STEPPED WEDGE CLUSTER RANDOMIZED CONTROLLED TRIALS: SAMPLE SIZE AND POWER DETERMINATIONS

by

Hsiang-Yu Chen

B.S., Taipei Medical University, Taiwan, 2006

M.S., University of Pittsburgh, 2009

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Hsiang-Yu Chen

It was defended on

July 31, 2014

and approved by

Thesis Advisor:
Chung-Chou H. Chang, PhD
Professor
Departments of Medicine, Biostatistics, and Clinical and Translational Science
School of Medicine and Graduate School of Public Health
University of Pittsburgh

Committee Member:
Ada Youk, PhD
Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Janice Zgibor, RPh, PhD
Associate Professor
Department of Epidemiology
Graduate School of Public Health
University of Pittsburgh

STEPPED WEDGE CLUSTER RANDOMIZED CONTROLLED TRIALS:

SAMPLE SIZE AND POWER DETERMINATIONS

Hsiang-Yu Chen, MS

University of Pittsburgh, 2014

ABSTRACT

Stepped wedge cluster randomized controlled trials (RCTs) are increasingly used in evaluating a causal-effect relationship between an intervention and an outcome. Sample size and power calculations are critical in designing a statistical study. Thus, the purpose of our study is to evaluate the power for both the continuous and binary responses in the context of the stepped wedge cluster design including three levels (such as hospital, physician, and individual levels).

The data structure of stepped wedge cluster RCTs is hierarchical and correlated, and we used the mixed models approach to account for the correlation of observations within each level. This approach, comprised of linear mixed models (LMM) and generalized linear mixed models (GLMM), is particularly appropriate for data with more than one level. For the continuous responses, we used LMM and GLMM with the identity link function; for the binary responses, we used GLMM with the logit link function. To compute the power of the hypothesis test for no intervention effect versus an assumed intervention effect, simulation studies were conducted and the empirical estimate of the power was obtained by calculating the percentage of the estimated intervention effect falling in the rejection region of the test when the hypothesis of no intervention effect is false.

From the results of the simulation studies, we found that the power increased as the intervention effect increased for both the continuous and binary responses, controlling for other parameters. As the overall sample size increased, a smaller minimum detectable difference with power at least 80% can be obtained. For the continuous responses only, the power increased as the within-individual correlation increased, controlling for other parameters. This increase of power with the correlation was prominent in the low intervention effect as compared to the high intervention effect.

In this study, we proposed statistical models, demonstrated power calculations, and discussed important features of sample size and power for the stepped wedge cluster design. The findings provide essential information in determining the optimal sample size and also can assure adequate power for stepped wedge cluster RCTs, which will have significant impact on research in public health in the future.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Randomized controlled trials (RCTs) are considered as the "gold standard" of study designs in evaluating a causal-effect of an intervention on the outcome of interest [1]. Through a randomization procedure using RCTs, one can prevent bias in allocating interventions and reduce confounding caused by observed and unobserved factors [2]. Cluster RCTs are a type of RCTs in which clusters of individuals, rather than independent individuals, are randomly allocated to intervention or control groups. Advantages of cluster RCTs over individual RCTs include the capability of studying those interventions which can only be administered in groups (e.g., promotion of lifestyle changes through radio broadcasting) and the capability of avoiding contamination between those with and without interventions (e.g., individuals who receive interventions may affect those who receive no interventions via communications) [3]. Because in the cluster design individuals are more similar within a cluster than between clusters, the correlation (or dependence) within the same cluster should be incorporated into the analysis to avoid making incorrect inferences [4].

Stepped wedge cluster RCTs employ a one-direction crossover design in which every cluster receives the control at baseline and then sequentially initiates the intervention at different time points. The order at which each cluster initiates the intervention is determined at random before the start of the trial. Figure 1 illustrates an example of a stepped wedge cluster RCT including five time points (i.e., four randomization steps). Every cluster receives the control at the first time point, and cluster 1 is randomly allocated to initiate the intervention at the second time point, and so on. At the last time point, all clusters will receive the intervention. More than one cluster may initiate the intervention at a time point. Once a cluster initiates the intervention, it will be on the intervention until the end of the trial. Data are collected longitudinally at each time point. The effect of the intervention can be evaluated by comparing the data collected at the time points in the intervention section with those in the control section. Brown et al. [5] and Mdege et al. [6] have provided articles of systematic review on this approach.

The advantages of choosing stepped wedge cluster RCTs over parallel or crossover cluster RCTs have been addressed in the literature. First, fewer ethical issues will arise when using stepped wedge cluster RCTs. The parallel or crossover designs require equipoise, a status of genuine uncertainty of whether the intervention is superior to the control [7]. If there exists a prior belief that the intervention will bring more benefits than harm, it may be unethical to withdraw the intervention (such as crossover design) or to allocate the control (such as parallel design) [8]. Stepped wedge cluster RCTs can solve this ethical dilemma because all clusters will eventually receive the intervention [5, 6, 9-12]. Second, stepped wedge cluster RCTs are useful if the intervention can only be implemented in stages because of logical, practical, and financial constraints [5, 12-15]. For example, the effect of a school-based anti-smoking campaign can be evaluated by a stepped wedge cluster RCT because there is only one team of trained educators traveling to each school and delivering the prevention curriculum in turn. Third, stepped wedge cluster RCTs may require fewer clusters because each cluster will receive both the intervention and the control, whereas in parallel cluster RCTs each cluster will receive either the intervention or the control [16]. Fourth, compared to parallel cluster RCTs where the analysis involves between-cluster comparisons only, stepped wedge cluster RCTs have higher statistical power because both within-cluster comparisons and between-cluster comparisons are considered in the analysis [17].

Stepped wedge cluster RCTs also have some limitations that warrant consideration. First, stepped wedge cluster RCTs may require longer trial duration than parallel cluster RCTs because it will take several steps to complete a stepped wedge cluster RCT [5, 6, 16]. If the primary outcome can be detected soon after clusters initiate the intervention, the total duration of stepped wedge cluster RCTs is likely to be acceptable. If it takes a long time for the primary outcome to occur, stepped wedge cluster RCTs may not be easily implemented because of a long total duration [6]. Second, it is almost impossible to blind the participants and those who administer the intervention because both would be aware of the changes from the control to the intervention in a stepped wedge cluster design. Thus, blinding those who assess the outcome becomes important in preventing information bias, in particular when the outcome is subjective [5, 6]. Third, stepped wedge cluster RCTs require that data are collected at each step where a new cluster initiates the intervention. If data collection is not easy or not done on a routine basis, the cost for implementation may be substantial [6, 18].

Power calculations and simulations of stepped wedge cluster RCTs have been demonstrated by Hussey et al [16]. Both a cluster-level model (addressing the responses as the mean of each cluster) and an individual-level model were proposed to characterize the design and analysis of stepped wedge cluster RCTs. The authors compared the three approaches: linear mixed models (LMM), generalized linear mixed models (GLMM), and generalized estimating equations (GEE). LMM can be used when the responses were continuous and normally distributed. It may also be used in the situation where the individual-level responses were not normal (e.g., binary) and the cluster sizes were equal, because the cluster-level responses (e.g., proportion) can then be modeled as continuous variables. However, if the individual-level responses were not normal but the cluster sizes were unequal, LMM may not be useful because the analysis at the cluster level would require weights. In this situation, using GLMM or GEE to directly model the individual-level responses will be preferable. The GLMM, an extension to LMM, can be used to model normal and non-normal responses via a link function. More specifically, the mean of responses with any distribution in the exponential family can be linked to the linear predictors through a link function. Using GEE as an alternative approach, one can handle normal and non-normal responses with flexibility. It can also provide more robust results when there is an issue of misspecification of the variance structure.

However, Hussey et al. presented the models which addressed the stepped wedge cluster design with only two levels (i.e., cluster and individual levels). This may not be applicable to the general setting of a clinical research where three levels (i.e., hospital, physician, and individual levels) exist; in other words, where individuals' repeated measures are nested within a physician who is nested within a hospital. Additionally, the power calculations and simulations in their studies were for the binary responses. The purpose of the current study is to evaluate the power for the continuous responses and the binary responses in the context of the stepped wedge cluster design with hospital, physician, and individual levels.

In Section 2, we propose the models for the continuous responses and the binary responses, and present the methods of power calculations. In Section 3, we carry out a set of simulation studies to demonstrate the methods shown in Section 2. In Section 4, we describe an example which uses the stepped wedge cluster design with market, dialysis center, and patient levels, and then apply our methods of power calculations to the example. In Section 5, we summarize our results and discuss the aspects of future studies.

2.0 METHODS

2.1 MODELS

The data structure of stepped wedge cluster RCTs is hierarchical (i.e., multi-level). As shown in Figure 1 as an example of study design, the repeated measures are nested within an individual who is nested within a cluster. Because of the nature of hierarchy, the data of stepped wedge cluster RCTs are correlated; in other words, the repeated measures are more similar within an individual than between individuals, and the individuals are more similar within a cluster than between clusters. The mixed models approach is widely used for hierarchical data to account for the dependence of observations within the same level. By adding random variation in the model, mixed models provide a more convenient approach to handle the dependence of observations and also allow us to make inferences on a wider population. Compared to generalized estimating equations (GEE), mixed models are particularly appropriate to be used on data with more than one level. For the current study, we specifically adopt the random effects models (also known as random intercept models) to control for the dependence within each level.

Mixed models are comprised of linear mixed models (LMM) and generalized linear mixed models (GLMM). LMM is used to analyze data with a normal distribution; GLMM is used to analyze data with any distribution of the exponential family (e.g., normal, Bernoulli, and Poisson distributions). GLMM is an extension to LMM, and utilizes a certain link functions to map the responses from the observational scale (e.g., 0 and 1 for the binary responses) to the real scale ($-\infty, +\infty$). In other words, the mean of responses with any distribution of the exponential family are transformed by a link function so that it can be related to the linear predictors. LMM is a special case of GLMM in which the distribution of responses is assumed to be normal and the identity link function is applied to relate the mean of responses to the linear predictors. In Sections 2.1.1 and 2.1.2, we will introduce the statistical models of the continuous responses and the binary responses for a stepped wedge cluster design. The model of the binary responses is shown as an example of the GLMM models because it is the most frequent situation where data are not normal. Similar results can be derived for other types of data using GLMM.

In the current study, individuals are nested within a physician who is nested within a hospital. We assume that in a stepped wedge cluster design we roll out I hospitals over J time points, and for each hospital there are K physicians, and for each physician we recruit L individuals. Let random variable Y_{ijkl} represent the response for individual l (l = 1, …, L) at time j (j = 1, …, J) and the individual is recruited from physician k (k = 1, …, K) at hospital i (i = 1, …, I). The hospital is considered as the unit of randomization (i.e., cluster); that is, all individuals in the same hospital will initiate the intervention at the same time point.

2.1.1 Model of the continuous responses

Suppose that Y_{ijkl} is a continuous response and assumed to be normally distributed, the model is presented as follows:

$$Y_{ijkl} = \mu + \alpha_i + \beta_{ik} + t_j + X_{ij}\theta + e_{ijkl,} \qquad (1)$$

where μ is the grand mean; α_i is the random effect of hospital i such that $\alpha_i \sim N(0, \tau^2)$; β_{ik} is the random effect of physician k nested within hospital i and $\beta_{ik} \sim N(0, \varphi^2)$; t_j is the fixed effect of time j; X_{ij} is the indicator of intervention for hospital i at time j (X_{ij} = 1 for intervention; X_{ij} = 0 for control); θ is the fixed effect of intervention; and e_{ijkl} is the residual error such that $e_{ikl} = [e_{i1kl}, e_{i2kl}, \ldots, e_{iJkl}]^T \sim N_J(0, \Sigma)$ and $\Sigma = \sigma^2 \begin{bmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{bmatrix}_{J\times J}$. We also assume that α_i and β_{ik} are independent of each other in addition to being independent of t_j, X_{ij}, and e_{ijkl}. In this model, α_i and β_{ik} provide the hospital- and physician- specific intercepts, which can be thought of as natural heterogeneity due to unmeasured factors between hospitals and between physicians, respectively. Besides, α_i can be used to account for the dependence between physicians in each hospital, and β_{ik} can be used to account for the dependence between individuals for each physician. The dependence between repeated measures for each individual is explained by the residual error, e_{ijkl}.

2.1.2 Model of the binary responses

Suppose that Y_{ijkl} is a binary response (taking values of 0 or 1), the model is presented as follows:

$$\log\left\{\frac{\mathbf{Pr(Y_{ijkl}=1)}}{\mathbf{Pr(Y_{ijkl}=0)}}\right\} = \mu + \alpha_i + \beta_{ik} + \gamma_{ikl} + t_j + X_{ij}\theta. \quad (2)$$

That is, the mean of the responses is related to the linear predictors through a logit link function. Most parameters are the same as we describe for model (1) except that γ_{ikl} is added and the residual error is removed. γ_{ikl} is the random effect of individual l that is nested within physician k and hospital i such that $\gamma_{ikl} \sim N(0, \zeta^2)$. Instead of using the residual error to control the dependence of repeated measures for each individual in model (1), an individual-specific random intercept is included in model (2) to account for natural heterogeneity between individuals and handle the dependence of repeated measures within an individual.

2.2 POWER CALCULATIONS

The power of a statistical test is the probability that the test rejects the null hypothesis when the null hypothesis is false. That is,

Power = Prob (reject the null hypothesis | the null hypothesis is false).

Thus, we first specify a significance level (known as type I error, α), and then determine the rejection region for which the null probability in the rejection region achieves α. Based on this rejection region, the power can be obtained from the alternative probability in the rejection region. For the current study, the goal is to test the hypothesis H_0: $\theta = 0$ (no intervention effect) versus H_a: $\theta = \theta_A$ where $\theta_A \neq 0$ for both model (1) and model (2) using the stepped wedge cluster design with I hospitals, J time points, K physicians under each hospital, and L individuals under each physician. After receiving $\hat{\theta}$ from LMM and/or GLMM approaches, we use the distribution of the Wald statistic $Z = \hat{\theta} / \sqrt{\mathrm{Var}(\hat{\theta})}$ to obtain the empirical estimate of the power by computing the percentage of $\hat{\theta}$ falling in the rejection region when the null hypothesis is false.

3.0 SIMULATION STUDIES

We carried out a set of simulation studies to assess the power of the hypothesis test H_0: $\theta = 0$ versus H_a: $\theta = \theta_A$ where $\theta_A \neq 0$. At each of repetitions, the simulated dataset consisted of 12 hospitals (I = 12), 5 time points (J = 5), 5 physicians under each hospital (K = 5), and 50 individuals under each physician (L=50). Because the hospital was the unit of randomization, all hospitals started with the control at the first time point (j = 1) and then every 3 hospitals switched to the intervention at each randomization step. Then, the simulated datasets were modeled using LMM with function lme() and/or GLMM with function glmmPQL() in R version 2.14.1. In addition, the performance of the simulation studies was evaluated using three measures: 95% coverage rate of the true θ, empirical standard error of $\hat{\theta}$ (i.e., $\mathrm{SE}(\hat{\theta})$), and the mean of standard error of $\hat{\theta}$ (i.e., $\widehat{\mathrm{SE}}(\hat{\theta})$). For the current simulation studies, we set the parameters based on the stepped wedge cluster design from Hussey et al. [16], and extended the design to the one with three levels. Other specifications can also be applied as we will describe in Section 4.

3.1 SIMULATION RESULTS OF THE CONTINUOUS RESPONSES

The continuous longitudinal profile was generated based on model (1) with between-hospital variance τ^2 = 0.01 and between-physician variance φ^2 = 0.5. The residual error was specified as compound symmetry with variance $\sigma^2 = 1$ and correlation $\rho = 0.2 - 0.9$. The grand mean μ was fixed as 0.5 and the intervention effect θ was -0.03, -0.05, and -0.07. We assumed no time effect, $t_j = 0$. The level of significance was specified as $\alpha = 0.05$. Then, the dataset was modeled using both LMM and GLMM with the identity link function. Each scenario was simulated with 1,000 repetitions.

Table 1 and Figure 2 summarize the results of the simulation studies for the continuous responses. The results from LMM and GLMM with the identity link function are similar to each other. When the intervention effect is -0.03 and the correlation is 0.9, the power is 0.894; when the intervention effect is -0.05 and the correlation is 0.7, the power is 0.872-0.873; when the intervention effect is -0.07 and the correlation is 0.3-0.4, the power is 0.811-0.868. That is, to achieve power 0.8-0.9, observations with the intervention, compared to those without the intervention, have lower responses by 0.03 when the correlation is 0.9; by 0.05 when the correlation is 0.7; by 0.07 when the correlation is 0.3-0.4. This shows that when the intervention effect is fixed, the power increases as the correlation increases. This increase of power with the correlation is steeper in the low intervention effect (e.g., $\theta = -0.03$) as compared to the high intervention effect (e.g., $\theta = -0.07$). When the correlation is fixed, the power increases as the intervention effect (absolute value of θ) increases. In all scenarios, the coverage rates approximate the nominal 95%, and $\mathrm{SE}(\hat{\theta})$ and $\widehat{\mathrm{SE}}(\hat{\theta})$ are close to each other. This evidence validates the simulation results of the continuous responses.

3.2 SIMULATION RESULTS OF THE BINARY RESPONSES

The binary longitudinal profile was generated based on model (2) with between-hospital variance τ^2 = 0.01, between-physician variance φ^2 = 0.05, and between-individual variance ζ^2 = 0.1. Under the logit link function, the grand mean μ was fixed as 0.5 and the intervention effect θ was varied between -0.10 and -0.20. We assumed no time effect, t_j = 0. The level of significance was specified as α = 0.05. Then, the dataset was modeled using GLMM with the logit link function. Each scenario was simulated with 2,000 repetitions.

Table 2 and Figure 3 summarize the results of the simulation studies for the binary responses. The results show that the power is 0.864 when the intervention effect is -0.18 under the logit link function (corresponding to the odds ratio = exp(-0.18) = 0.84). That is, when the power is 0.864, the odds of the responses decrease 16% for those with the intervention compared to those without the intervention. As we expected, the power increases as the intervention effect (absolute value of θ) increases. In all scenarios, the coverage rates are close to the nominal 95%, and $\mathrm{SE}(\hat{\theta})$ and $\widehat{\mathrm{SE}}(\hat{\theta})$ are similar to each other. Thus, we can conclude that the simulation results of the binary responses are appropriate.

4.0 APPLICATIONS

Hospice care is underused in end-stage renal disease (ESRD) patients on dialysis. Among annual 65,000 deaths of ESRD patients on dialysis in the United States, only 20% of them receive hospice care in the last month of life, which is remarkably lower than patients with cancer (55%) and patients with congestive heart failure (39%) [19]. Because of the underuse of hospice care in ESRD patients on dialysis, three-quarters of them die with distressing symptoms and half of them are admitted to an intensive care unit at the end of life [20, 21]. This situation is due to the current Medicare payment policy which structurally separates ESRD benefits (i.e., dialysis) from hospice benefits. More specifically, patients who have a terminal diagnosis as ESRD are required to withdraw from dialysis in order to receive hospice care. Since withdrawal from dialysis for ESRD patients leads to death in 4 days for 50% of patients [22], most ESRD patients decide to continue dialysis, rather than receive hospice care but withdraw from dialysis.

To address this situation, the Center for Research on Health Care at the University of Pittsburgh School of Medicine proposed concurrent supportive care for ESRD patients on dialysis. In collaboration with dialysis and hospice providers, ESRD patients will concurrently receive dialysis and hospice care, and a payment model will be developed to cover the expenses by employing a capped bundled payment with shared provider risks and saving to control costs. The goal of concurrent supportive care was to increase advance care planning, enhance hospice use, improve qualify of dying, and reduce total Medicare costs mainly through decreasing acute care utilization.

A stepped wedge cluster design was planned for evaluating the effect of concurrent supportive care in ESRD patients on dialysis. Corresponding to Section 2, hospital-physician-individual relation was equivalent to market-dialysis center-patient relation in this application. That is, the patients were nested within a dialysis center which was nested within a market. For a stepped wedge cluster design, we planned to have about 250 patients in 12 markets (I = 12) nationwide over 7 time points (J = 7), and every 2 markets will initiate to receive concurrent supportive care at each randomization step. In each market, there will be 6 dialysis centers (K = 6). Two designs were considered and compared; one with total patients fewer than 250 and one with total patients more than 250. In comparison, Design 1 included 12 markets (I = 12), 6 dialysis centers within each market (K = 6), and 3 patients within each dialysis center (L = 3) (i.e., 216 patients in total), whereas Design 2 included 12 markets (I = 12), 6 dialysis centers within each market (K = 6), and 4 patients within each dialysis center (L = 4) (i.e., 288 patients in total). The major difference between the two designs was the number of patients within each dialysis center.

Similar to Section 3, a set of simulation studies was conducted to assess the power of the hypothesis test for the effect of concurrent supportive care, H_0: $\theta = 0$ versus H_a: $\theta = \theta_A$ where $\theta_A \neq 0$. In fact, the main purpose was to determine the minimum detectable difference for the effect of concurrent supportive care in order to reach at least 80% power. We implemented LMM with function lme() and/or GLMM with function glmmPQL() in R. Besides, 95% coverage rate of the true θ, empirical standard error of $\hat{\theta}$ (i.e., $\mathrm{SE}(\hat{\theta})$), and the mean of standard error of $\hat{\theta}$ (i.e., $\widehat{\mathrm{SE}}(\hat{\theta})$) were used to evaluate the results of the simulation studies.

4.1 APPLICATION RESULTS OF THE CONTINUOUS RESPONSES

The continuous responses in this application were 1) Medicare costs, and 2) Consumer Assessment and Reports of End-of-life Care (CARE) score, which is a measurement of quality of dying. For ease of interpretation, all the parameters were standardized as mean = 0 and variance = 1, and thus the effect of concurrent supportive care θ can be expressed as the averaged change in the continuous responses of one standard deviation above and below the mean, comparing those with and without concurrent supportive care. Accordingly, based on model (1) between-market variance τ^2 and between-dialysis center variance φ^2 were standardized as 1. The residual error was assumed to be compound symmetry with standardized variance $\sigma^2 = 1$ and correlation $\rho = 0.1, 0.3, 0.5$ based on preliminary results. The grand mean μ was standardized as 0. We assumed no time effect, $t_j = 0$. Varied effects of concurrent supportive care θ were used to find the situation with power ≥ 80%. The level of significance was specified as $\alpha = 0.05$. Analyses were conducted using both LMM and GLMM with the identity link function. Each scenario was simulated with 1,000 repetitions.

Table 3 and Table 4 present the simulation results of the continuous responses for Design 1 and Design 2, respectively. In both designs, the results from LMM and GLMM with the identity link function are similar to each other. In Design 1, to achieve power ≥ 80%, the minimum detectable difference of one standard deviation above and below the mean is 0.26 (absolute value of θ) when the correlation is 0.1; 0.22, when the correlation is 0.3; 0.20, when the correlation is 0.5. In Design 2, the minimum detectable difference of one standard deviation above and below the mean with power ≥ 80% is 0.22, when the correlation is 0.1; 0.20, when the correlation is 0.3; 0.16, when the correlation is 0.5. Comparing the results from Design 1 (smaller overall sample size) with those from Design 2 (larger overall sample size), we can expect that the greater the overall sample size, the smaller the minimum detectable difference. In addition, the coverage rates approximate the nominal 95%, and $\mathrm{SE}(\hat{\theta})$ and $\widehat{\mathrm{SE}}(\hat{\theta})$ are close to each other in all scenarios. Thus, the simulation results of the continuous responses are validated.

4.2 APPLICATION RESULTS OF THE BINARY RESPONSES

The binary responses in this application were 1) advance care plan (Y/N), and 2) hospice enrollment (Y/N). Similar to the continuous responses, the setting of the parameters was standardized as mean = 0 and variance = 1. Based on model (2), between-market variance τ^2, between-dialysis center variance φ^2, and between-patient variance ζ^2 were standardized as 1. Under the logit link function, the grand mean μ was standardized as 0. We assumed no time effect, $t_j = 0$. We studied several effects of concurrent supportive care θ in order to find the situations in which power ≥ 80%. The level of significance was specified as $\alpha = 0.05$. Analysis was conducted using GLMM with the logit link function. Each scenario was simulated with 2,000 repetitions.

Table 5 and Table 6 show the simulation results of the binary responses for Design 1 and Design 2, respectively. In Design 1, to achieve power ≥ 80%, the minimum detectable difference for the effect of concurrent supportive care is 0.60 (absolute value of θ) under the logit link function, which can be expressed as an odds ratio of $\exp(\theta) = \exp(-0.60) = 0.55$. In Design 2, the minimum detectable difference for the effect of concurrent supportive care with power ≥ 80% is 0.52 under the logit link function, corresponding to an odds ratio of $\exp(\theta) = \exp(-0.52) = 0.59$. Like the continuous responses, as the overall sample size increases from Design 1 to Design 2, we can detect a smaller difference between those with and without concurrent supportive care when power $\geq$ 80%. The coverage rates are around 0.92, which is slightly smaller than 0.95. To further improve the coverage rates, a possible solution is to use a better estimator for $\mathrm{SE}(\hat{\theta})$.

5.0 DISCUSSION

For a stepped wedge cluster RCT with three levels (e.g., an individual is nested within a physician who is nested within a hospital), we proposed statistical models for the continuous responses and the binary responses, and computed the power for testing the intervention effect using simulation studies. For a continuous outcome response, parameters required to calculate the power included the sample size of each level, the number of time points (or randomization steps), between-hospital variance, between-physician variance, residual variance (such as correlation structure, variance, and correlation), grand mean of the responses, intervention effect, and time effect. For a binary outcome response, parameters required to calculate the power included the sample size of each level, the number of time points (or randomization steps), between-hospital variance, between-physician variance, between-individual variance, percent responses for the population, intervention effect, and time effect. We also found that the power increased as the intervention effect increased for both the continuous and binary responses, controlling for other parameters. Besides, as the overall sample size increased, a smaller minimum detectable difference with a sufficient power can be achieved. For the continuous responses only, the power increased as the within-individual correlation increased, controlling for other parameters. This increase of power with the correlation was prominent in the low intervention effect as compared to the high intervention effect.

We assumed an equal sample size for each level in the present study; that is, there were an equal number of individuals within every physician and an equal number of physicians within every hospital. In addition, we also assumed that an equal number of hospitals started to receive the intervention at every randomization step. Although this setting with equal sample size for each level would enhance the statistical power [16], this may bring about some restrictions on the study design of a stepped wedge cluster RCT. For example, each physician was required to recruit 50 individuals in Section 3, and thus a physician who recruited only 40 individuals was not eligible to participate in the study. Therefore, relaxing the assumption of equal sample size for each level would introduce flexibility in the study design, and power calculations for the situation with varied sample sizes for each level are necessary in future research.

Stepped wedge cluster RCTs are increasingly used in a variety of areas, such as cardiovascular diseases [23, 24], nutrition [12, 25], respiratory diseases [14], maternal and child health [26, 27], and cancers [13, 28]. Sample size and power calculations are critical in designing a trial. However, the available research regarding sample size and power for stepped wedge cluster RCTs [16] did not include three levels like a general clinical research (i.e., hospital, physician, and individual levels), and did not discuss power calculations and simulations for the continuous responses. Therefore, our study is unique and provides essential information to determine optimal sample size and assure adequate power for a stepped wedge cluster RCT.

APPENDIX: TABLES AND FIGURES

Table 1. Simulation results for the continuous responses to assess the power of the hypothesis test

H_0: $\theta = 0$ vs. H_a: $\theta = \theta_A$ where $\theta_A \neq 0$*

		Using LMM				Using GLMM with the identity link function			
θ	ρ	Power	$SE(\hat{\theta})$	$\widehat{SE}(\hat{\theta})$	95% coverage rate	Power	$SE(\hat{\theta})$	$\widehat{SE}(\hat{\theta})$	95% coverage rate
-0.03	0.2	0.198	0.0266624	0.0266110	0.949	0.198	0.0266643	0.0266031	0.948
	0.3	0.216	0.0249608	0.0248981	0.949	0.216	0.0249624	0.0248910	0.949
	0.4	0.247	0.0231128	0.0230563	0.950	0.245	0.0231141	0.0230500	0.949
	0.5	0.287	0.0210593	0.0210527	0.951	0.288	0.0210603	0.0210471	0.951
	0.6	0.349	0.0188270	0.0188344	0.951	0.349	0.0188278	0.0188296	0.951
	0.7	0.454	0.0163377	0.0163147	0.951	0.454	0.0163382	0.0163108	0.950
	0.8	0.614	0.0133531	0.0133243	0.949	0.614	0.0133534	0.0133212	0.949
	0.9	0.894	0.0094431	0.0094239	0.949	0.894	0.0094432	0.0094218	0.949
-0.05	0.2	0.470	0.0266999	0.0266112	0.949	0.470	0.0267017	0.0266033	0.948
	0.3	0.512	0.0249608	0.0248981	0.949	0.513	0.0249624	0.0248910	0.949
	0.4	0.570	0.0231080	0.0230564	0.950	0.570	0.0231093	0.0230501	0.949
	0.5	0.663	0.0210535	0.0210520	0.950	0.662	0.0210545	0.0210465	0.950
	0.6	0.775	0.0188346	0.0188340	0.951	0.777	0.0188353	0.0188293	0.951
	0.7	0.873	0.0163351	0.0163147	0.951	0.872	0.0163356	0.0163108	0.950
	0.8	0.960	0.0133681	0.0133245	0.949	0.960	0.0133685	0.0133215	0.949
	0.9	0.999	0.0094232	0.0094238	0.949	0.999	0.0094233	0.0094218	0.949
-0.07	0.2	0.771	0.0266624	0.0266110	0.949	0.771	0.0266643	0.0266031	0.948
	0.3	0.811	0.0249603	0.0248981	0.949	0.812	0.0249619	0.0248910	0.949
	0.4	0.868	0.0231204	0.0230563	0.950	0.868	0.0231216	0.0230500	0.949
	0.5	0.916	0.0210613	0.0210524	0.951	0.916	0.0210623	0.0210468	0.951
	0.6	0.960	0.0188523	0.0188346	0.951	0.960	0.0188531	0.0188298	0.951
	0.7	0.991	0.0163221	0.0163149	0.951	0.991	0.0163226	0.0163110	0.950
	0.8	0.999	0.0133463	0.0133243	0.949	0.999	0.0133466	0.0133212	0.949
	0.9	1.000	0.0094345	0.0094238	0.949	1.000	0.0094346	0.0094217	0.949

*We assumed a stepped wedge cluster design with 12 hospitals, 5 time points, 5 physicians under each hospital, and 50 individuals under each physician. Based on model (1), the parameters were fixed as $\tau^2 = 0.01$, $\varphi^2 = 0.5$, $\mu = 0.5$, and $t_j = 0$ (no time effect). The residual error was specified as compound symmetry with $\sigma^2 = 1$ and varied ρ. Type I error was set as 0.05. Each scenario was simulated with 1,000 repetitions.

Abbreviations: LMM, linear mixed models; GLMM, generalized linear mixed models; τ^2, between-hospital variance; φ^2, between-physician variance; σ^2, residual variance; ρ, correlation; μ, grand mean; θ, intervention effect; t_j, time effect.

Table 2. Simulation results for the binary responses to assess the power of the hypothesis test H_0: $\theta = 0$ vs. H_a: $\theta = \theta_A$ where $\theta_A \neq 0$[*]

		Using GLMM with the logit link function			
θ	Odds ratio[†]	Power	$SE(\hat{\theta})$	$\widehat{SE}(\hat{\theta})$	95% coverage rate
-0.10	0.90	0.396	0.0581411	0.0573366	0.948
-0.12	0.89	0.525	0.0576691	0.0572949	0.954
-0.14	0.87	0.658	0.0580020	0.0572549	0.946
-0.16	0.85	0.770	0.0578155	0.0572142	0.950
-0.18	0.84	0.864	0.0578998	0.0571781	0.945
-0.20	0.82	0.926	0.0578106	0.0571418	0.945

[*] We assumed a stepped wedge cluster design with 12 hospitals, 5 time points, 5 physicians under each hospital, and 50 individuals under each physician. Based on model (2), the parameters were fixed as $\tau^2 = 0.01$, $\varphi^2 = 0.05$, $\zeta^2 = 0.1$, $\mu = 0.5$, and $t_j = 0$ (no time effect). Type I error was set as 0.05. Each scenario was simulated with 2,000 repetitions.

[†] Effect size can be expressed as odds ratio, which is $\exp(\theta)$.

Abbreviations: GLMM, generalized linear mixed models; τ^2, between-hospital variance; φ^2, between-physician variance; ζ^2, between-individual variance; μ, grand mean; θ, intervention effect; t_j, time effect.

Table 3. Application of the continuous responses for a stepped wedge cluster design with 12 markets, 7 time points, 6 dialysis centers under each market, and 3 patients under each dialysis center to assess the power of the hypothesis test H_0: $\theta = 0$ vs. H_a: $\theta = \theta_A$ where $\theta_A \neq 0$*

θ	ρ	Using LMM				Using GLMM with the identity link function			
		Power	$SE(\hat{\theta})$	$\widehat{SE}(\hat{\theta})$	95% coverage rate	Power	$SE(\hat{\theta})$	$\widehat{SE}(\hat{\theta})$	95% coverage rate
-0.24	0.1	0.789	0.0856657	0.0863325	0.953	0.791	0.0856874	0.0860788	0.953
-0.26		**0.840**	**0.0856030**	**0.0863342**	**0.953**	**0.840**	**0.0856232**	**0.0860805**	**0.953**
-0.28		**0.898**	**0.0856056**	**0.0863341**	**0.953**	**0.901**	**0.0856268**	**0.0860804**	**0.953**
-0.30		0.949	0.0856030	0.0863342	0.953	0.949	0.0856232	0.0860805	0.953
-0.20	0.3	0.735	0.0754826	0.0761870	0.953	0.735	0.0754985	0.0759671	0.952
-0.22		**0.819**	**0.0754256**	**0.0761874**	**0.953**	**0.820**	**0.0754400**	**0.0759676**	**0.952**
-0.24		**0.877**	**0.0755268**	**0.0761905**	**0.953**	**0.878**	**0.0755411**	**0.0759707**	**0.952**
-0.26		0.940	0.0754256	0.0761874	0.953	0.940	0.0754400	0.0759676	0.952
-0.18	0.5	0.793	0.0638411	0.0644316	0.955	0.796	0.0638503	0.0642491	0.953
-0.20		**0.865**	**0.0638429**	**0.0644302**	**0.955**	**0.866**	**0.0638521**	**0.0642477**	**0.953**
-0.22		0.940	0.0638241	0.0644318	0.955	0.940	0.0638330	0.0642493	0.953

*Based on model (1), the parameters were standardized as $\tau^2 = 1$, $\varphi^2 = 1$, $\mu = 0$, and $t_j = 0$ (no time effect). The residual error was specified as compound symmetry with $\sigma^2 = 1$ and varied ρ. Type I error was set as 0.05. Each scenario was simulated with 1,000 repetitions.

Abbreviations: LMM, linear mixed models; GLMM, generalized linear mixed models; τ^2, between-market variance; φ^2, between-dialysis center variance; σ^2, residual variance; ρ, correlation; μ, grand mean; θ, effect of concurrent supportive care; t_j, time effect.

Table 4. Application of the continuous responses for a stepped wedge cluster design with 12 markets, 7 time points, 6 dialysis centers under each market, and 4 patients under each dialysis center to assess the power of the hypothesis test H_0: $\theta = 0$ vs. H_a: $\theta = \theta_A$ where $\theta_A \neq 0$*

		Using LMM				Using GLMM with the identity link function			
θ	ρ	Power	$SE(\hat{\theta})$	$\widehat{SE}(\hat{\theta})$	95% coverage rate	Power	$SE(\hat{\theta})$	$\widehat{SE}(\hat{\theta})$	95% coverage rate
-0.20	0.1	0.777	0.0747652	0.0748367	0.947	0.779	0.0747727	0.0746718	0.947
-0.22		**0.848**	**0.0747652**	**0.0748367**	**0.947**	**0.848**	**0.0747727**	**0.0746718**	**0.947**
-0.24		0.902	0.0747652	0.0748367	0.947	0.904	0.0747727	0.0746718	0.947
-0.18	0.3	0.793	0.0659512	0.0660304	0.947	0.793	0.0659562	0.0658876	0.947
-0.20		**0.865**	**0.0659512**	**0.0660304**	**0.947**	**0.864**	**0.0659562**	**0.0658876**	**0.947**
-0.22		0.921	0.0659512	0.0660304	0.947	0.921	0.0659562	0.0658876	0.947
-0.14	0.5	0.724	0.0557517	0.0558315	0.947	0.725	0.0557546	0.0557129	0.947
-0.16		**0.827**	**0.0557517**	**0.0558315**	**0.947**	**0.830**	**0.0557546**	**0.0557129**	**0.947**
-0.18		0.908	0.0557517	0.0558315	0.947	0.909	0.0557546	0.0557129	0.947

*Based on model (1), the parameters were standardized as $\tau^2 = 1$, $\varphi^2 = 1$, $\mu = 0$, and $t_j = 0$ (no time effect). The residual error was specified as compound symmetry with $\sigma^2 = 1$ and varied ρ. Type I error was set as 0.05. Each scenario was simulated with 1,000 repetitions.

Abbreviations: LMM, linear mixed models; GLMM, generalized linear mixed models; τ^2, between-market variance; φ^2, between-dialysis center variance; σ^2, residual variance; ρ, correlation; μ, grand mean; θ, effect of concurrent supportive care; t_j, time effect.

Table 5. Application of the binary responses for a stepped wedge cluster design with 12 markets, 7 time points, 6 dialysis centers under each market, and 3 patients under each dialysis center to assess the power of the hypothesis test H_0: $\theta = 0$ vs. H_a: $\theta = \theta_A$ where $\theta_A \neq 0$[*]

		Using GLMM with the logit link function			
θ	Odds ratio[†]	Power	$SE(\hat{\theta})$	$\widehat{SE}(\hat{\theta})$	95% coverage rate
-0.58	0.56	0.779	0.2170503	0.1936428	0.918
-0.60	**0.55**	**0.802**	**0.2162086**	**0.1937347**	**0.921**
-0.62	**0.54**	**0.828**	**0.2150595**	**0.1938322**	**0.927**
-0.64	**0.53**	**0.850**	**0.2148441**	**0.1939434**	**0.923**
-0.66	**0.52**	**0.872**	**0.2148470**	**0.1940510**	**0.919**
-0.68	**0.51**	**0.888**	**0.2151393**	**0.1941631**	**0.922**
-0.70	**0.50**	**0.898**	**0.2155592**	**0.1942769**	**0.918**
-0.72	0.49	0.913	0.2145534	0.1943834	0.915

[*] Based on model (2), the parameters were standardized as $\tau^2 = 1$, $\varphi^2 = 1$, $\zeta^2 = 1$, $\mu = 0$, and $t_j = 0$ (no time effect). Type I error was set as 0.05. Each scenario was simulated with 2,000 repetitions.

[†] Effect size can be expressed as odds ratio, which is $\exp(\theta)$.

Abbreviations: GLMM, generalized linear mixed models; τ^2, between-market variance; φ^2, between-dialysis center variance; ζ^2, between-patient variance; μ, grand mean; θ, effect of concurrent supportive care; t_j, time effect.

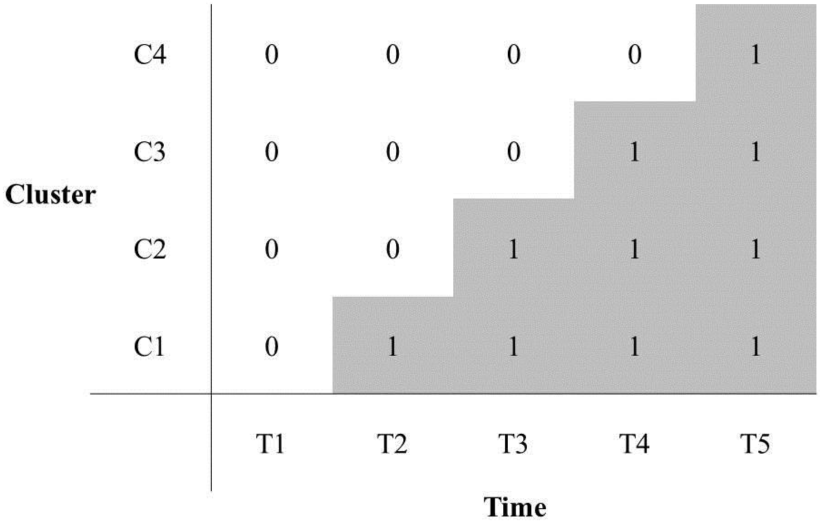
Table 6. Application of the binary responses for a stepped wedge cluster design with 12 markets, 7 time points, 6 dialysis centers under each market, and 4 patients under each dialysis center to assess the power of the hypothesis test H_0: $\theta = 0$ vs. H_a: $\theta = \theta_A$ where $\theta_A \neq 0$[*]

		Using GLMM with the logit link function			
θ	Odds ratio[†]	Power	$SE(\hat{\theta})$	$\widehat{SE}(\hat{\theta})$	95% coverage rate
-0.50	0.61	0.785	0.1843711	0.1682908	0.919
-0.52	**0.59**	**0.812**	**0.1853929**	**0.1683681**	**0.918**
-0.54	**0.58**	**0.836**	**0.1852366**	**0.1684518**	**0.918**
-0.56	**0.57**	**0.860**	**0.1854834**	**0.1685291**	**0.913**
-0.58	**0.56**	**0.884**	**0.1851372**	**0.1686089**	**0.916**
-0.60	0.55	0.901	0.1853810	0.1686902	0.918

[*] Based on model (2), the parameters were standardized as $\tau^2 = 1$, $\varphi^2 = 1$, $\zeta^2 = 1$, $\mu = 0$, and $t_j = 0$ (no time effect). Type I error was set as 0.05. Each scenario was simulated with 2,000 repetitions.

[†] Effect size can be expressed as odds ratio, which is $\exp(\theta)$.

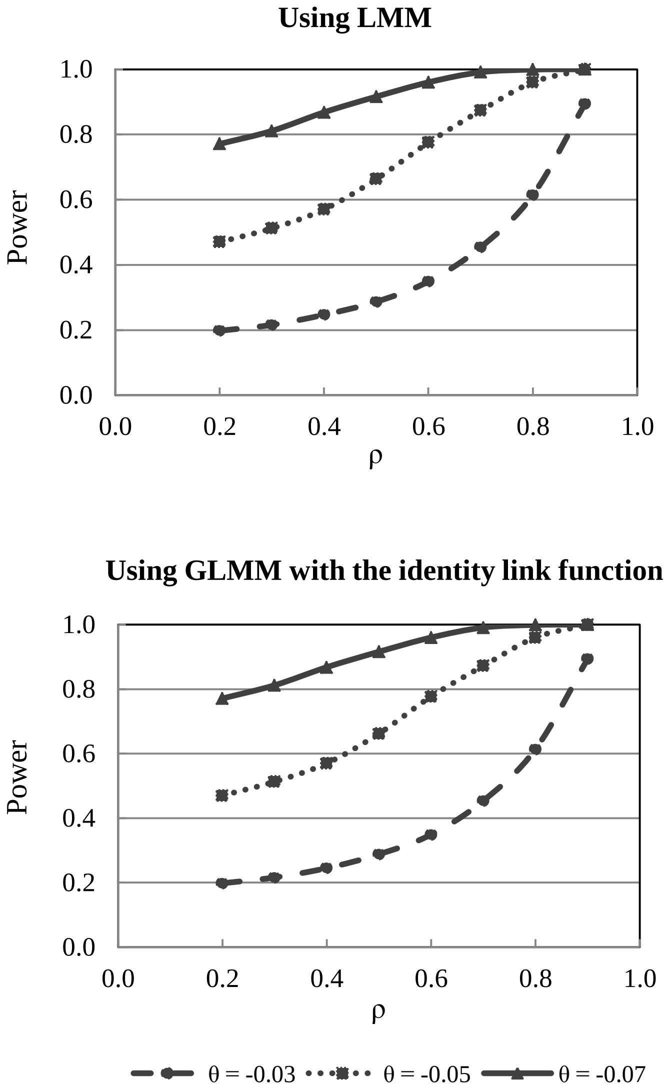
Abbreviations: GLMM, generalized linear mixed models; τ^2, between-market variance; φ^2, between-dialysis center variance; ζ^2, between-patient variance; μ, grand mean; θ, effect of concurrent supportive care; t_j, time effect.



Cluster					
C4	0	0	0	0	1
C3	0	0	0	1	1
C2	0	0	1	1	1
C1	0	1	1	1	1
Time	T1	T2	T3	T4	T5

“0” represents control periods; “1” represents intervention periods.

Figure 1. Illustration of a stepped wedge cluster design

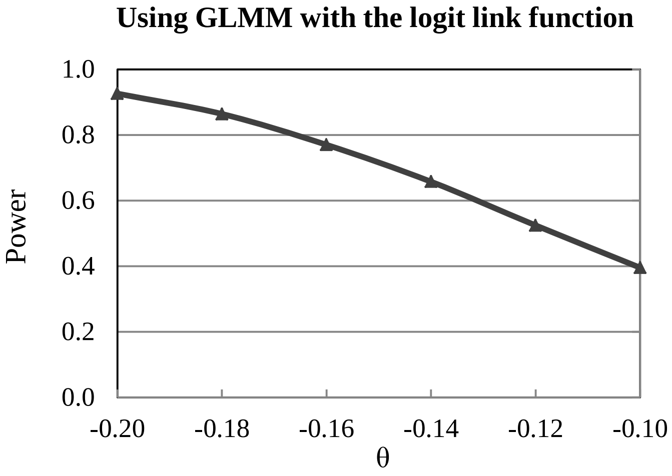


Using LMM

Using GLMM with the identity link function

Abbreviations: LMM, linear mixed models; GLMM, generalized linear mixed models; θ, intervention effect; ρ, correlation

Figure 2. Simulation results for the continuous responses to assess the power of the hypothesis test H_0: θ = 0 vs. H_a: θ = $θ_A$ where $θ_A \neq 0$



Abbreviations: GLMM, generalized linear mixed models; θ, intervention effect

Figure 3. Simulation results for the binary responses to assess the power of the hypothesis test H_0: $\theta = 0$ vs. H_a: $\theta = \theta_A$ where $\theta_A \neq 0$

BIBLIOGRAPHY

1. Meldrum, M.L., *A brief history of the randomized controlled trial: from oranges and lemons to the gold standard.* Hematology/Oncology Clinics of North America, 2000. **14**(4): p. 745-760.
2. Piantadosi, S., *Clinical trials: a methodologic perspective*. Vol. 593. 2005: John Wiley & Sons.
3. Edwards, S.J., et al., *Ethical issues in the design and conduct of cluster randomised controlled trials.* British Medical Journal, 1999. **318**(7195): p. 1407.
4. Mazor, K.M., et al., *Cluster randomized trials: opportunities and barriers identified by leaders of eight health plans.* Medical Care, 2007: p. S29-S37.
5. Brown, C.A. and R.J. Lilford, *The stepped wedge trial design: a systematic review.* BMC Medical Research Methodology, 2006. **6**(1): p. 54.
6. Mdege, N.D., M.-S. Man, and D.J. Torgerson, *Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation.* Journal of Clinical Epidemiology, 2011. **64**(9): p. 936-948.
7. Freedman, B., *Equipoise and the ethics of clinical research.* The New England Journal of Medicine, 1987.
8. Stolberg, H.O., G. Norman, and I. Trop, *Randomized controlled trials.* American Journal of Roentgenology, 2004. **183**(6): p. 1539-1544.
9. Levy, R., et al., *Multidisciplinary HIV adherence intervention: a randomized study.* AIDS Patient Care and STDs, 2004. **18**(12): p. 728-735.
10. Cook, R.F., A. Back, and J. Trudeau, *Substance abuse prevention in the workplace: recent findings and an expanded conceptual model.* Journal of Primary Prevention, 1996. **16**(3): p. 319-339.
11. Grant, A.D., et al., *Effect of routine isoniazid preventive therapy on tuberculosis incidence among HIV-infected men in South Africa.* The Journal of the American Medical Association, 2005. **293**(22): p. 2719-2725.
12. Ciliberto, M.A., et al., *Comparison of home-based therapy with ready-to-use therapeutic food with standard therapy in the treatment of malnourished Malawian children: a controlled, clinical effectiveness trial.* The American Journal of Clinical Nutrition, 2005. **81**(4): p. 864-870.
13. Group, G.H.S., *The Gambia hepatitis intervention study.* Cancer Research, 1987. **47**(21): p. 5782-5787.
14. Somerville, M., et al., *From local concern to randomized trial: the Watcombe Housing Project.* Health Expectations, 2002. **5**(2): p. 127-135.

15. Bailey, I. and L. Archer, *The impact of the introduction of treated water on aspects of community health in a rural community in Kwazulu-Natal, South Africa.* Water Science and Technology, 2004. **50**(1): p. 105-110.
16. Hussey, M.A. and J.P. Hughes, *Design and analysis of stepped wedge cluster randomized trials.* Contemporary Clinical Trials, 2007. **28**(2): p. 182-191.
17. Pearson, D., et al., *Parable of two agencies, one of which randomizes.* The ANNALS of the American Academy of Political and Social Science, 2010. **628**(1): p. 11-29.
18. Brown, C., et al., *An epistemology of patient safety research: a framework for study design and interpretation. Part 2. Study design.* Quality and Safety in Health Care, 2008. **17**(3): p. 163-169.
19. Wong, S.P., W. Kreuter, and A.M. O'Hare, *Treatment intensity at the end of life in older adults receiving long-term dialysis.* Archives of Internal Medicine, 2012. **172**(8): p. 661-663.
20. Cohen, L.M., et al., *Dying well after discontinuing the life-support treatment of dialysis.* Archives of Internal Medicine, 2000. **160**(16): p. 2513-2518.
21. Weisbord, S.D., et al., *Prevalence, severity, and importance of physical and emotional symptoms in chronic hemodialysis patients.* Journal of the American Society of Nephrology, 2005. **16**(8): p. 2487-2494.
22. O'Connor, N.R., et al., *Survival after dialysis discontinuation and hospice enrollment for ESRD.* Clinical Journal of the American Society of Nephrology, 2013. **8**(12): p. 2117-2122.
23. Viera, A.J. and J.M. Garrett, *Preliminary study of a school-based program to improve hypertension awareness in the community.* Family Medicine, 2008. **40**(4): p. 264.
24. Liddy, C., et al., *Improved delivery of cardiovascular care (IDOCC) through outreach facilitation: study protocol and implementation details of a cluster randomized controlled trial in primary care.* Implement Science, 2011. **6**: p. 110.
25. Patel, M.P., et al., *Supplemental feeding with ready-to-use therapeutic food in Malawian children at risk of malnutrition.* Journal of Health, Population and Nutrition, 2005: p. 351-357.
26. Winani, S., et al., *Use of a clean delivery kit and factors associated with cord infection and puerperal sepsis in Mwanza, Tanzania.* Journal of Midwifery & Women's Health, 2007. **52**(1): p. 37-43.
27. Bashour, H.N., et al., *The effect of training doctors in communication skills on women's satisfaction with doctor–woman relationship during labour and delivery: a stepped wedge cluster randomised trial in Damascus.* BMJ open, 2013. **3**(8).
28. Husaini, B.A. and M.C. Reece, *A church-based program on prostate cancer screening for African American men: reducing health disparities.* Ethnicity & Disease, 2008. **15**: p. 16.