1] and Robert E Johnson[2]

Abstract

Experiments with multiple nested levels where randomization can take place at any level bring challenges to the computation of sample sizes. Formulas derived under simple single-level experiments must be adjusted using multiplicative factors or design effects. In this work, we take a unified approach to finding the design effects in terms of intracluster correlations and present formulas to compute sample sizes of different levels. Equal cluster sample sizes and homogeneous within cluster variances are assumed.

Keywords

Sample size, design effects, three-level design

1 Introduction

Experiments involving a three-level data structure are increasingly common in many areas of research. For example, evaluation of a new health intervention usually includes patients who are nested within physicians and physicians who, in turn, are nested within health centers. Three-level designs are also often encountered in educational studies where students are nested within classrooms and classrooms are nested within schools.

The analytic complexity of three-level data arises from the need to understand the relationships between and within the nested levels and the appropriate specification of the variance structure. Despite the growing literature on technical and statistical issues related to three-level data, one of the most problematic issues remains to be the estimation of statistical power and sample size.[1,2]

In the context of three-level designs, there are multiple challenges in computing sample sizes and conducting power analyses. First, the multilevel nesting effect introduces more than one intracluster correlation into the model. Second, the variance of the estimated treatment effect is more complicated than that of a single level or two-level design. Third, three sample sizes are required—one for each level—and each sample size affects the power differently.

[1]Graduate Program in Public Health, Eastern Virginia Medical School, Norfolk, VA, USA
[2]Department of Biostatistics, School of Medicine, Vanderbilt University, Nashville, TN, USA

Corresponding author:
Tina D Cunningham, Graduate Program in Public Health, Eastern Virginia Medical School, 651 Colley Avenue, Room 418, Norfolk, VA 23501, USA.
Email: CunninTD@evms.edu

Since units in the lower levels of a nested study design may not be independent of each other, the fundamental sample size assumption for simple randomized control trials does not hold. For example, patients who visit the same physician might share common characteristics which vary from patients who visit a different physician. In addition, physicians within the same health center tend to be homogeneous while physician groups across centers are heterogeneous. The existing literature suggests that one way to account for the effect of this intracluster dependence is to inflate the standard sample size estimates by a clustering effect, also known as the variance inflation factor or design effect.

Methods to estimate the design effect in two-level data are well-established.[3,4] A few authors have suggested formulas for the design effect in three-level data structure.[5–7] However, these formulas were developed under the premise that random assignment occurs at the highest level (health centers or schools, for example), whereas in reality treatments could be randomized at level two (physicians or classes) or level one (patients or students). Our goal is to present formulas for the design effects that can be used in sample size computation for three-level experiments where randomization might occur at different levels. We begin with reviewing the literature on two/three level designs. Here, we provide three practical examples to illustrate how the randomization of treatment might be directed towards different levels. We then review design effects for both two- and three-level studies. Next, we develop formulas for the design effects under different randomization schemes, assuming that the statistical analysis is based on an (approximate) *t*-test by way of an additive Generalized Linear Mixed Model (GLMM). To narrow the focus of the discussion, we limit our attention to continuous outcomes with identity link function and compound symmetric error structure. Finally, we provide a practical example to illustrate the application of our formulas.

2 Background

2.1 Literature review and examples

In the past few decades, randomized trials with multilevel designs have become common in many areas of research such as psychology, education, and medical sciences. This can be seen in numerous publications in which the choice of the unit of analysis is a major concern.[1,2,8,9]

Although the literature on multilevel studies mainly focuses on two-level designs, more recent attention has been devoted to the three-level designs. Pituch et al.[10] viewed the three-level design as a modification of a two-level *cluster randomized trial* where the treatments are assigned to the whole cluster or group, or a *multi-site trial* where the treatments are assigned to the individuals within each site. Treatments may be assigned to whole clusters, but clusters may have subclusters, such as teachers nested within schools, thus resulting in a three-level design. Treatments may be assigned to clusters of students or patients within multiple sites, thus resulting in a multi-site cluster randomized design. More on issues in the design and analysis of three-level data is found in Heo and Leon,[5] Teerenstra et al.,[6,7] and Pituch et al.[10]

Choosing the unit of randomization is a primary concern in the designs of a multi-level study. Treatment assignment can be done at a cluster level or individual level. According to Glynn et al.,[11] the selection of an optimal randomization scheme is a debatable topic. These authors argued that randomization may be at the individual patient level, the patients' provider, a group of providers within a practice, or several practices in a health system or specific geographic area. If the outcomes are measured separately in each individual, then randomization at the lowest level will minimize the

risk of covariate imbalance. On the other hand, if it is not feasible to conduct multiple study arms within a cluster, due in part to contamination across the arms, then the cluster should be chosen over the individual as a better unit of randomization.

Three recent studies serve to motivate this paper. In the first example, Ruf et al.[12] examined the dissemination strategies of an online quality improvement program for alcohol-related disorders in South Baden and South Württemberg, Germany. A hundred and twelve general practices were randomized into three groups, with 2647 practitioners participating in the study. The primary outcome was the acceptance and the use of the program. This cluster randomized trial assumes a three-level structure in which randomization takes place at third level (practice).

In the second example, researchers in Spain conducted a study to assess the effectiveness of Experimental Program for Physical Activity Promotion (PEPAF), a program designed to increase physical activity in patients who did not meet recommended aerobic physical activity levels.[13] The primary outcome measure was the change in physical activity from baseline to 6 months using the 7-Day Physical Activity Recall (PAR) semi-structured interview. Recruitment involved 15 research groups associated with the Health Promotion Primary Care Research Network. Physicians were randomized into two groups. Those in the PEPAF program provided patients with advice on using health promotion websites and health educational materials and those in the control group delivered usual care. This study is an example of a three-level design in which randomization took place at the second level (physician).

In the third example, a randomized, controlled study was conducted by the Virginia Ambulatory Care Outcomes Research Network to evaluate the effectiveness of an interactive web-based personalized healthcare record compared to usual delivery of preventive services.[14] Eight primary care practices in Northern Virginia were recruited and 4500 active patients were randomly assigned to intervention and control groups. Patients in the intervention group were invited to use *myPreventiveCare*, a web-based personal health record that provides patients personalized prevention plans and individualized educational material about preventive services and chronic disease management. Control patients received ''usual'' preventive care. Outcome measurements included the percentage of intervention patients who visited *myPreventiveCare*, percentage of patients who were up-to-date with indicated preventive service, and an aggregated percentage of preventive services that were up to date. This study is an example of a three-level design in which randomization took place at the first level (patient).

The above examples illustrate randomization at each level of a three-level study. Regardless of the unit of randomization, the clustering effects between patients within the same physician and physicians within the same practice must be accounted for by the study design. An important concern in computing sample sizes for such situations is how to take these clustering effects into account. This may be accomplished by computing the design effect specific to the level at which treatment is randomly assigned.

2.2 Design effect for two-level data

In a two-level design, the calculation of the design effect involves the intracluster correlation (ICC). Detailed discussion of the ICC can be found in the writings of Murray,[3] Donner and Klar,[4] and elsewhere.[15–17]

The ICC can be expressed as $\rho = \sigma_B^2/(\sigma_B^2 + \sigma_W^2)$. In this formula, σ_B^2 is the random level two (between cluster) variance and σ_W^2 is the random level one (between subjects within the same

cluster) variance. The design effect in balanced two-level models is defined as:

$$1 + (m - 1)\rho$$

where m is the number of subjects per cluster. Application examples using this two-level design effect formula can be found in Campbell et al.[15] Compilations of ICC values and methods to compute them can be found in the existing literature.[16–18]

2.3 Design effect for three-level data

While the literature on sample size computations for data from two-level experiments is abundant, relatively few publications exist for three-level designs. Heo and Leon[5] presented a closed form power function and formulas for three-level data. Forming a test statistic using maximum likelihood estimates under a mixed models approach, they derived a design effect to compute sample sizes for three-level design where randomization takes place at the highest level. Tereenstra et al.[6] proposed formulas for the variance inflation factors in three-level designs as a function of the sample sizes, the Pearson correlation, and the correlation between two individual measurements (level one) within the same subject (level two). The authors suggested exact solutions to compute the required sample size for one level when the sample sizes for the other two levels are known. The sample size is the product of the sample size in a model without correlation and two variance inflation factors that describe the clustering of level one units within level two and level two units within level three.

Although the formulas for the design effect derived in the aforementioned studies were very helpful in guiding the sample allocation for three-level experiments, they were constructed based on the assumption that the treatments/interventions were randomized at the highest level. In the following section, we present alternate forms for the design effects when randomization occurs in level two and level one. We begin first by presenting the general notations in linear mixed models in the context of three-level data. Next, we review the derivation of the design effects when treatments are randomized at the third level. Finally, we move to suggesting the formulas for the design effect when treatments are randomized at either level two or level one.

3 Statistical model

3.1 The general model

Consider a three-level study design in which, for example, patients are nested within physicians and physicians are nested within centers. Let Y_{ijk} be the response for the k^{th} patient ($1 \le k \le n$) who was seen by the j^{th} physician ($1 \le j \le p$) within the i^{th} center ($1 \le i \le N$). The total sample size of patients is $T = Npn$.

Assuming random effects at both level two (physician) and level three (center), as well as random error at level one (patient), the total variance of the outcome, $\sigma_T^2 = \sigma_c^2 + \sigma_p^2 + \sigma_e^2$, is composed of three components: the between patients and within physicians variance σ_e^2, the between physicians and within centers variance σ_p^2, and the between centers variance σ_c^2. The intraclass correlation between patients nested within the same physician/center is specified by

$$corr\left(Y_{ijk}, Y_{ijk'}\right) = \frac{\sigma_p^2 + \sigma_c^2}{\sigma_T^2} = r \tag{1}$$

Similarly, the intraclass correlation between patients within the same center but under different physicians is specified by:

$$corr\left(Y_{ijk}, Y_{ij'k'}\right) = \frac{\sigma_c^2}{\sigma_T^2} = \rho \tag{2}$$

Using mixed models notation and assuming the error term and random effects have independent normal distributions, we write

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where

- $\mathbf{Y}$ is the $(T \times 1)$ vector of responses
- $\mathbf{X}$ is the $(T \times m)$ fixed effects full rank design matrix for the fixed effects
- $\boldsymbol{\beta}$ is the $(m \times 1)$ vector of regression fixed effects coefficients
- $\mathbf{Z}$ is the $(T \times q)$ random effects full rank design matrix for the random effects
- $\boldsymbol{\gamma}$ is the $(q \times 1)$ vector of random effects, $\boldsymbol{\gamma} \sim N(\mathbf{0}, \mathbf{G})$
- G is the variance matrix corresponding to $\boldsymbol{\gamma}$
- $\boldsymbol{\varepsilon}$ is the error vector, $\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma_e^2\mathbf{I}\right)$, where $\mathbf{I}$ is the identity matrix

Following the properties of the multivariate normal distribution, we have $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ where the positive definite variance matrix is given by

$$\mathbf{W} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \sigma_e^2\mathbf{I} \tag{3}$$

Let $\mathbf{I}_r$ be the $r \times r$ identity matrix and $\mathbf{J}_s$ the $s \times s$ matrix where every element is equal to 1. The matrix $\mathbf{W}$ is a block diagonal and can be written as

$$\mathbf{W} = \mathbf{I}_c \otimes \mathbf{V} \tag{4}$$

where $\mathbf{V}$ takes the following form (see Appendix A)

$$\mathbf{V} = \mathbf{I}_p \otimes \left(\sigma_e^2\mathbf{I}_n + \sigma_p^2\mathbf{J}_n\right) + \sigma_c^2\mathbf{J}_{pn} \tag{5}$$

Factor out the total variance σ_T^2 to obtain $\mathbf{V} = \sigma_T^2\mathbf{R}$, where, using (1) and (2),

$$\mathbf{R} = \mathbf{I}_p \otimes [(1-r)\mathbf{I}_n + (r-\rho)\mathbf{J}_n] + \rho\mathbf{J}_{pn} \tag{6}$$

In this mixed model, the sample size and power estimates depend on the variances of the fixed effects, which can be found by

$$Var\left(\widehat{\boldsymbol{\beta}}\right) = \left(\sum_{i=1}^{N} \mathbf{X}_i^T\mathbf{V}^{-1}\mathbf{X}_i\right)^{-1} \tag{7}$$

where $\mathbf{X}_i$ is the design matrix corresponding to i^{th} center.

Assuming the treatment effects are fixed and the model has no other covariates ($m = 2$), then $\boldsymbol{\beta} = (\beta_0, \beta)^T$ where β_0 is the expected response for a patient in the control group and $\beta_0 + \beta$ is the expected response for a patient in the intervention group. The incremental effect of the intervention is thus β. The hypotheses of interest can be written as

$$\mathbf{H_0} : \beta = 0$$
$$\mathbf{H_A} : \beta = d \neq 0$$

To test this hypothesis, a Wald-type test based on asymptotically normal distributions can be used. The asymptotic variance of $\sqrt{N}(\widehat{\beta} - \beta)$ is determined by the right lower corner element of the estimated variance-covariance matrix $\widehat{\Sigma} = Var[\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$.

The power to detect an effect of size d with a two-sided type I error rate of α is

$$power = 1 - \mathcal{T}_{\xi,\lambda}\left(t_{\alpha/2,\xi}\right) + \mathcal{T}_{\xi,\lambda}\left(-t_{\alpha/2,\xi}\right) \tag{8}$$

where $\mathcal{T}_{\xi,\lambda}$ is the cumulative distribution function of the t-distribution with ξ degrees of freedom and noncentrality parameter $\lambda = d/\sqrt{Var(\hat{\beta})}$. The value $t_{\alpha/2,\xi}$ is the $100\alpha/2$ percentile from the central t-distribution with ξ degrees of freedom. The values of ξ and λ depend on the level where randomization takes place (see Appendix B).

Our problem remains how to compute the variance estimator $Var(\widehat{\beta})$, which is determined by the fixed effects design matrix $\mathbf{X}$. The form of this matrix depends on the level at which randomization occurs. Once this value is obtained, we can easily plug it into the power formula. Note that in a three-level design, power depends on the sample sizes at all three levels, i.e. the values of N, p, and n, which are embedded in the computation of $Var(\widehat{\beta})$. In the following sections, we derive a computationally feasible form of $Var(\widehat{\beta})$ separately for each randomization case.

3.2 Randomize at level three

Suppose the centers are randomized such that πN centers are in the treatment arm and $(1-\pi)N$ centers are in the control arm. The design matrix for the i^{th} center is $\mathbf{X}_i = \mathbf{X}_{treat} = (\mathbf{1}_{pn}, \mathbf{1}_{pn})$ if it is randomized into the treatment group, and $\mathbf{X}_i = \mathbf{X}_{control} = (\mathbf{1}_{pn}, 0_{pn})$ if it is randomized into the control group where $\mathbf{1}_{np}$ is the $np \times 1$ matrix with all elements equal one, and $\mathbf{0}_{np}$ is the $np \times 1$ matrix with all elements equal zero.

The estimated variance is:

$$\begin{aligned}\widehat{\Sigma} &= \sigma_T^2 \lim_{N\to\infty} N\left[N\pi\left(\mathbf{X}_{treat}^T\mathbf{R}^{-1}\mathbf{X}_{treat}\right) + N(1-\pi)\left(\mathbf{X}_{control}^T\mathbf{R}^{-1}\mathbf{X}_{control}\right)\right]^{-1} \\ &= \sigma_T^2\left[\pi\left(\mathbf{X}_{treat}^T\mathbf{R}^{-1}\mathbf{X}_{treat}\right) + (1-\pi)\left(\mathbf{X}_{control}^T\mathbf{R}^{-1}\mathbf{X}_{control}\right)\right]^{-1}\end{aligned}$$

Following the derivation procedure presented by Shih,[19] we then can write

$$\widehat{\Sigma} = \sigma_T^2\left[\begin{pmatrix}1 & \pi\\ \pi & \pi\end{pmatrix}\left(\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}\right)\right]^{-1} = \frac{\sigma_T^2}{\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}}\left(\frac{1}{\pi(1-\pi)}\right)\begin{pmatrix}\pi & -\pi\\ -\pi & 1\end{pmatrix}$$

which yields

$$Var\left(\widehat{\beta}\right) = \frac{\sigma_T^2}{\pi(1-\pi)N\left(\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}\right)} \tag{9}$$

The quantity $\left(\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}\right)^{-1}$, which is proportional to the design effect (variance inflation factor), is a scalar value equal to the sum of all elements in the matrix $\mathbf{R}^{-1}$. Following Teerenstra et al.,[7] the exact expression of $\mathbf{R}^{-1}$ is given by

$$\mathbf{R}^{-1} = \frac{1}{1-r}\left[I_{pn} - \frac{\rho}{\varphi}J_{pn}\right]\left[I_p \otimes \left(I_n - \frac{r-\rho}{\gamma}J_n\right)\right] \tag{10}$$

where $\varphi = 1 + (n-1)r + n(p-1)\rho$ and $\gamma = 1 + (n-1)r - n\rho$. Thus, $\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1} = \frac{pn}{\varphi}$.

Hence, when randomization takes place at the third level, the design effect is given by

$$\varphi = 1 + (n-1)r + n(p-1)\rho,$$

And

$$Var\left(\widehat{\beta}\right) = \frac{\sigma_T^2}{\pi(1-\pi)Npn}\varphi \tag{11}$$

3.3 Randomize at level two

Assume that for each center p physicians are randomized such that πp physicians are in the treatment arm and $(1-\pi)p$ physicians are in the control arm. Furthermore, assume that there is no interaction effect between treatment and center, i.e., any difference due to treatment is the same in every center. Under these assumptions, the correlation matrix $\mathbf{R}$ remains the same as described in (6). The covariate matrix $\mathbf{X}_i$ for the i^{th} center is a $pn \times 2$ matrix with the first column of all ones and the second column with ones in the first πpn rows, and zeros in the remaining $(1-\pi)pn$ rows. We then have:

$$\widehat{\Sigma} = \sigma_T^2\left[\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i\right]^{-1} \tag{12}$$

The matrix $\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i$ is a 2×2 matrix with the following elements

$$\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i = \begin{pmatrix} \sum \text{ all elements in } \mathbf{R}^{-1} & \sum \text{ elements in first } \pi pn \text{ columns of } \mathbf{R}^{-1} \\ \sum \text{ elements in first } \pi pn \text{ rows of } \mathbf{R}^{-1} & \sum \text{ elements in first } \pi pn \text{ columns and first } \pi pn \text{ rows of } \mathbf{R}^{-1} \end{pmatrix}$$

Hence,

$$\left(\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i\right)^{-1} = \frac{\gamma}{spn\pi(1-\pi)}\begin{pmatrix} \frac{pn\pi(1-\pi)}{\gamma} + \pi^2 s & -\pi s \\ -\pi s & s \end{pmatrix}$$

where s is the sum of all elements in $\mathbf{R}^{-1}$ (see Appendix C). The right lower corner of the above matrix is:

$$Var(\widehat{\beta}) = \frac{\sigma_T^2}{\pi(1-\pi)Npn}\gamma$$

Therefore, when randomization takes place at the second level, assuming no interaction between center and treatment effects, the design effect is given by $\gamma = 1 + (n-1)r - n\rho$.

3.4 Randomize at level one

For the j^{th} physician in the i^{th} center, suppose the patients are randomized such that πn patients are in the treatment arm and $(1-\pi)n$ patients are in the control arm. In addition, assume that there is no interaction effect, i.e., the treatment effect is the same in every center–physician pair. The covariate matrix $\mathbf{X}_i$ for the i^{th} center is a $pn \times 2$ matrix of the form $\mathbf{X}_i = \mathbf{1}_p \otimes \mathbf{X}^*$. The first column of the $n \times 2$ matrix $\mathbf{X}^*$ contains all ones, whereas the second column contains ones in the first πn rows, and zeros in the remaining $(1-\pi)n$ rows.

The matrix $\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i$ is now a 2×2 matrix with the following elements

$$\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i = \begin{pmatrix} \sum \text{ all elements in } \mathbf{R}^{-1} & \sum \text{ elements in treatment columns} \\ \sum \text{ elements in treatment rows} & \sum \text{ elements in both treatment columns and treatment rows} \end{pmatrix}$$

We then can write

$$\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i = \begin{pmatrix} s & \pi s \\ \pi s & t \end{pmatrix}$$

where s is, again, the sum of all elements in $\mathbf{R}^{-1}$ and t is the sum of all elements that are in both first columns and first πpn rows of $\mathbf{R}^{-1}$ (see Appendix D).

$$\text{thus} \left(\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i\right)^{-1} = \frac{1}{st - \pi^2 s^2} \begin{pmatrix} t & -\pi s \\ -\pi s & s \end{pmatrix}$$

$$\text{and } Var\left(\widehat{\beta}\right) = \frac{\sigma_T^2}{\pi(1-\pi)Npn}(1-r)$$

Therefore, when randomization takes place at the first level without interaction effect, the design effect is $1-r$.

4 Application example

Consider an experiment to evaluate the effect of an on-line program on reducing anxiety and depression symptoms. The study is conducted on samples of patients drawn from physician patient panels where each group of physicians is drawn from a center which, in turn, is drawn from a population of centers. In such a design, depending on the context of the problem, randomization to treatment or control could occur at level one (patients), level two (physicians), or level three (centers). The sample size calculation is based on detecting a post-intervention effect size of 1.8, with a power of 0.90 and alpha = 0.05. The ICC coefficient between physicians is assumed to be $r = 0.02$ and the ICC between centers is assumed to be $\rho = 0.03$.

The above derived formulas may be applied under these assumptions to compute sample sizes for different randomization scenarios. In Table 1, the required sample sizes for level three are computed given different combinations of sample sizes from levels one and two. This is because in most multilevel design studies, the number of units in the highest level is usually limited. However,

Table 1. Example of sample size computations

		Number of centers		
Number of patients	Number of physicians	Randomize level one	Randomize level two	Randomize level three
10	4	12	15	30
	6	8	10	26
	8	6	7	24
20	4	6	9	24
	6	4	6	22
	8	3	5	22
30	4	4	7	24
	6	3	5	22
	8	2	4	20

a similar procedure can be used to compute sample sizes for level one or level two given fixed sample sizes at the other levels.

5 Discussion

Motivated by the need of a proper study design, we provided a unified method to compute sample size or power for three-level cluster randomized trials where randomization could occur at any one of the levels. We considered a two-group comparison involving one treatment arm and one control arm. We constructed our power functions by deriving the variance of the treatment effect. We focused on the test of contrast between two treatment groups since, in most published literature, multilevel designs are powered for a primary treatment contrast even when several treatments are being compared.

In some previous work, the sample size formulas were derived from the Z-test based on the standard normal distribution.[6,15] We instead chose the *t*-distribution as the sampling distribution of the test statistic, accounting for appropriate degrees of freedom. Our decision was based on the fact that in application the variance is usually unknown, and in such situation more accurate power will be obtained based on a central or noncentral *t*-distributions instead of the normal distribution.[5] Most statistical software use the *t*-test.

We presented additive models here, but our approach applies when there is variation in the treatment effect across physicians (level 2) or centers (level 3). For example, when randomization occurs at individual levels, the intervention effect might vary with different practices due to different levels of adherence to a protocol or physicians' clinical skills. If this is the case, an interaction between practices (or physicians) and treatment could be considered by adding variance components to account for interaction between the treatment effect and levels of hierarchy. Some examples of the variance structures when interaction exists can be seen in Appendix E.

Although we only focused on continuous data, our method can be adapted for other data structures, such as binary. An important issue is the choice of the analytic methods. Two common approaches are the subject- or cluster-specific and population-averaged models. Given the context of our study, a subject-specific approach utilizing GLMM can be considered,

especially when treatment is randomized at either level two or level one. Discussion on the choice of GLMM can be found in Dang et al.[20] Under this framework, different link functions can be selected and different variance structures can be constructed. Future work will include extending our approach to ordinal, nominal, or count data.

Our methods have some limitations. Firstly, our sample size formulas were derived under the assumption of balanced cluster size, whereas in practice unbalanced sizes may occur by design or by attrition. If this is the case, the variance matrix $\mathbf{W}$ remains a block diagonal matrix, where

$$\mathbf{W} = \begin{bmatrix} \mathbf{V_1} & 0 & 0 & 0 \\ 0 & \mathbf{V_2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{V_N} \end{bmatrix}.$$

However, the matrices $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_N$ no longer have the same dimensions. A closed form for the design effect when the cluster size varies may be difficult to derive, though the general structure of the variance matrix can still be used to estimate sample size by way of numerical methods. Despite this drawback, it is worthwhile to point out that the best designs in the presence of homogeneous variance are the ones where cluster sizes are equal. Imbalanced cluster size leads to a discrepancy in precision when estimations are done within clusters with smaller sizes and clusters with larger sizes. Besides, the power of the study suffers when the cluster sizes are less balanced.[21]

Secondly, our formulas for the design effects are functions of the intracluster correlations. Users should be cautioned that the use of any simple design effect in computing sample sizes depends in large part on the estimation of the ICCs. Systematic efforts have been spent by many researchers to obtain information on the values of ICC in practice[16–18]; however, sample size calculations are sensitive to small perturbations of the ICCs, especially for large cluster size.

Finally, this paper has not addressed the common issues that might be encountered even in simple sample size calculations. These complications include missing values and adjustment for loss to follow-up, studies with more than two treatments, studies with repeated measures, and situations when the between cluster variations are not equal in the two treatment groups. These and other issues are worth further consideration.

In summary, we proposed methods to derive the design effects and to compute sample sizes in three-level studies. Although many research methodologists have discussed similar topics, most previous work applied strictly to designs where randomization takes place at the highest cluster level. Much of the existing sample size methods and software for multilevel studies rely heavily on simulation algorithms. Such methods are limited for the GLMM approach since the convergence rate is very slow for data with small or medium sample sizes. Our methods focus on very specific context, although the basic technique is quite general and can be extended to various applications.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Maas CJM and Hox JJ. Sufficient sample sizes for multilevel modeling. *Eur J Res Meth Behav Soc Sci* 2005; **1**: 85–91.
2. Scherbaun CA and Ferreter JM. Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organ Res Meth* 2009; **12**: 347–367.
3. Murray DM. *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press, 1998.
4. Donner A and Klar N. *Design and analysis of cluster randomization trials in health research*. London: Arnold, 2000.
5. Heo M and Leon AC. Statistical power and sample size requirement for three level hierarchical cluster randomize trials. *Biometric* 2008; **64**: 1256–1262.
6. Teerenstra S, Moerbeek M, Achterberg T, et al. Sample size calculations for three-level cluster randomized trials. *Clin Trials* 2008; **5**: 486–495.
7. Teerenstra S, Lu B, Preisser JS, et al. Sample size considerations for GEE analyses of three-level cluster randomized trials. *Biometrics* 2010; **66**: 1230–1237.
8. Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol* 1978; **108**: 100–102.
9. Raudenbush SW and Liu X. Statistical power and optimal design for multisite randomized trials. *Psychologic Meth* 2000; **5**(2): 99–213.
10. Pituch KA, Murphy DL and Tate RL. Three-level models for indirect effects in school- and class-randomized experiments in education. *J Exp Educ* 2010; **78**(1): 60–95.
11. Glynn RJ, Brookhart MA, Stedman M, et al. Design of cluster-randomized trials of quality improvement interventions aimed at medical care providers. *Med Care* 2007; **45**: S38–43.
12. Ruf D, Berner M, Kriston L, et al. Cluster-randomized controlled trial of dissemination strategies of an online quality improvement programme for alcohol-related disorders. *Alcohol Alcohol* 2010; **45**: 70–78.
13. Grandes G, Sanchez A, Sanchez-Pinilla RO, et al. Effectiveness of physical activity advice and prescription by physicians in routine primary care. *Arch Intern Med* 2009; **7**: 694–701.
14. Krist AH, Woolf SH, Rothemich SF, et al. Randomized trial of an interactive preventive health record to enhance the delivery of recommended care. *Ann Fam Med* 2012; **10**: 312–319.
15. Campbell MK, Thomson S, Ramsay RC, et al. Sample size calculator for cluster randomized trials. *Comput Biol Med* 2004; **34**: 113–125.
16. Hedges L and Hedberg EC. Intraclass correlation values for planning group-randomized trials in education. *J Educ Eval Pol Anal* 2007; **29**: 60–87.
17. Thompson DM, Fernald DH and Mold JW. Intraclass correlation coefficients typical of cluster-randomized studies: estimates from the Robert Wood Johnson Prescription for Health Projects. *Ann Fam Med* 2012; **10**(3): 235–240.
18. Cook AJ, Bruckner T, MacLennan SG, et al. Clustering in surgical trials – database of intracluster correlations. *Trials* 2012; **13**: 2–8.
19. Shih WJ. Sample size and power calculations for periodontal and other studies with clustered sample using the method of generalized estimating equations. *Biometric J* 1997; **39**: 899–908.
20. Dang Q, Mazumdar S and Houck PR. Sample size and power calculations based on generalized linear mixed models with correlated binary outcomes. *Comput Meth Program Biomed* 2008; **91**: 122–127.
21. Eldridge SM, Ashby D and Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol* 2006; **35**: 1292–1300.

Appendix A

Consider a simple case with two centers, each center has two physicians, and each physician has two patients. Assuming random effects at both level two (physician) and level three (center), as well as random error at level one (patient), the total variance of the outcome, $\sigma_T^2 = \sigma_c^2 + \sigma_p^2 + \sigma_e^2$, is composed of three components: the between patients and within physicians variance σ_e^2, the between physicians and within centers variance σ_p^2, and the between centers variance σ_c^2

The matrix $\mathbf{Z}$ and $\mathbf{G}$ can be written as:

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{G} = \begin{bmatrix} \sigma_c^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_c^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_p^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_p^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_p^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_p^2 \end{bmatrix}$$

Under mixed model notation, the total variance of $\mathbf{Y}$ is computed by $Var(\mathbf{Y}) = \mathbf{W} = \mathbf{ZGZ}^T + \sigma_e^2\mathbf{I}$, where $\mathbf{I}$ is the 8×8 identity matrix,

$$\mathbf{ZGZ}^T = \begin{bmatrix} \sigma_c^2+\sigma_p^2 & \sigma_c^2+\sigma_p^2 & \sigma_c^2 & \sigma_c^2 & 0 & 0 & 0 & 0 \\ \sigma_c^2+\sigma_p^2 & \sigma_c^2+\sigma_p^2 & \sigma_c^2 & \sigma_c^2 & 0 & 0 & 0 & 0 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2+\sigma_p^2 & \sigma_c^2+\sigma_p^2 & 0 & 0 & 0 & 0 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2+\sigma_p^2 & \sigma_c^2+\sigma_p^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_c^2+\sigma_p^2 & \sigma_c^2+\sigma_p^2 & \sigma_c^2 & \sigma_c^2 \\ 0 & 0 & 0 & 0 & \sigma_c^2+\sigma_p^2 & \sigma_c^2+\sigma_p^2 & \sigma_c^2 & \sigma_c^2 \\ 0 & 0 & 0 & 0 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2+\sigma_p^2 & \sigma_c^2+\sigma_p^2 \\ 0 & 0 & 0 & 0 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2+\sigma_p^2 & \sigma_c^2+\sigma_p^2 \end{bmatrix},$$

and $\mathbf{W} = \mathbf{ZGZ}^T + \sigma_e^2\mathbf{I}$

$$\mathbf{W} = \begin{bmatrix} \sigma_T^2 & \sigma_c^2+\sigma_p^2 & \sigma_c^2 & \sigma_c^2 & 0 & 0 & 0 & 0 \\ \sigma_c^2+\sigma_p^2 & \sigma_T^2 & \sigma_c^2 & \sigma_c^2 & 0 & 0 & 0 & 0 \\ \sigma_c^2 & \sigma_c^2 & \sigma_T^2 & \sigma_c^2+\sigma_p^2 & 0 & 0 & 0 & 0 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2+\sigma_p^2 & \sigma_T^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_T^2 & \sigma_c^2+\sigma_p^2 & \sigma_c^2 & \sigma_c^2 \\ 0 & 0 & 0 & 0 & \sigma_c^2+\sigma_p^2 & \sigma_T^2 & \sigma_c^2 & \sigma_c^2 \\ 0 & 0 & 0 & 0 & \sigma_c^2 & \sigma_c^2 & \sigma_T^2 & \sigma_c^2+\sigma_p^2 \\ 0 & 0 & 0 & 0 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2+\sigma_p^2 & \sigma_T^2 \end{bmatrix}$$

Let

$$\mathbf{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{J}_2 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \text{and} \quad \mathbf{J}_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

$\mathbf{W}$ is then a block diagonal matrix with two blocks $\mathbf{W} = \mathbf{I}_2 \otimes \mathbf{V}$, where

$$\mathbf{V} = \begin{bmatrix} \sigma_T^2 & \sigma_c^2+\sigma_p^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2+\sigma_p^2 & \sigma_T^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_T^2 & \sigma_c^2+\sigma_p^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2+\sigma_p^2 & \sigma_T^2 \end{bmatrix} = \mathbf{I}_2 \otimes \left(\sigma_e^2\mathbf{I}_2 + \sigma_p^2\mathbf{J}_2\right) + \sigma_c^2\mathbf{J}_4.$$

In general, when there are n patients nested within p physicians, and the physicians are nested within c centers, the structure of $\mathbf{V}$ can be written as:

$$\mathbf{V} = \mathbf{I}_p \otimes \left(\sigma_e^2\mathbf{I}_n + \sigma_p^2\mathbf{J}_n\right) + \sigma_c^2\mathbf{J}_{pn}$$

Appendix B

The table below shows the breakdown of the degrees of freedom for the different scenarios. Here t is the number of treatment arms to which units are randomized. Balanced sample sizes are assumed.

	Degrees of Freedom, sum of each row (total df) = cpn				
	Mean	Treatment	Level 3	Level 2	Level 1
No treatment	1		c−1	c(p−1)	cp(n−1)
Randomize to treatment at level 3	1	t−1	t(c/t−1)	c(p−1)	cp(n−1)
Randomize to treatment at level 2	1	t−1	c−1	c(p−1) − (t−1)	cp(n−1)
Randomize to treatment at level 1	1	t−1	c−1	c(p−1)	cp(n−1) − (t−1)

Appendix C

To derive the design effect when randomization takes place at level two, we need to find $\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i$. Rewrite $\mathbf{R}^{-1}$ by expanding elements in the brackets:

$$\mathbf{R}^{-1} = \frac{1}{1-r}\left\{I_p \otimes \left[I_n - \frac{r-\rho}{\gamma}J_n\right] - \frac{\rho}{\varphi}J_{pn}\left[I_m \otimes \left(I_n - \frac{r-\rho}{\gamma}J_n\right)\right]\right\}$$

For simplification, let $\frac{r-\rho}{\gamma} = a$ and $\frac{\rho}{\varphi} = b$, thus

$$\mathbf{R}^{-1} = \frac{1}{1-r}\left\{I_p \otimes [I_n - aJ_n] + b[na-1]J_{pn}\right\}.$$

The matrix $\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i$ is a 2×2 matrix with the following elements

$$\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i = \begin{pmatrix} \sum \text{all elements in } \mathbf{R}^{-1} & \sum \text{elements in first } \pi pn \text{ columns of } \mathbf{R}^{-1} \\ \sum \text{elements in first } \pi pn \text{ rows of } \mathbf{R}^{-1} & \sum \text{elements in first } \pi pn \text{ columns and first } \pi pn \text{ rows of } \mathbf{R}^{-1} \end{pmatrix}$$

Denoting s as the sum of all elements in $\mathbf{R}^{-1}$ and writing s in terms of a, b, n, and r we have

$$\begin{aligned} s &= \frac{1}{1-r}\left[p\left(n - n^2a\right) + p^2n^2b(na-1)\right] \\ &= \frac{1}{1-r}\left[pn - pn^2a + p^2n^3ab - p^2n^2b\right] \end{aligned}$$

Denote t as the sum of all elements that are in both first πpn columns and first πpn rows of $\mathbf{R}^{-1}$, $t = 1/(1-r)\left[p\pi\left(n - n^2a\right) + p^2n^2\pi^2b(na-1)\right]$ which reduces to $t = (pn\pi(1-\pi))/\gamma + \pi^2 s$

Since $\mathbf{R}^{-1}$ is symmetric then

$$\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i = \begin{pmatrix} s & \pi s \\ \pi s & t \end{pmatrix} = \begin{pmatrix} s & \pi s \\ \pi s & \dfrac{pn\pi(1-\pi)}{\gamma} + \pi^2 s \end{pmatrix}$$

and

$$\left(\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i\right)^{-1}=\frac{\gamma}{spn\pi(1-\pi)}\begin{pmatrix}\frac{pn\pi(1-\pi)}{\gamma}+\pi^2 s & -\pi s\\ -\pi s & s\end{pmatrix}$$

Appendix D

When randomization takes place at level one, the matrix $\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i$ has the following elements

$$\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i=\begin{pmatrix}\sum\text{all elements in }\mathbf{R}^{-1} & \sum\text{elements in treatment columns}\\ \sum\text{elements n treatment rows} & \sum\text{elements in both treatment columns and treatment rows}\end{pmatrix}$$

$$\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i=\begin{pmatrix}s & \pi s\\ \pi s & t\end{pmatrix}$$

$$\left(\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i\right)^{-1}=\frac{1}{st-\pi^2 s^2}\begin{pmatrix}t & -\pi s\\ -\pi s & s\end{pmatrix}$$

Note that $t-\pi^2 s=\frac{1}{1-r}\left[p\left(n\pi-n^2\pi^2 a\right)+p^2n^2\pi^2 b(na-1)-\pi^2 p\left(n-n^2 a\right)-\pi^2p^2n^2b(na-1)\right]$

$$=\frac{pn\pi(1-\pi)}{1-r}$$

Thus, the right hand corner element of $\left(\mathbf{X}_i^T\mathbf{R}^{-1}\mathbf{X}_i\right)^{-1}$ is:

$$Var\left(\widehat{\beta}\right)=\frac{\sigma_T^2}{\pi(1-\pi)Npn}(1-r)$$

Appendix E

Example of variance structure for randomization at level two with interaction

Denote the variance of the interaction as σ_{ct}^2, which allows for the impact of treatment on the outcome measurement to vary across centers (level 2). the total variance is now

$$\sigma_T^2=\sigma_c^2+\sigma_p^2+\sigma_e^2+\sigma_{ct}^2$$

The structure of $\mathbf{V}$ can be written as follows:

$$\mathbf{V}=block(\mathbf{U},\mathbf{T})+J_{np}\left(\sigma_c^2\right),\ \text{where:}$$

$$\mathbf{U}=I_{\pi p}\otimes\left[I_n\left(\sigma_e^2\right)+J_n\left(\sigma_p^2\right)\right]+J_{\pi np}\left(\sigma_{ct}^2\right)$$

$$\mathbf{T}=I_{(1-\pi)p}\otimes\left[I_n\left(\sigma_e^2\right)+J_n\left(\sigma_p^2\right)\right]+J_{(1-\pi)np}\left(\sigma_{ct}^2\right).$$

To illustrate, suppose there are two centers, each center has three physicians and each physician has two patients. Furthermore, suppose randomization takes place at the physician level where the first two physicians are in treatment arm and the last physician is in control arm. For one center,

$$\mathbf{V} = \left[\begin{array}{cc|cc|cc} \sigma_T^2 & \omega^2 & \upsilon^2 & \upsilon^2 & \sigma_c^2 & \sigma_c^2 \\ \omega^2 & \sigma_T^2 & \upsilon^2 & \upsilon^2 & \sigma_c^2 & \sigma_c^2 \\ \hline \upsilon^2 & \upsilon^2 & \sigma_T^2 & \omega^2 & \sigma_c^2 & \sigma_c^2 \\ \upsilon^2 & \upsilon^2 & \omega^2 & \sigma_T^2 & \sigma_c^2 & \sigma_c^2 \\ \hline \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_T^2 & \omega^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 & \omega^2 & \sigma_T^2 \end{array}\right]$$

where $\upsilon^2 = \sigma_c^2 + \sigma_{ct}^2$ and $\omega^2 = \sigma_c^2 + \sigma_p^2 + \sigma_{ct}^2$.

Example of variance structure for randomization at level one with interaction

Considers the situation where the treatment effect varies across different centers. Denote the variance of the interaction as σ_{ct}^2, the total variance is now given by $\sigma_T^2 = \sigma_c^2 + \sigma_p^2 + \sigma_e^2 + \sigma_{ct}^2$

Under mixed model theory, the structure of $\mathbf{V}$ can be written as follows:

$$\mathbf{V} = I_p \otimes \mathbf{U} + J_p \otimes \mathbf{T} - I_p \otimes \mathbf{T} \text{ where}$$

$$\mathbf{U} = block\left\{\left[I_{n\pi}\left(\sigma_e^2\right) + J_{n\pi}\left(\sigma_{ct}^2\right)\right], \left[I_{n(1-\pi)}\left(\sigma_e^2\right) + J_{n(1-\pi)}\left(\sigma_{ct}^2\right)\right]\right\} + J_n\left(\sigma_c^2 + \sigma_p^2\right)$$

$$\mathbf{T} = block\left\{J_{n\pi}\left(\sigma_{ct}^2\right), J_{n(1-\pi)}\left(\sigma_{ct}^2\right)\right\} + J_n\left(\sigma_c^2\right)$$

To illustrate, consider the i^{th} center with two physicians, each physician has four patients. Suppose the patients are randomized equally to the two treatment groups, the $\mathbf{V}$ matrix takes the following form:

$$\mathbf{V} = \left[\begin{array}{cc|cc|cc|cc} \sigma_T^2 & \omega^2 & \tau^2 & \tau^2 & \upsilon^2 & \upsilon^2 & \sigma_c^2 & \sigma_c^2 \\ \omega^2 & \sigma_T^2 & \tau^2 & \tau^2 & \upsilon^2 & \upsilon^2 & \sigma_c^2 & \sigma_c^2 \\ \hline \tau^2 & \tau^2 & \sigma_T^2 & \omega^2 & \sigma_c^2 & \sigma_c^2 & \upsilon^2 & \upsilon^2 \\ \tau^2 & \tau^2 & \omega^2 & \sigma_T^2 & \sigma_c^2 & \sigma_c^2 & \upsilon^2 & \upsilon^2 \\ \hline \upsilon^2 & \upsilon^2 & \sigma_c^2 & \sigma_c^2 & \sigma_T^2 & \omega^2 & \tau^2 & \tau^2 \\ \upsilon^2 & \upsilon^2 & \sigma_c^2 & \sigma_c^2 & \omega^2 & \sigma_T^2 & \tau^2 & \tau^2 \\ \hline \sigma_c^2 & \sigma_c^2 & \upsilon^2 & \upsilon^2 & \tau^2 & \tau^2 & \sigma_T^2 & \omega^2 \\ \sigma_c^2 & \sigma_c^2 & \upsilon^2 & \upsilon^2 & \tau^2 & \tau^2 & \omega^2 & \sigma_T^2 \end{array}\right]$$

where $\tau^2 = \sigma_c^2 + \sigma_p^2$, $\upsilon^2 = \sigma_c^2 + \sigma_{ct}^2$, and $\omega^2 = \sigma_c^2 + \sigma_p^2 + \sigma_{ct}^2$.