POWER ANALYSIS FOR STEPPED WEDGE TRIALS WITH DELAYED TREATMENT INITIATION AND LONGITUDINAL MEASUREMENTS

by

Peng Liu

BMed, Sun Yat-sen University, China, 2012

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

2016

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Peng Liu

It was defended on

April 15, 2016

and approved by

Thesis Advisor:
Jonathan G. Yabes, PhD
Assistant Professor
Department of Medicine, Biostatistics
School of Medicine and Graduate School of Public Health
University of Pittsburgh

Committee Member:
Manisha Jhamb, MD, MPH
Assistant Professor
Department of Medicine
School of Medicine
University of Pittsburgh

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

Jonathan G. Yabes, PhD

POWER ANALYSIS FOR STEPPED WEDGE TRIALS WITH DELAYED TREATMENT INITIATION AND LONGITUDINAL MEASUREMENTS

Peng Liu, MS

University of Pittsburgh, 2016

ABSTRACT

Stepped wedge trials (SWT) is a special type of crossover cluster randomized trials (CRT) in which clusters are randomized to initiate treatment at different points in time. This design is proposed for a future trial that aims to study the effectiveness of a population health management (PHM) intervention based on Electronic Health Record (EHR) among patients with CKD. This study will collect multi-level data with longitudinal kidney function measurements from patients nested within practices. There are two challenges in powering this trial: (1) existing literature to power SWTs focus on cross-sectional design and single level of clustering; and (2) patients enrolled in the EHR-PHM trial may experience delayed treatment initiation in which they receive treatment later than expected. The effect of delayed treatment initiation to power has not been discussed in the current literature.

The goal of this thesis is to develop a simulation-based method for power and sample size analysis for SWTs with longitudinal outcomes and delayed treatment initiation. We assumed random effects mixed models to account for correlation introduced by multiple levels of clustering. Simulation parameters are acquired from preliminary EHR data and verified by nephrologists. We determined the power and sample size requirements with varying levels of delayed treatment initiation.

We have found that delayed treatment initiation results in slight loss of power. The impact of varying levels of subject delay for a fixed time delay is similar to the impact of varying levels of time delay for a fixed subject delay. Simulation-based power calculation is a flexible and practical tool in designing SWT with longitudinal measurements.

Public health significance: In clinical trials, the simulation-based power calculation method provides a practical and flexible approach for power calculation and sample size determination in designing SWTs with longitudinal outcomes while incorporating the effect of delayed treatment initiation. This method will be useful in the design of the EHR-PHM trial which could potentially improve care for and outcomes of high risk CKD patients.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to thank everyone who have helped me directly or indirectly in completion of my thesis. I would especially thank my thesis advisor Dr. Jonathan Yabes, for guiding me through with patience and encouragement. I would also like to thank my academic advisor Dr. Ada Youk and my committee member, Dr. Manisha Jhamb for their suggestions and support throughout the producing of my thesis.

I would like to thank my wife, Shan, back home for her unconditional love and support all the time.

1.0 INTRODUCTION

Chronic kidney disease (CKD) is one of the major causes of morbidity and mortality in the United States with 26 million adults having the disease and an increasing number of people are at risk [1-3]. Glomerular filtration rate (GFR) is the best indicator for classifying CKD into different stages. Low GFR indicates poor renal function. Patients with GFR<60 mL/min/1.73m2 for greater than 3 months are classified as CKD; Patients with GFR<15 mL/min/1.73m2 for >3 months are classified as end stage renal disease (ESRD) and usually treated with dialysis or kidney transplantation [3,4]. Most early stage CKD patients are provided treatment by the primary care providers (PCPs). Due to the limited knowledge and experience of the PCPs in nephrology, many patients are insufficiently educated and treated, which leads to suboptimal disease outcomes [1,5,6]. To close the gap between the generalist and nephrologist and improve the care of CKD patients at primary care settings, a study with EHR-based population health management (PHM) intervention is planned.

The proposed EHR PHM trial will be a large pragmatic trial studying the effectiveness of the PHM intervention. The study is aimed to slow down the decline of GFR by implementing EHR-based PHM intervention that includes patient education, medication review, and unsolicited electronic nephrology consultation. In contrast, usual care patients only receive care from their primary physicians with an education component. The intervention will target CKD patients with high-risk of progression to renal failure. These patients are more likely to benefit from PHM potentially leading to a more cost-effective intervention.

The EHR PHM trial will use a Stepped Wedge Design to assess the intervention effect. Stepped wedge trials (SWTs) are studies where clusters randomly and sequentially roll-out to initiate the intervention over multiple time points. One cluster or multiple clusters can be randomized to initiate the intervention at a time point. The roll-out process is unidirectional, which means clusters can only cross-over from control to intervention. At the end of the trial, there will be a time period in which all clusters have initiated intervention. Randomization will be at the practice level because individual randomization raises concerns about implementation and cross-contamination.

Traditional parallel cluster randomized trials (CRT) can also be used to assess the intervention effect when the randomization and implementation of the intervention are at the cluster level. However, parallel CRTs have some limitations when applied to EHR PHM study. One limitation is that practices initiate intervention at the baseline, which is not practical in some scenarios. Another limitation is that in parallel CRTs, practices stay in either control or intervention group for the whole duration of the study. Whereas we want all practices to be exposed to intervention by the end of the study because of the high expectation that the intervention will work. In this way, the likelihood of practices participation and engagement will also increase.

The power calculation approach and the sample size formula have been provided for SWTs with one level of clustering and cross-sectional outcomes [7, 16]. Woertman and colleagues (2013) have shown that cross-sectional SWTs are more efficient than CRTs in terms of sample size [16]. In the more complicated case where longitudinal measurements are obtained from each individual and individuals are nested within the clusters, there is much less literature discussing the power and sample size calculation for SWTs.

Delayed treatment effect occurs when the length of time interval between two transition time points is not long enough to allow the full treatment effect to be observed. The full treatment

effect may not be fully realized for several time intervals. Delayed treatment effect has been shown to negatively impact the power of SWTs [7].

Delayed treatment initiation arises from the recruiting process. During the time of the study, we continuously recruit new patients. Because of the recruitment strategy, patients do not enter the study exactly at the time when the first visit is scheduled. There is a lag between when the patient is expected to be exposed to intervention and the time patients actually receive intervention. This is called delayed treatment initiation, which is different from delayed treatment effect.

The EHR PHM trial has two unique challenges: (1) the outcome is longitudinal; (2) there is a delayed treatment initiation.

In this study, we will examine impact of delayed treatment initiation on the sample size and power for SWT with longitudinal outcomes. We will focus on studies with continuous outcomes but our approach can be easily extended to categorical outcomes including the binary case. In section 2.1, we review traditional designs for randomized trials. Section 2.2 covers the design, analysis and power of stepped wedge trial. In section 2.3, we focus on the stepped wedge trial with longitudinal measurements and outline the simulation procedure for this type of design. In section 3, we describe the implementation of simulation-based power analysis on the EHR PHM Trial.

2.0 METHODS

2.1 REVIEW OF TRADTIONAL DESIGNS

2.1.1 Individually Randomized Trials

Randomized controlled trials (RCTs) are studies in which participants are randomly allocated to intervention or control arm. They are considered the "gold standard" of study designs and provide the strongest evidence in evaluating the effect of an intervention on the response [1]. Randomization provides a mechanism to minimize bias in allocating subjects to interventions and control for confounding caused by observed and unobserved factors [2]. Although randomizing at the individual level is the most prevalent RCT design, it can be impractical or infeasible to allocate individual patients to the intervention arms. For example, evaluating the effectiveness of a new vaccine for typhoid fever in India, selected and randomized geographically clustered communities rather than individual persons [23]. Had this study been individually randomized, conducting the trial would have been a lot more expensive because distribution of the vaccine needs to span all communities rather few selected ones. In addition, health centers are usually located within a community that all members have potential access too, increasing the likelihood of cross-contamination, in which subjects in one arm learn and adopt the treatment of subjects in the other arm.

2.1.2 Cluster Randomized Trials

CRTs are a special type of RCT in which natural clusters (e.g. clinical practices, communities) instead of individuals are randomized to the intervention group or control group. CRTs are suitable for use in several scenarios, such as natural group-based interventions or when individualized RCT cannot be applied for ethical, administrative or feasibility reasons [21, 22]. This design minimizes cross-contamination caused by the allocation of different treatment arms to subjects who can potentially interact with each other because they belong to the same cluster.

Parallel arm CRTs are the most commonly used design to assess the effectiveness of interventions that are delivered at cluster level. In this design, treatment is assigned to clusters at baseline, and all selected individuals in the cluster receive the treatment assigned. Clusters remain in the assigned group until the end of the study. Subjects in the two groups are followed simultaneously and measurements are taken concurrently.

Although this parallel CRT is simple and easy to implement, it has some limitations. In the case where there is compelling evidence that the intervention is effective, it is potentially unethical to withhold the intervention from the control patients. Moreover, the risk of low participation rate may be higher because patients are unwilling to be randomized to the control group. If they get randomized and the intervention is unblinded, a higher drop-out rate may occur due to loss of interest of patients randomized to the control. In the case where implementation of the intervention requires staggered roll-out because of logistical or practical concerns, it would be impractical to utilize parallel CRT. A variant of CRT called Stepped wedge trials address these limitations.

2.2 STEPPED WEDGE TRIALS

	Time period			
Clusters	Period 1	Period 2	Period 3	Period 4
1	0	1	1	1
2	0	0	1	1
3	0	0	0	1

0 and blank cells represent control time period, 1 and shaded cells represent intervention time period

Figure 1. A Stepped wedged trial with 3 clusters and 3 transition time points

A stepped wedge trial design is a special case of CRT [7-11] in which clusters crossover in one direction from control to intervention during the study period. In SWT, all clusters start in the control group, and the intervention is sequentially rolled-out at pre-specified cross-over time points or "steps". One cluster or a group of clusters may initiate intervention at a time point. The time at which clusters cross-over to the intervention is assigned randomly. The cross-over process is unidirectional, which means once a cluster turns on the intervention, it will be exposed to intervention until the end of the study. All clusters would be in the intervention arm by the end of the study. Figure 1 illustrates a stepped wedge design with 3 clusters and 3 transition points.

Compared to parallel CRTs, SWTs have some unique features [7, 8, 10, 17]. 1) SWTs allows intervention to be implemented sequentially. This feature makes SWTs especially useful when intervention cannot be delivered simultaneously because of limited resources or logistical reasons. 2) At the last step, all clusters would have received the intervention. This feature will help alleviate ethical issues when the benefit of the intervention is clear. 3) SWT may require less sample size compared to parallel CRT to achieve the same power [16]. This is because we have

more measurements and the design allows both cross-cluster comparison and within-cluster comparison. Some argued that the improvement of power is merely because of the repeated measurements, not from the design of the study [17, 20]. In their example, the power advantage of SWT is case-dependent. When the intra-cluster correlation (ICC) is small, the power advantage becomes minor.

SWTs have disadvantages as well [17]. SWTs usually require much longer time than CRTs. In the cases where time-varying confounding factors are present, it becomes more difficult to find the true intervention effectiveness. Furthermore, the relatively more complex study design and longer study time will impose challenges on study implementation, for example, patients and researchers may not be able to afford the increased cost and resources to collect data. In addition, when the effect of the intervention is not clear, it is not safe for everyone to be exposed to intervention at the end of the study.

Considering the pros and cons, SWT is recommended under the following conditions: 1) When the intervention has been proven to be effective in individual level studies, and we want to test the effectiveness of the population level. 2) When the intervention and data collection can be integrated into the routine medical services, and does not need patients' extra participation. 3) When the ICC is expected to be high.

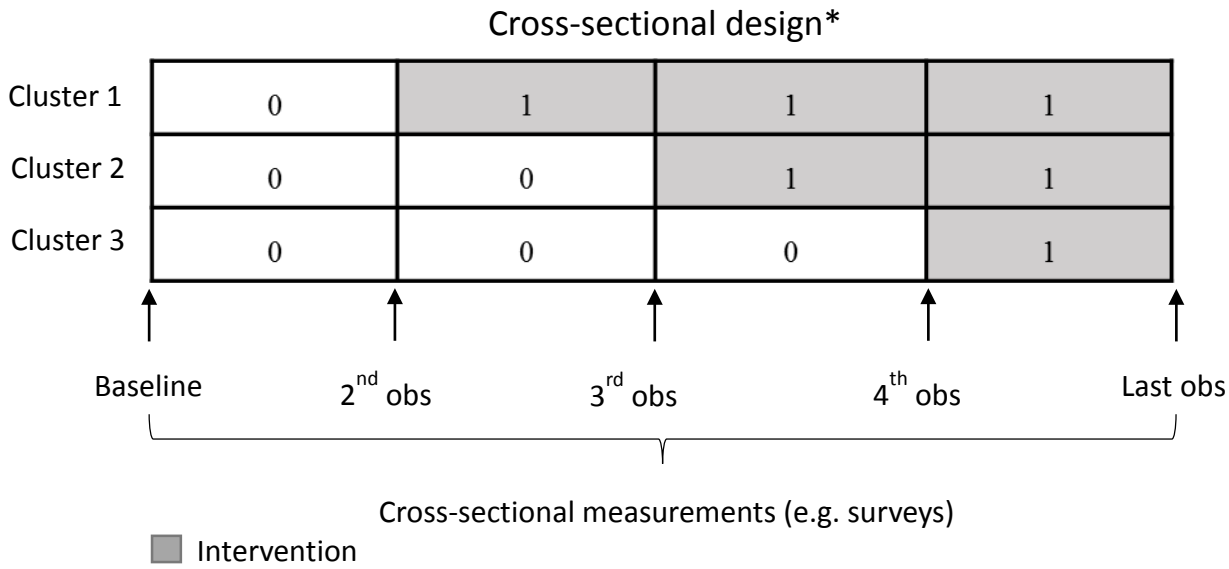
2.2.1 Design of SWTs

Cross-sectional and cohort are the two common types of designs for SWTs [14]. In a cross-sectional design, a snapshot of the outcome across groups is taken at each of the cross-over time point. Observations obtained at different time points are from different subjects. The subjects are randomly sampled from the continuously changing population in the clusters. Outcome can be an

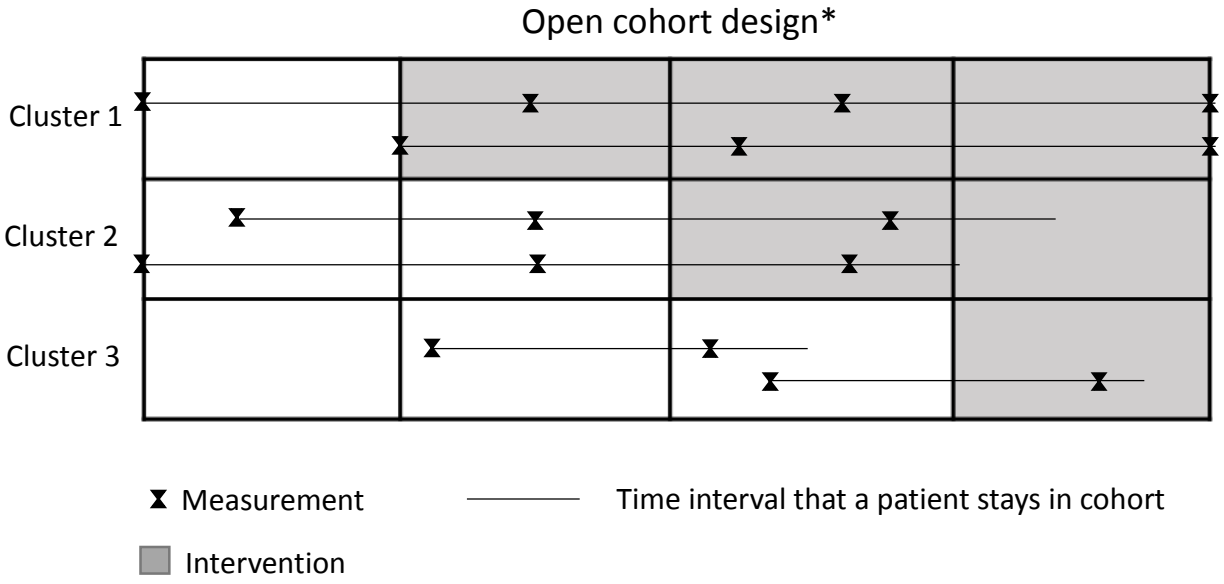
event aggregate at the cluster level. For example, the outcome can be the incidence or prevalence of a disease. Figure 2a shows a SWT with a cross sectional design. In a cohort design, on the other hand, repeated observations are collected for every individual over time. This allows analysis of the intervention effect at both cluster and individual levels. The cross-sectional design is suitable for studying the population-level intervention effect [15]. Meanwhile, the cohort design has an advantage in establishing time relationship between the outcome and intervention for individuals. A drawback however is an increased risk of loss to follow-up particularly when the observation time period is much longer than CRT [15].

Copas and colleagues (2015) have summarized three major types of cohorts: (1) open cohort; (2) closed cohort; and (3) cohort with continuous recruitment and short exposure. In an open cohort, patients are recruited or drop-out during the course of the study as shown in Figure 2b. In contrast, a closed cohort involves a fixed population as shown in Figure 2c [14]. In both open and closed cohort, individuals' intervention status changes at the roll-out time points. In continuous recruitment and short exposure designs, there is ongoing recruitment during the study and patients will receive a short period of intervention. Each patient is followed and have one or more repeated measurements over time. Individual's intervention status is determined at the entry of the trial and does not change over time. For example, if an individual enrolled in a cluster that had already transitioned to intervention, then that individual would belong to the intervention group. Figure2-d shows a SWT with continuous recruitment and short exposure design where repeated measurements are taken for each patients until the end of the study.

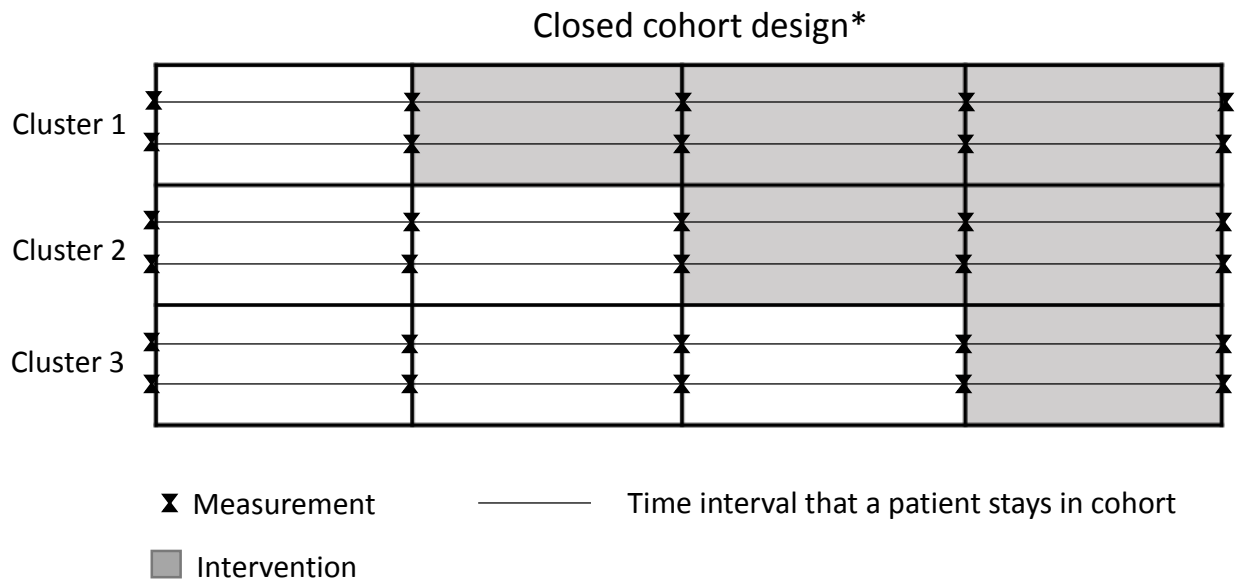
a)


Cross-sectional design*
Cluster 1
Cluster 2
Cluster 3
0
1
1
1
0
0
1
1
0
0
0
1
Baseline
2nd obs
3rd obs
4th obs
Last obs
Cross-sectional measurements (e.g. surveys)
Intervention

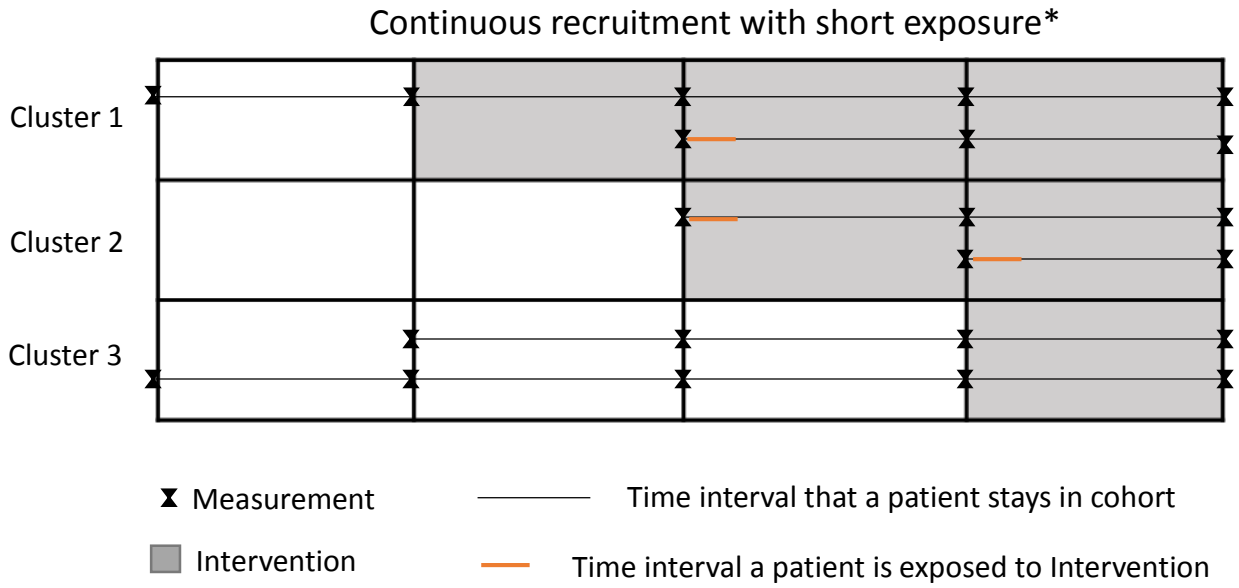
b)


Open cohort design*
Cluster 1
Cluster 2
Cluster 3
Measurement
Time interval that a patient stays in cohort
Intervention

c)



d)



* The above figures illustrate different types of SWTs with 3 clusters and 3 transition time points. The vertical lines represent the time points, including 3 transition time points, one baseline time point and the last time point.

Figure 2. Different designs of SWT

2.2.2 Analysis of SWTs with Cross-sectional Data

Observations derived from an SWT are naturally correlated within clusters. Analysis then requires accounting for the hierarchical nature of the design. Linear mixed models (LMM) are used for analyzing SWTs with continuous outcome. For example, in the SWT aimed to assess the effect of a strategy to address malnourished hospitalized adults, the primary outcome was the daily energy and protein intake [24]. Generalized linear mixed models (GLMM) were used for analyzing SWTs with binary, ordinal or count outcomes. For example, the enhanced peri-operative care for high-risk patients (EPOCH) trial is a SWT with cross-sectional design studying the effect of a healthcare quality improvement intervention. The primary outcome is the 90-day mortality of patients undergoing emergency laparotomy [25]. If we are interested in the marginal or population average effect, generalized estimating equations (GEE) could be an alternative for SWTs analysis.

2.2.3 Sample Size and Power of SWTs with Cross-sectional Data

For power calculation, a method for cross-sectional SWT data based on the Wald test has been described in the literature [7, 11]. This method assumes a SWT with K clusters, J time points and N subjects sampled from each cluster at each time points. The trial is aimed to study whether treatment will improve the outcome. For the Gaussian distributed outcome, the method is based on a random effects mixed model given by:

$$Y_{jk} = \mu + \beta_1 T_j + \theta X_{jk} + \zeta_k + e_{jk} \quad (1)$$

where $k=1,2\ldots K$, $j=1,2,\ldots J$, Y_{jk} denotes the mean response for kth cluster at jth time point, the cluster random effect $\zeta_k \sim N(0, \tau^2)$ accounts for the variation between clusters and the term $e_{jk} \sim$

$N(0, \sigma_e^2)$ accounts for the variation of the mean outcome within clusters. T_j is the jth crossover time and X_{ik} is the treatment status for cluster k at time T_j.

Let θ be the treatment effect. We wish to test the hypothesis H_0: $\theta=0$ versus H_a: $\theta\neq0$. The power formula is given by

$$Power= \Phi\left(\frac{\hat{\theta}}{\sqrt{Var(\hat{\theta})}} - Z_{1-\alpha/2}\right) \quad (2)$$

Where Φ is the cumulative standard normal distribution and $Z_{1-\alpha/2}$ is the (1-α/2)-th quantile. $\hat{\theta}$ and are, respectively, the point estimate and variance of the intervention effect. This formula for power is derived based on the following assumptions: 1) The SWT study design is cross-sectional, 2) There is one level of clustering, and 3) there is no time by treatment interaction.

For sample size calculation, Woertman and colleagues (2013) proposed a closed formula based on Hussey and Hughes' model and assumptions [11, 16]. This method starts with the sample size that would be needed under an individually randomized RCT design (N_u). The design effect of SWTs (denoted as DE_{sw} is derived). The sample size of SWTs can be obtained by multiplying the N_u with design effect DE_{sw} to adjust for the extra complexity of design and clustering.

Hemming and colleagues (2015) provided a conservative approach to estimate the power for multiple levels of clustering [8, 18]. In a simulation experiment, they showed that SWT is robust to the mis-specification of the hierarchical structure (e.g., 2-levels versus 3-levels of clustering) compared to CRT. They suggested a conservative approach by assuming only one level of clustering to simplify the problem.

2.3 STEPPED WEDGE TRIALS WITH LONGITUDINAL OUTCOMES

It is not uncommon to have a stepped wedge trial with a longitudinal data collection scheme for each individual within a cluster. For example, the EHR population management trial will utilize an open cohort with continuous recruitment and short exposure design. Each patient within each practice will be recruited continuously until the end of study period. Their GFR measurements will be collected at baseline (i.e., their enrollment day) and every 6 months thereafter to monitor decline in renal function. In these types of data, we will have observations across time points nested within a patient, nested within a practice. To account for the correlation of measurements within a patient within a practice, we can use a multi-level linear mixed model with random patient and random practice intercept.

The treatment assigned to a patient is determined by whether or not the practice caring for him/her has already transitioned to the intervention at the time the patient is enrolled. For example, if the patient was enrolled at the very beginning of the study, then that patient will be in the control group because all practices are in the control phase. On the other hand, if a patient was recruited after the last transition point, then that patient will be in the treatment group because all practices would have already transitioned to the intervention. Note that although the intervention status of a practice changes over time, the intervention group of a patient does not change over time.

In the next section and thereafter, we will focus on trials with continuous and normal outcomes. However, we can take the generalized linear mixed models approach to extend the methods to non-normal outcomes such as binary or count data.

2.3.1 Analysis and Model

We can model the mean outcome μ_{ijk} at the i-th time point T_{ijk} of a patient j in practice k whose treatment group is X_{jk} using a generalized linear mixed model formulation as:

$$g(\mu_{ijk}) = \beta_0 + \beta_1 T_{ijk} + \beta_2 X_{jk} + \theta(T_{ijk} X_{jk}) + \zeta_{j(k)} + \zeta_k \quad (3)$$

where $g(.)$ is the link function, β_0 is the overall intercept, β_1 is the effect per unit time, β_2 represents the difference between intervention and control group at baseline, and θ is the treatment effect per unit time. The random effects $\zeta_k \sim N(0, \tau_1^2)$ is the level 3 random intercept (practice specific random intercept), $\zeta_{j(k)} \sim N(0, \tau_2^2)$ is the level 2 random intercept (patient-specific random intercept). For continuous outcomes, we assume the error term follows normal distribution and can use a linear mixed model formulation given by

$$g(\mu_{ijk}) = \mu_{ijk}, \text{ and } Y_{ijk} = \mu_{ijk} + e_{ijk},$$

where Y_{ijk} is the response value for the i^{th} time point of the j^{th} patient in the k^{th} practice, $e_{ijk} \sim N(0, \sigma_e^2)$ is the error term assumed to be independent of the patient-specific and practice specific random intercepts.

We are interested in the intervention effect over time θ, adjusted for time effect β_1 and baseline outcome differences between intervention patients and control patients β_2. In the EHR PHM trial, CKD patients generally have GFR that declines over time. It is also possible that GFR between control and intervention patients are different at baseline because randomization was not performed at the individual level. Therefore, we need to adjust for time as well as baseline GFR in the analysis of intervention effect [10]. In this case, β_1 is the rate of GFR decline over time of control patients, β_2 is the GFR mean difference between control and intervention patients at

baseline (time 0), and θ is the difference in the rate of GFR decline between control and intervention patients.

In our model, there are two levels of clustering, repeated measurements over time (level 1) that are clustered in patients (level 2), and patients nested within practice (level 3). Without loss of generality, suppose we have balanced data in which there are *I* repeated measurements for each of the *J* patients within *K* practices, we can write our model for the outcome in matrix form as:

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\alpha}+\mathbf{Z}\boldsymbol{\gamma}+\mathbf{e} \tag{5}$$

where **Y** is a ($I*J*K \times 1$) outcome vector, **X** ($I*J*K \times 4$) is the design matrix for the fixed effects, $\boldsymbol{\alpha}=\begin{bmatrix}\beta_0 & \beta_1 & \beta_2 & \theta\end{bmatrix}^T$ is a 4 × 1 vector of fixed effects parameters, **Z** is the design matrix for the random effects, $\boldsymbol{\gamma} = \begin{bmatrix}\zeta_1 & \cdots & \zeta_K & \zeta_{11} & \cdots & \zeta_{KJ}\end{bmatrix}$ is a ($K+K*J$) × 1 vector of parameters for the random effects, and **e** is the vector of the random errors. The **Z** matrix have $I*J*K$ rows and $K*J+K$ columns. The first *K* columns are the random parameters for the *K* practices, and the remaining $J*K$ columns are the random parameters for all the $J*K$ patients in the practices. That is:

$$\mathbf{Z} = \begin{bmatrix} 1_{IJ} & 0_{IJ} & \cdots & 0_{IJ} & Z_1 & 0_{IJ\times J} & \cdots & 0_{IJ\times J} \\ 0_{IJ} & 1_{IJ} & \cdots & 0_{IJ} & 0_{IJ\times J} & Z_2 & \cdots & 0_{IJ\times J} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{IJ} & 0_{IJ} & \cdots & 1_{IJ} & 0_{IJ\times J} & 0_{IJ\times J} & \cdots & Z_K \end{bmatrix}_{IJK\times(KJ+K)}$$

where 1_{IJ} is a $I*J \times 1$ vector where all entries equal to 1, 0_{IJ} is a $I*J \times 1$ vector where all entries equal to 0

$$\mathrm{Z_k} = \begin{bmatrix} 1_I & 0_I & 0_I & 0_I \\ 0_I & 1_I & 0_I & 0_I \\ 0_I & 0_I & \ddots & 0_I \\ 0_I & 0_I & 0_I & 1_I \end{bmatrix}_{IJ\times J}$$

The covariance structure of our model is given by:

$$\mathbf{V} = \mathrm{var}(\mathbf{Y}) = \mathrm{var}(\mathbf{X\alpha}+\mathbf{Z\gamma}+\mathbf{e})$$

$$= \mathrm{var}(\mathbf{Z\gamma}) + \mathrm{var}(\mathbf{e})$$

$$= \mathbf{ZGZ} + \mathbf{R} \qquad (6)$$

G matrix is a diagonal matrix with *K* elements of τ_1^2 and *KJ* elements of τ_2^2

$$\mathrm{G} = \begin{bmatrix} \tau_1^2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \tau_1^2 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \tau_2^2 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & \tau_2^2 \end{bmatrix}_{(K+KJ)\times(K+KJ)}$$

The variance matrix of the random error is

$\mathbf{R} = diag(\sigma_e^2)_{IJK\times IJK}$.

Therefore, the variance-covariance matrix of our model is

$$\mathbf{V} = \begin{bmatrix} V_1 & 0_{IJ\times IJ} & \cdots & & 0_{IJ\times IJ} \\ 0_{IJ\times IJ} & V_2 & & & \vdots \\ \vdots & & \ddots & & \\ & & & & 0_{IJ\times IJ} \\ 0_{IJ\times IJ} & \cdots & & 0_{IJ\times IJ} & V_K \end{bmatrix}_{IJK\times IJK}$$

where $\mathbf{V}_\mathrm{k}$ is a matrix denoting the variance-covariance structure within each practice. Different practices are not correlated. The covariance pattern within each practice is

$$\mathrm{V_k} = \begin{bmatrix} U_1 & \tau_1^2 J_{IxI\backslash} & \cdots & & \tau_1^2 J_{IxI} \\ \tau_1^2 J_{IxI} & U_2 & & & \vdots \\ & & \ddots & & \\ \vdots & & & & \tau_1^2 J_{IxI} \\ \tau_1^2 J_{IxI} & \cdots & & \tau_1^2 J_{IxI} & U_J \end{bmatrix}_{IJ\times IJ}$$

where J_{IxI} denote the $I \times I$ matrix with all entries equals to 1. As the random effect of the practice

is $\zeta_k \sim N(0, \tau_1^2)$, the covariance between different patients within the same practice is τ_1^2. **U**j is the covariance pattern within each patient,

$$\mathbf{U}_j=\begin{bmatrix} \tau_1^2+\tau_2^2+\sigma_e^2 & \tau_1^2+\tau_2^2 & \cdots & \tau_1^2+\tau_2^2 \\ \tau_1^2+\tau_2^2 & \tau_1^2+\tau_2^2+\sigma_e^2 & \cdots & \tau_1^2+\tau_2^2 \\ \vdots & \vdots & \ddots & \vdots \\ \tau_1^2+\tau_2^2 & \tau_1^2+\tau_2^2 & \cdots & \tau_1^2+\tau_2^2+\sigma_e^2 \end{bmatrix}_{I\times I}$$

Notice that the covariance between two observations within each patient is $\tau_1^2+\tau_2^2$, which accounts for both the covariance within practice and covariance within patient. The variance of each observation is $\tau_1^2+\tau_2^2+\sigma_e^2$, where the additional component is the variance of the random error. Our model assumes compound symmetric covariance pattern between repeated measures in each patient.

The hierarchical structure of the data along with the covariance structure above leads to two ICCs: (1) the correlation at the cluster level, where observations from different subjects but within the same cluster are correlated, is as follows:

$$\rho_a = \frac{\tau_1^2}{\tau_1^2+\tau_2^2+\sigma_e^2}$$

(2) the correlation at the patient level, where different observations within the same subject in the same practice are correlated, is as follows:

$$\rho_b = \frac{\tau_1^2+\tau_2^2}{\tau_1^2+\tau_2^2+\sigma_e^2}$$

2.3.2 Sample size and Power

Closed formulas for power and sample size of SWTs with cross-sectional data collection have been described in the literature (see Section 2.2.3). Unfortunately, none of these methods directly apply to multiple levels of clustering, including the longitudinal design in which measurements are repeated within a subject nested within a cluster.

There are three main obstacles in applying these formulas to our model: a) the data for our model is longitudinal, while the formulas are derived based on the assumption that the outcome is cross-sectional; b) there are two levels of clustering in our model while these formulas only assume a single level of clustering; c) the formulas assume that the time effect is constant over treatment and control (no time by intervention interaction) [11].

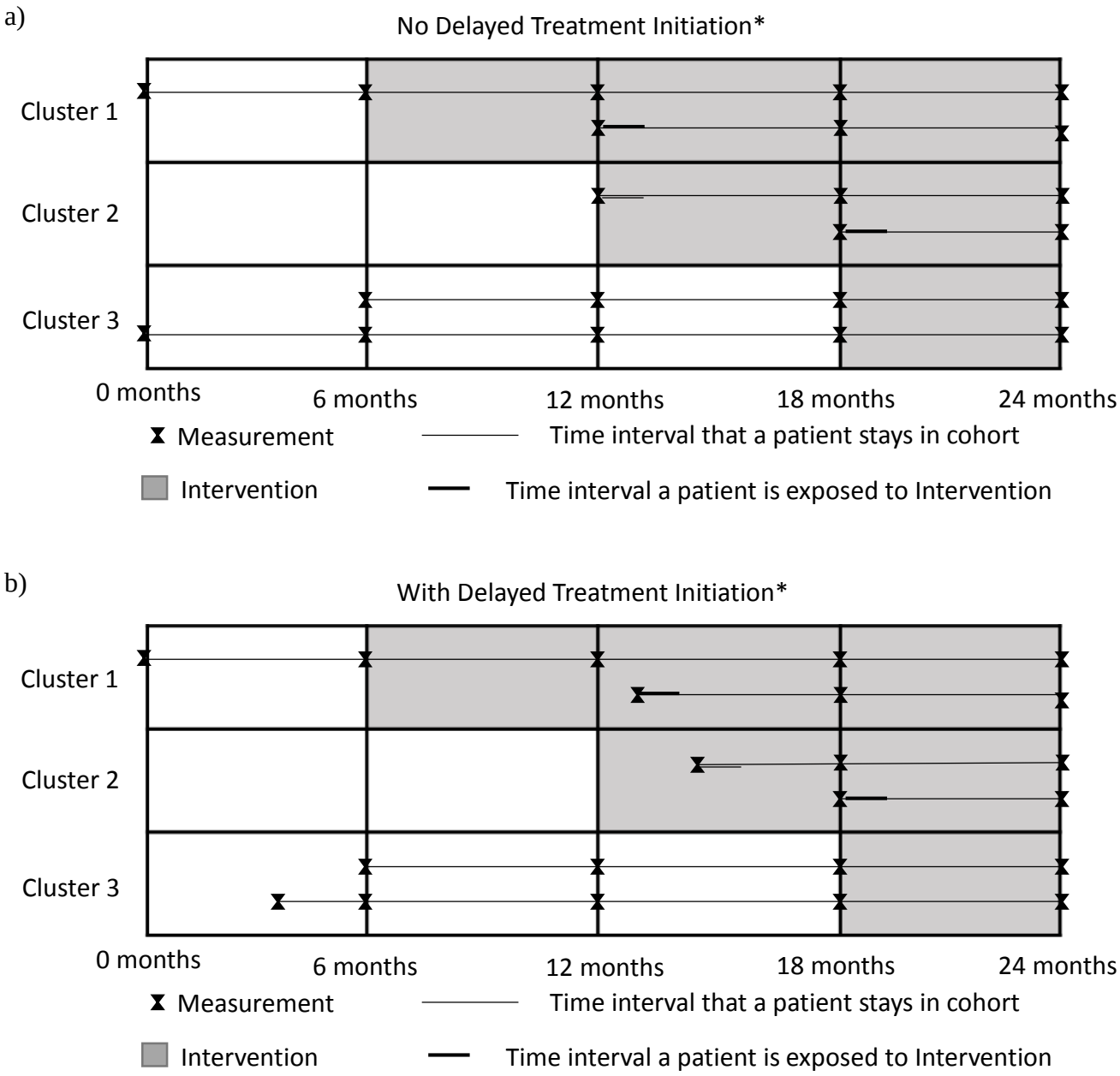
To address the limitations of these analytical methods, we instead use simulations to determine sample size and power for SWTs with longitudinal outcomes. This approach, although computationally intensive, offers flexibility and is amenable to the various kinds of study designs [11, 14]. Despite the computational demand, the lack of analytical formulas afford simulations as a practical and efficient approach for power calculations.

2.3.3 Delayed Treatment Effect and Delayed Treatment Initiation

Ideally, patients' treatment effect will be fully realized within the time interval between two transition time points in a SWT. However, in practice, the treatment effect may not reach its full impact until several time intervals later. This phenomenon is called *delayed treatment effect*. For cross-sectional data, delayed treatment effect has been discussed by Hussey and Hughes [7]. By simulation, they showed that delayed treatment effect decreases the power of the study. Extending the time interval long enough to enable the treatment effect to be fully realized is a possible solution to alleviate the delayed treatment effect. However, the cost for data collection and loss to follow-up will increase that may lead to a higher risk of subject noncompliance.

Another factor that may influence power relates to when the patient actually received the treatment. In the ideal setting, patients consenting to participate in the study would have their first clinic visit scheduled exactly at one of the roll-out time points or baseline. This would allow them to receive the treatment immediately after the practice transitioned to intervention. In reality, however, patients may show up later than their expected visit times. This delays their actual receipt of treatment (or control condition). This is called *delayed treatment initiation* as illustrated by Figure 3. Patients with delayed treatment initiation have less follow-up time compared to the non-delayed patients.

Delayed treatment effect and delayed treatment initiation are related. Both of them defer the influence of intervention on the patients' outcome, but in different ways. The effect of delayed treatment initiation on the power has not been studied before. We will use simulation to study the effect of delayed treatment initiation.



* The figures above illustrate SWTs with 3 clusters and 3 transition time points. The vertical lines represent the time points, including 3 transition time points, one baseline time point and the last time point.

Figure 3. Delayed treatment initiation

2.3.4 Simulation process for SWTs with longitudinal outcomes

The rationale for the simulation is that by building up a data generating process, we will be able to mimic the real-world data sufficiently many times. To calculate power, we fit the specified model based on the datasets we have generated and determine the proportion of times that the model correctly rejects the null hypothesis. The simulation procedure will be performed as follows:

A. No delayed treatment initiation

1. Specify the number of practices, and the number of patients in each practice. Choose a suitable number of observations for each patient.
2. Set the parameters based on preliminary data or previous studies. The parameters include:
    a. The overall mean GFR β_0
    b. Time effect β_1
    c. Baseline difference between intervention and control β_2
    d. Intervention effect θ
    e. The level-3 random effect variance τ_1^2 (practice specific random intercept)
    f. The level-2 random effect variance τ_2^2 (patient specific random intercept)
    g. The within patient variance or random error term σ_e^2.
3. Generate a dataset based on the study design and the model specified in Equation (3).
    a. Randomly select practices at each step to initiate intervention.
    b. Simulate a value for each of the practice specific random effect $\zeta_1, \zeta_2 \ldots \zeta_K$ and each of the patient specific random effect $\zeta_{1(k)}, \zeta_{2(k)} \ldots \zeta_{J(k)}$. Simulate a value for each of the observation's random error.

c. Randomly select an entry time for each patient from the set of transition time points including baseline, i.e., $\{t_0 = 0, t_1, \ldots, t_P\}$ where P is the total number of transition points. Generate the time of follow-up measurements according to the patient entry time and the specified measurement schedule until the end of study .

d. Under a cohort with continuous enrollment and short-exposure SWT, the intervention status of each patient does not change over time and is defined by the intervention status of the practice at the time of patient entry. Under a closed-cohort design in which each patient may transition to intervention as the practice transitions to the intervention, the intervention status of each patient is determined by the intervention status of the practice at the time of measurement and hence may vary across the measurement occasions.

e. Calculate the linear predictor using the pre-specified model parameters and the simulated intervention status and measurement times. Use this to generate the outcome value. If the outcome is continuous and follows a normal distribution, add the random error to obtain the outcome. For binary outcome, we can randomly generate the data using the Bernoulli distribution with probability equal to the inverse logit of the linear predictor.

4. Fit the three level random effects model Equation (3) and record the p-value of the intervention effect θ.

5. Repeat the steps 3 and 4 sufficiently many times. Set the type I error rate to be α. Power is calculated as the proportion of times that the null hypothesis was rejected (p-value $\leq \alpha$) across all of the simulated datasets.

B. With delayed treatment initiation

When there is delayed treatment initiation, the first clinic visit of patients may not exactly be at the start of study or the transition time points. There are two delay parameters that may impact power:

a. the proportion of patients delayed, p_d;

b. the length of time delay, l_d.

A patient j in practice k with delayed treatment initiation has entry time $T_{1jk} + l_d$ rather than T_{1jk}. SWT with no delay is a special case in which either $p_d = 0$ (no patient is delayed) or $l_d = 0$ (no time delay). If all patients are delayed, then $p_d = 1$. In this study, we assume the maximum time delay is equal to the length of time between any two transition points so that the range of potential delay is the same across all patients regardless of entry time. For simplicity of terminologies, we will refer to p_d as ***subject delay*** and l_d as ***time delay***.

To simulate data with delayed treatment initiation, we will follow the same data generation procedure above with a modified Step 3c. That is, after randomly selecting the entry time T_{1jk} for each patient j in practice k from the set of transition time points, $\{t_0 = 0, t_1, \ldots, t_P\}$, we will randomly pick a proportion p_d of subjects out of the total number of subjects. Observed entry times for these subjects will be set to $T_{1jk} + l_d$. The follow-up time points will be the same as those without delay. The model fitted and power calculation in Steps 4 and 5 will be the same as the no delayed treatment initiation case.

3.0 APPLICATION: EHR POPULATION MANAGEMENT TRIAL

3.1 STUDY DESIGN OVERVIEW

The EHR Population management trial is a stepped wedge randomized trial with continuous recruitment and short exposure. The primary outcome is GFR decline. The goal is to evaluate whether a population management intervention significantly slows down GFR decline compared to usual care. There are two levels of clusters: practices and patients. We simplified the problem by assuming that 1) we have a balanced design (i.e., the same number of patients in each practice and the same number of observations for each patient), 2) the same number of practices roll-out at each cross-over time point, and 3) measurements are equally spaced, which means the length of time period between two successive measurements are the same. Practices will be grouped into three, and each group will randomly roll-out to initiate intervention at one of the three time points: 5, 10 or 15 months after the beginning of the study. These transition points were chosen so that 24 month follow-up for all patients will be completed within 39 months from the study initiation. When a practice is in the control phase, patients who enroll in that practice will only be exposed to usual care. When a practice has already transitioned to the treatment phase, patients who enroll in that practice will be exposed to the PHM intervention for a short period of time. A baseline measurement will be taken from each patient at the time of recruitment. After the patients have been recruited, four follow-up GFR measurements will be collected for each patient at 6, 12, 18 and 24 months after the baseline measurement. Since this study will be conducted in a real clinical setting, we expect that the actual time a patient enrolls in the study may not coincide with the transition time points. Thus, a practice may have already transitioned to intervention, but may not

enroll a patient to be in the intervention until a few months later. In this case, delayed treatment initiation will happen. In this study, we assume that the maximum time that a patient could be delayed is 5 months, the length of time between two successive transition time points.

3.2 SIMULATION BASED POWER CALCULATION

To assess power and arrive with a sample size estimate needed for the EHR PM trial, we conducted simulations following the procedures described in Section 2.3.4. We simulated 5,000 datasets based on the random effects model Equation (3) which accounts for the natural clustering of observations within patients in each practice. The parameters were set to values based on preliminary data. However because we do not have preliminary estimate of the between-practice variability τ_1^2, we instead specified the within practice ICC to be 0.2 based on our experience from other studies. Table 1 shows the parameter values used in the simulations.

For each dataset, the model parameters were estimated by fitting the model with the same fixed effects and random effects specification as the data generating mechanism. The lme command from package nlme in R via REML was used for model fitting. The estimated treatment effect p-values were recorded and assessed for statistical significance at $\alpha = 0.05$ level. Power was estimated as the proportion of rejections among the 5,000 datasets. We assumed a balanced design, and varied the total number of practices from 9 to 36, and the practice size (i.e., number of patients per practice) from 5 to 20.

To examine the impact of delayed treatment initiation, we calculated power under combinations of varying degrees of subject delay and time delay. The percent of subjects delayed, p_d*100%, was set to 0%, 25%, 50%, 75%, 100% representing, respectively, none, mild, moderate,

major, and severe subject delay. The length of time delay l_d was set to 0%, 25%, 50%, 75%, 100% of the maximum time of delay (5 months) representing, respectively, none, mild, moderate, major, and severe time delay. For instance, a mild and major time delay is equivalent to a 1.25 and 3.75 month delayed treatment initiation, respectively.

Table 1. Parameters for data generation

Notation	Meaning	Value
σ_e^2	Variance of residual	*45.14*
τ_2^2	variance of subject specific random intercepts	*120.43*
ρ_a	ICC with respect to within practice correlation $\rho_a = \frac{\tau_1^2}{\tau_1^2 + \tau_2^2 + \sigma_e^2}$	*0.2*
β_0	Overall mean GFR at baseline	*46.45* mL/min/1.73 m^2
β_1	The rate of GFR decrease per month in control group	*0.49* mL/min/1.73 m^2/month
β_2	the GFR difference in treatment and control at baseline	*0*
θ	GFR decline rate in intervention group – GFR decline rate in control group	-*0.125* mL/min/1.73 m^2/month

3.3 RESULTS

The first column of Table 2 shows the power of SWTs without delayed treatment initiation. As expected, increasing the sample size (either the number of practices or the practice size) increases the power of the study. The power varies from 0.240 to 0.997 as the total sample size varies from 45 to 720. The rest of the columns in Table 2 display power as the severity of time delay increases, fixing the subject delay to 50%. Overall, power decreases with delay severity but only slightly. The decrease in power from no delay to major delay ranged from 0.002 to 0.095.

In planning the EHR PM trial, we are interested in determining the sample size required to achieve adequate power. We can also use Table 2 for this purpose. If we assume an acceptable power of 80%, we need to recruit 18 practices and 15 patients in each practice, or 27 practices and 10 patients in each practice under our study design and assumptions.

In general, we can also find other combinations of number of practices and number of patients per practice that achieves 80% power other than those shown in Table 2. With the data generating and power calculation method we have developed, it is convenient to vary the numbers and choose the optimal combination of number of practices and practice size that achieves adequate power within the bounds of logistical and implementation constraints.

Figures 4 and 5 show visually the influence of sample size on the power under different levels of delay. The number of practices were set to be 18 and the number of repeated measurements were set to 5 for each patient. In Figure 4, subject delay was fixed to 50% and the curves represent different degrees of time delay. In Figure 5, meanwhile, time delay was fixed to 50% and the curves represent different levels of subject delay. Both figures show similar levels and trend in power as the sample size increases. In the no delay case, we need about 14 patients in each practice to obtain 80% power, while under major delay case, we need about 16 patients in

each practice to obtain 80% power. In mild and moderate case, around 15 patients are required to obtain 80% power. The powers in different delay levels are almost parallel. The power increases more rapidly when the number of practice is small and power is not high, but when power gets large, the rate of increase becomes smaller.

Table 2. Sample size and power

		No delay	**Mild delay**	**Moderate delay**	**Major delay**
Practices	**Patients***	p_d=0 l_d=0**	p_d=0.5 l_d=1.25	p_d=0.5 l_d=2.5	p_d=0.5 l_d=3.75
9	**5**	0.240	0.212	0.202	0.197
	10	0.381	0.356	0.347	0.334
	15	0.580	0.512	0.498	0.485
	20	0.667	0.626	0.613	0.592
18	**5**	0.387	0.380	0.377	0.358
	10	0.659	0.603	0.587	0.569
	15	0.820	0.799	0.781	0.772
	20	0.918	0.914	0.905	0.894
27	**5**	0.540	0.511	0.496	0.485
	10	0.813	0.825	0.812	0.802
	15	0.930	0.927	0.922	0.913
	20	0.990	0.981	0.976	0.973
36	**5**	0.655	0.649	0.629	0.611
	10	0.913	0.918	0.903	0.885
	15	0.981	0.982	0.974	0.973
	20	0.997	0.999	0.998	0.995

* number of patients in each practice ** p_d denotes subject delay, l_d denotes time delay

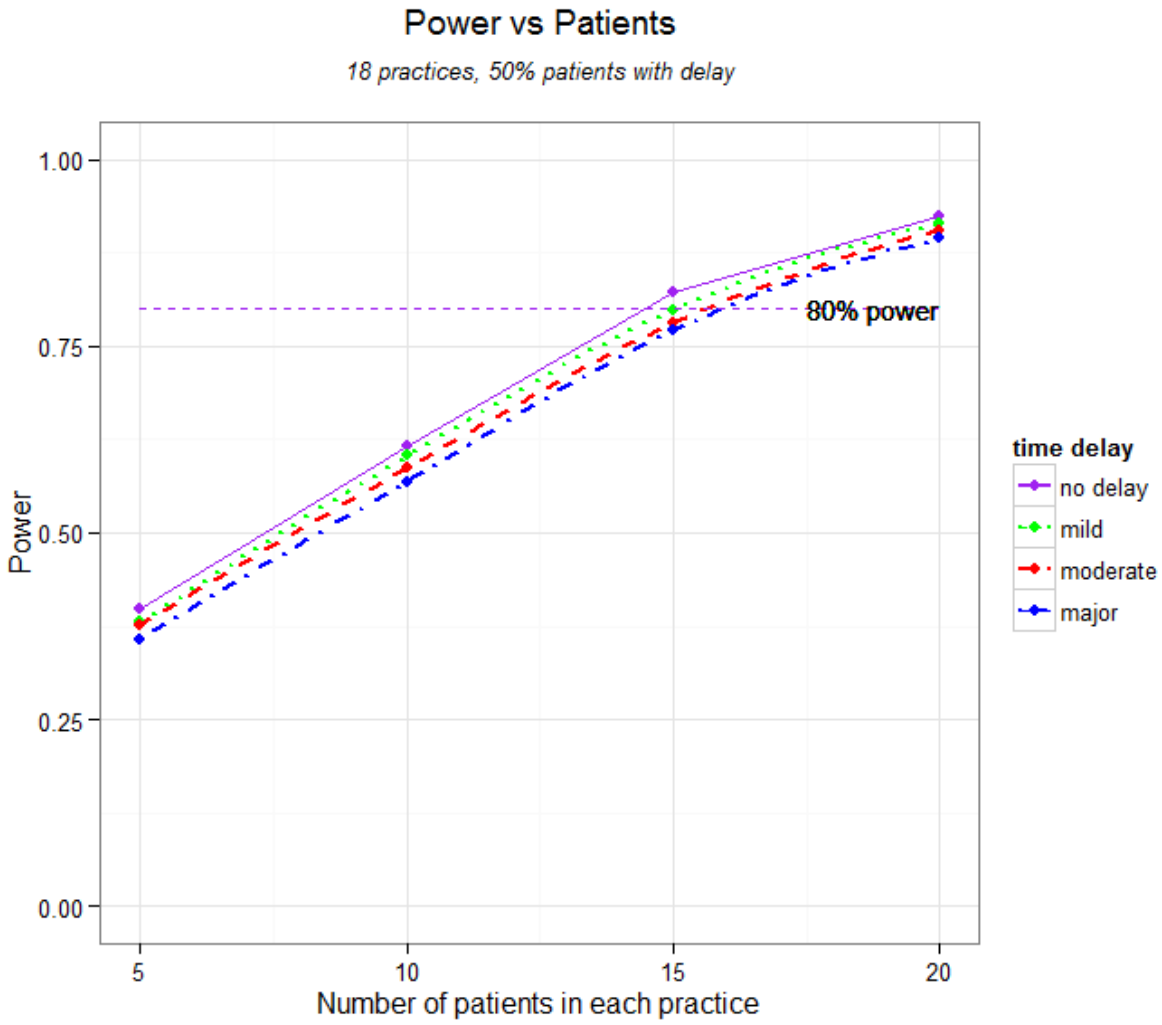


Figure 4. Power vs Patients (fix subject delay)

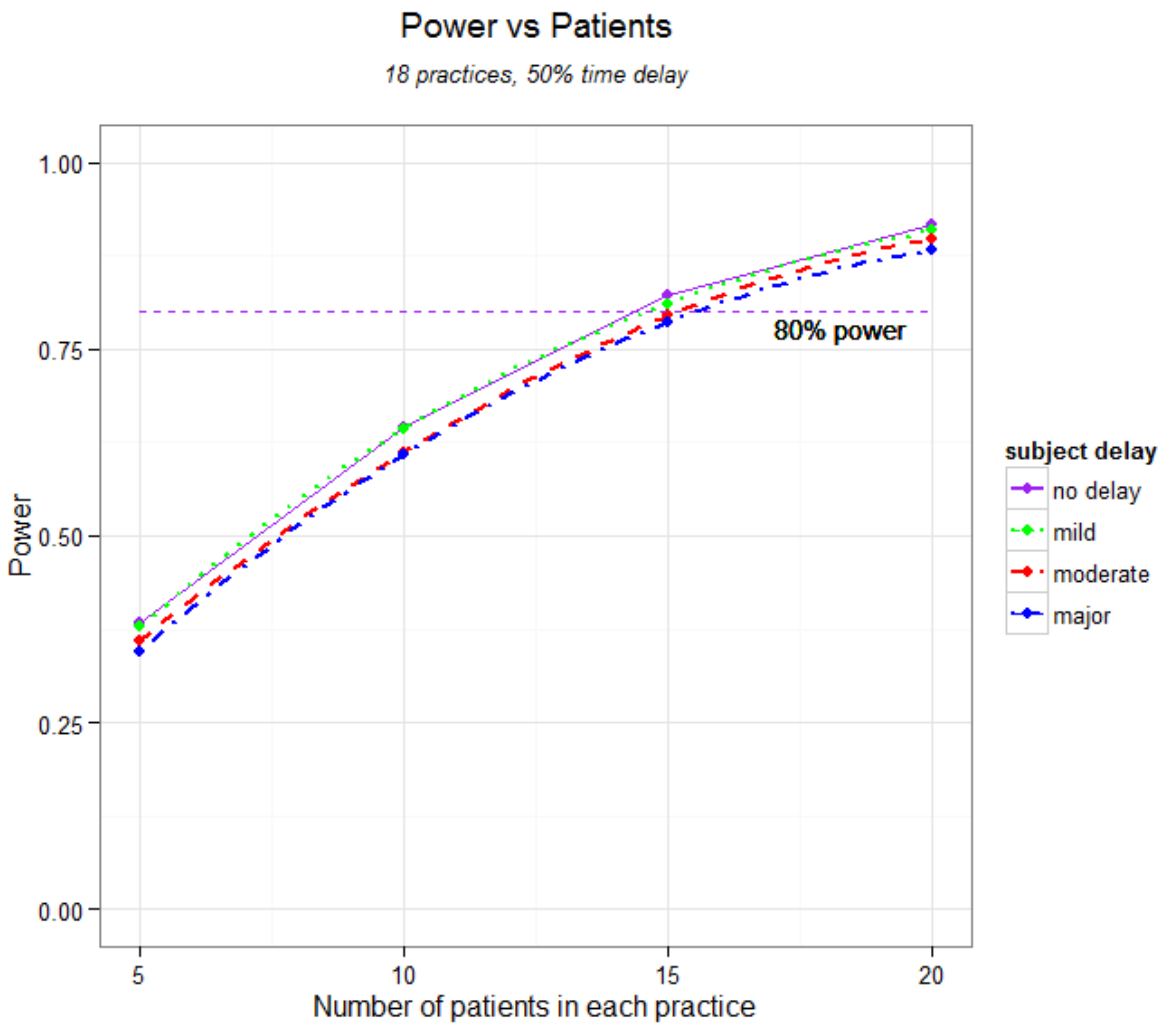


Figure 5. Power vs Patients (fix time delay)

Table 3 shows the joint effect of subject delay and time delay on the power. This table was generated assuming that there are 18 practices and 15 patients in each practice. Although marginally either time delay or subject delay had small impact on power based on Table 2 and Figures 4 and 5, Table 3 indicates that the loss in power could be substantial when they are taken jointly. The reduction in power could be as high as 32.1%, as can be seen in the mildest case (subject and time delay are both 25%) to compare with the worst case (both subject and time delay

are 100%). From mild to major (subject and time delay are both 75%), the reduction in power was 6.5%.

It also appears in Table 3 that the impact of time delay and subject delay on power are similar. For example, when time delay was fixed to 25%, the reduction in power due to subject delay (comparing 75% versus 25%) was only 2.7%; similarly when subject delay was fixed to 25%, the reduction in power due to time delay (75% versus 25%) was 2.6%. The reduction in power may be slightly greater if either subject delay or time delay was major or worst (i.e., 75%-100% delay). This is much easier seen in Figure 6, which is just the plot of Table 3. For example, the reduction in power was 3.9% comparing 75% versus 25% subject delay when time delay is 75%; the reduction in power was 4.6% comparing 75% versus 25% time delay when subject delay is 75%.

Table 3. Different combinations of subject and time delay

Subject delay	**Time delay (l_d/5*100%)**			
(p_d*100%)	**25%**	**50%**	**75%**	**100%**
25%	0.815	0.808	0.801	0.789
50%	0.805	0.791	0.777	0.738
75%	0.808	0.785	0.762	0.666
100%	0.788	0.755	0.719	0.553

18 practices, 15 patients in each practice, power is 0.820 when there is no delay

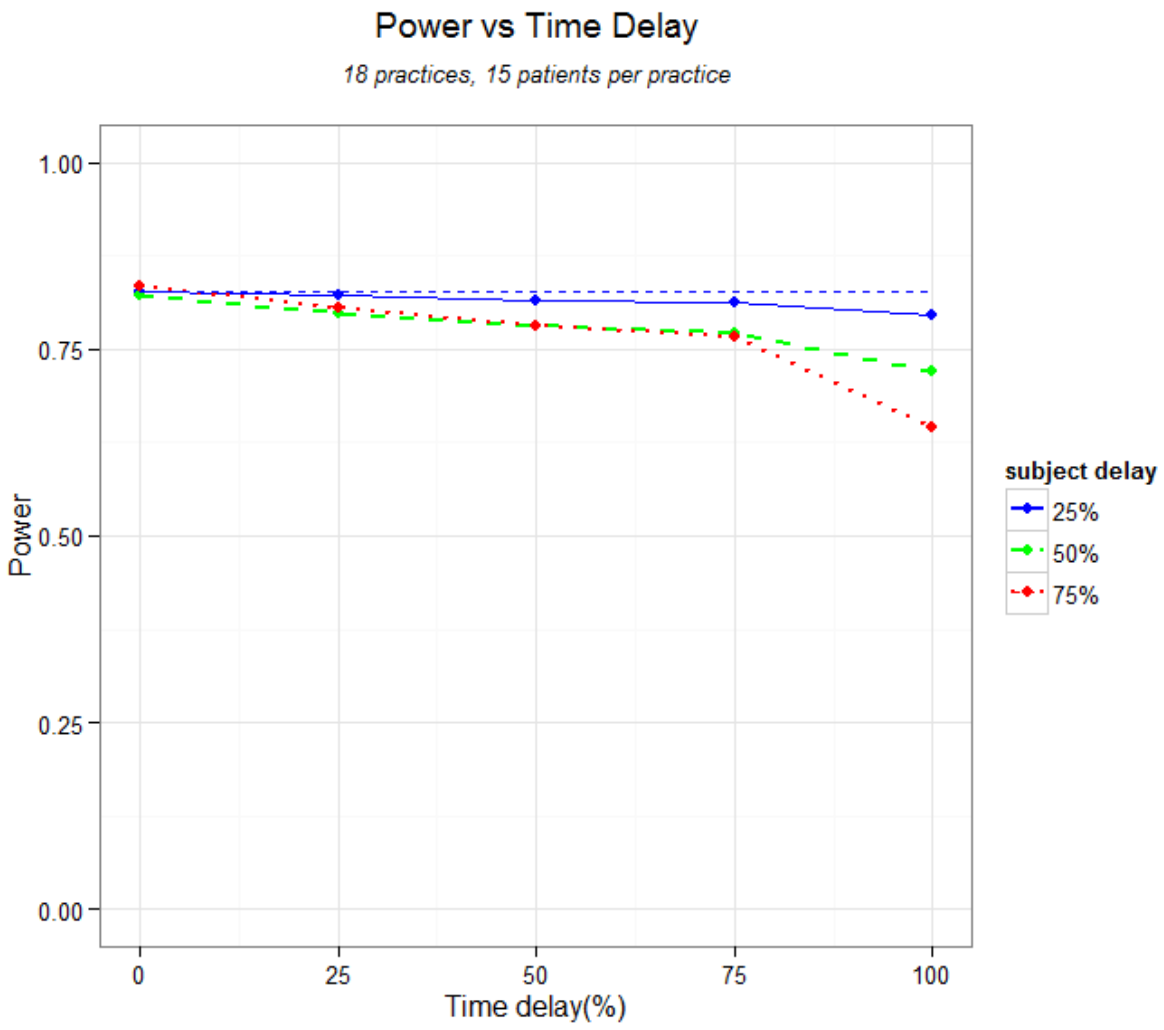


Figure 6. Power vs Time delay

4.0 DISCUSSION AND CONCLUSION

In this study, a simulation-based power and sample size calculations for SWTs with repeated measurements and short exposure design was developed. Existing sample size formulas for SWTs are applicable to studies with cross-sectional outcomes. Although convenient, these methods cannot be directly used to design SWTs with longitudinal outcomes. Simulation based calculations on the other hand are flexible and can accommodate multiple features of a trial such as repeated measurements, multilevel data, and delayed treatment initiation. The model we used accounts for clustering of repeated observations within a patient, and clustering of patients within a practice. The simulation design was motivated by the EHR PHM trial, a proposed future study that aims to assess whether a population management intervention is effective in slowing down the decline in GFR of high risk CKD patients.

The simulation-based approach allowed us to study the impact of delayed treatment initiation on power. Delayed treatment initiation is a concern in the EHR PHM trial. We found that delayed treatment initiation has a negative effect on the power. When either subject delay or time delay is mild, we do not need to worry about the loss of power. However, when both subject delay and time delay are major, the loss of power may be substantial. The impact of varying levels of subject delay for a fixed time delay is similar to the impact of varying levels of time delay for a fixed subject delay. In general, the loss in power due to delayed treatment initiation is small compared to the loss in power due to delayed treatment effect as reported in the literature [7]. This may be because the maximum length of time delay we used in the simulations is short (5 months). If we allow the delay to be longer, the impact on power may be larger. This warrants further investigation.

In our model, we assumed that each practice has the same number of patients and each patient has the same number of measurements. Also, equal number of clusters roll-out at each of the transition time points. However, it is possible that the data collected is not balanced, especially in a study with continuous recruitment design. Furthermore, we assumed that the measurements of each patients are equally spaced. In reality, this may not be necessarily true. Power analysis of the SWT with unbalanced or unequally spaced repeated measurements could be further topics of study.

By using the random effects mixed model, we implicitly assumed compound symmetric covariance pattern between observations within a subject. This covariance structure assumes that the correlations of observations within the same subject between any two time points, regardless of how far apart they are, are equal. Other types of correlation structure, such as Toeplitz or Autoregressive, may be alternatives for this study. If our interest is the intervention effect on population-level, other options of the model could be covariance pattern model or using GEE, where more complex correlation structures between the repeated measurements could be assumed. Finally, our approach focused on continuous outcomes assuming normal distribution. In general, however, the data generating approach can be easily expanded to generate datasets with binary outcomes. This is possible by taking the generalized linear mixed model approach.

In clinical studies, SWTs with longitudinal outcomes have been increasingly used in recent times. A simulation-based power calculation method provides a practical guidance for power calculation and sample size determination in designing the SWTs with longitudinal outcomes while incorporating the effect of delayed treatment initiation. This method will be useful in planning the EHR-PHM trial which could potentially improve care for and outcomes of high risk CKD patients. This is a significant public health concern given that CKD affects 26 million adults and is one of the major causes of morbidity and mortality in the United States.

APPENDIX: R MARKDOWN CODE USED FOR SIMULATION

````
---
title: "Simulation-based power analysis for SWTs with repeated measurements"
author: "Peng"
date: "March 8, 2016"
output: html_document
---

```{r setup, include=FALSE}
require(lme4)  # mixed models
require(nlme)  # mixed models
require(compiler)  # to compile our function
require(dplyr); require(tidyr) # for efficiently manipulating datasets
require(ggplot2) # draw the graph

setwd("C:/2016 spring/thesis/results")
```


```{r results='hide'}
# generate diagonal matrix
dmat <- cmpfun(function(i) {
  j <- length(i)
  n <- sum(i)
  index <- cbind(start = cumsum(c(1, i[-j])), stop = cumsum(i))
  H <- matrix(0, nrow = n, ncol = j)
  for (i in 1:j) {
    H[index[i, 1]:index[i, 2], i] <- 1L
  }
  return(H)
})
```

## functions
- Set up a simulation function for continuous case
```{r}
## continuous outcome without delay
sim_con<-function(iter=1000,j=5,k=18,o=0.5,q=0.5){
  #iter<-1000
  pow.r<-matrix(NA,iter)
  set.seed(123)
  for(rep in 1:iter){
    # 9 practices, 5 subjects for each practice, 4 obs for each subject
    i<-5
    #j<-5
    #k<-9
    # set variances
    sigma2_k<- 41.39
    sigma2_j<- 120.43
    sigma2_e <-45.14
    # number of subjects
````
````

```
n<- j*k
# number of observations
m<-n*i

# proportion of pts delay
#o<-0.3
# proportion of delay time
#q<-0.5

# random effect of practice
T<- rep(i*j,k)
#z1<- diag(k) %x% rep(1,i*j)
z1<-dmat(T)
a<- rnorm(k,0,sqrt(sigma2_k))
random_k<-z1 %*% a
# random effect of subject
z2<- diag(n) %x% rep(1,i)
b<- rnorm(n,0,sqrt(sigma2_j))
random_j<-z2 %*% b
# error term
e<- rnorm(m,0,sqrt(sigma2_e))

# fixed effects
# overall mean
mu=46.45
# randomly select three practice at each crossover time point
k1<- sample(1:k,k/3,replace=F)
k2<- sample(setdiff(1:k,k1),k/3,replace=F)
k3<- setdiff(1:k,c(k1,k2))
# practice
c<-1:k
N<-rep(j,k)
M<-rep(i,n)
practice<-dmat(M) %*% dmat(N) %*% c
# patient
d<-rep(1:j,k)
patient<-dmat(M) %*% d
# visit
visit<-rep(0:(i-1),n)
# time point of each subject
tp<-c(0,5,10,15)
enterm<-sample(tp,n,replace=TRUE)

# the following observations time

t1<-enterm+6
t2<-t1+6
t3<-t2+6
t4<-t3+6

# simulate delayed treatment effect
delay<-sample(1:n,floor(n*o),replace=FALSE)
enterm[delay]<-round(enterm[delay]+q*5,digits=2)

# create colomn delay
delayflag<-rep(0,n)
delayflag[delay]<-1
```

```
    # transform enterm vector to match the dimension
    delayflag<-dmat(M) %*% delayflag

    # convert to a vector
    time<-cbind(enterm,t1,t2,t3,t4)

    time<-as.vector(t(time))

    # transform enterm vector to match the dimension
    enterm<-dmat(M) %*% enterm

    # combine practice,patient,visit,enterm and time to one dataset

    data<-data.frame(practice,patient,visit,delayflag,enterm,time)

    #create trt and time_dif columns
    data<-within(data, {
      trt=ifelse(practice %in% k1 & enterm>=5,1,
                 ifelse(practice %in% k2 & enterm>=10,1,
                        ifelse(practice %in% k3 & enterm>=15,1,0)
                 )
      )
      time_dif=time-enterm
    })

    #Set parameters
    beta1<- -0.49
    beta2<- 0
    theta<- 0.125

    trt<-data$trt
    time_dif<-data$time_dif
    #  time_dif0<-rep(c(6,12,18,24),n)

    # outcome
    y<-mu+beta1*time_dif+beta2*trt+theta*(trt*time_dif)+random_k+random_j+e
    #final dataset
    data<-data.frame(data,y)

    tryCatch({
    # fit lmm
    #fit <- lmer(y ~ time_dif*trt + (1 | practice) + (1 |practice:patient),
data = data)
    #fits<-summary(fit)
    #fitt<-fits$coefficients[3,3]
    # fit lmm
    fit<- lme(y ~ trt*time_dif, random= ~ 1|practice/patient , data=data)
    fits<- summary(fit)
    fitp<- fits$tTable[4,5]
    rej<-fitp<0.05

    # record power
    pow.r[rep]<-rej
    }, warning = function(war) {
      # warning handler picks up where error was generated
      print(paste("MY_WARNING:  ",war))
    },error = function(err) {
```

```
      # error handler picks up where error was generated
      print(paste("MY_ERROR:  ",err))
    }
    )

    # count the iterations
    if( rep %% 100 == 0 ) cat(paste("iteration", rep, "complete\n"))
  }
  mean(pow.r[complete.cases(pow.r)])
}
```

Analysis

Table 1 different combinations of % of patients and % of delay time
- Change the % of delayed patients and the % of delayed time
```{r}
# continuous case with delay, o from 0.25 to 1, q from 0.25 to 1
possible.o<-seq(from=0.25,to=1,by=0.25)
possible.q<-seq(from=0.25,to=1,by=0.25)
powers<-matrix(NA,length(possible.o),length(possible.q))
colnames<-c("0.25","0.5","0.75","1")
rownames<-c("0.25","0.5","0.75","1")
iter<-1000
for(Z in 1: length(possible.o)){
    for(Y in 1: length(possible.q)){

      o=possible.o[Z]
      q=possible.q[Y]
      powers[Z,Y]<-sim_con(iter=iter,k=45,o=o,q=q)
    }
}
powers
```

Table 2 sample size and power in different levels of delayed treatment
- Change the sample size in no delay case
```{r}

possible.k<-seq(from=9,to=36,by=9)
possible.j<-seq(from=5,to=20,by=5)
powers_kj_no<-matrix(NA,length(possible.k),length(possible.j))
colnames(powers_kj_no)<-c("5","10","15","20")
rownames(powers_kj_no)<-c("9","18","27","36")
iter<-1000
for(Z in 1: length(possible.k)){
    for(Y in 1: length(possible.j)){

      k=possible.k[Z]
      j=possible.j[Y]
      powers_kj_no[Z,Y]<-sim_con(iter=iter,o=0,q=0,k=k,j=j)
    }
}
powers_kj_no
```

- Change the sample size in mild case

```{r}

possible.k<-seq(from=9,to=36,by=9)
possible.j<-seq(from=5,to=20,by=5)
powers_kj_mild<-matrix(NA,length(possible.k),length(possible.j))
colnames(powers_kj_mild)<-c("5","10","15","20")
rownames(powers_kj_mild)<-c("9","18","27","36")
iter<-1000
for(Z in 1: length(possible.k)){
    for(Y in 1: length(possible.j)){

      k=possible.k[Z]
      j=possible.j[Y]
      powers_kj_mild[Z,Y]<-sim_con(iter=iter,o=0.5,q=0.25,k=k,j=j)
    }
}
powers_kj_mild
```

- Change the sample size in moderate case

```{r}

possible.k<-seq(from=9,to=36,by=9)
possible.j<-seq(from=5,to=20,by=5)
powers_kj_mod<-matrix(NA,length(possible.k),length(possible.j))
colnames(powers_kj_mod)<-c("5","10","15","20")
rownames(powers_kj_mod)<-c("9","18","27","36")
iter<-1000
for(Z in 1: length(possible.k)){
    for(Y in 1: length(possible.j)){

      k=possible.k[Z]
      j=possible.j[Y]
      powers_kj_mod[Z,Y]<-sim_con(iter=iter,o=0.5,q=0.5,k=k,j=j)
    }
}
powers_kj_mod
```

- Change the sample size in major case

```{r}

possible.k<-seq(from=9,to=36,by=9)
possible.j<-seq(from=5,to=20,by=5)
powers_kj_maj<-matrix(NA,length(possible.k),length(possible.j))
colnames(powers_kj_maj)<-c("5","10","15","20")
rownames(powers_kj_maj)<-c("9","18","27","36")
iter<-1000
for(Z in 1: length(possible.k)){
    for(Y in 1: length(possible.j)){

      k=possible.k[Z]
      j=possible.j[Y]
```

```
      powers_kj_maj[Z,Y]<-sim_con(iter=iter,o=0.5,q=0.75,k=k,j=j)
    }
}
powers_kj_maj
```

Graph 1 subject delay vs. power under different levels of patients
- j set to be 15, k set to be 18
- subject delay set to be 25% of total subjects
```{r}
# 25% subject delay
possible.q<-seq(from=0,to=1,by=0.25)
powers25<-rep(NA,length(possible.q))
iter<-5000
for(Z in 1: length(possible.q)){
  q=possible.q[Z]
  powers25[Z]<-sim_con(iter=iter,j=15,k=18,o=0.25,q=q)

}
powers25
```
- subject delay set to be 50% of total subjects
```{r}
# 50% subject delay
possible.q<-seq(from=0,to=1,by=0.25)
powers50<-rep(NA,length(possible.q))
iter<-5000
for(Z in 1: length(possible.q)){
  q=possible.q[Z]
  powers50[Z]<-sim_con(iter=iter,j=15,k=18,o=0.5,q=q)

}
powers50
```
- subject delay set to be 75% of total subjects
```{r}
# 75% subject delay
possible.q<-seq(from=0,to=1,by=0.25)
powers75<-rep(NA,length(possible.q))
iter<-5000
for(Z in 1: length(possible.q)){
  q=possible.q[Z]
  powers75[Z]<-sim_con(iter=iter,j=15,k=18,o=0.75,q=q)

}
powers75
```


```{r results='hide'}
# combine all powers to one dataframe
df1 = data.frame(possible.q,powers25,powers50,powers75)
```

- plot the powers of different levels of delayed treatment
```{r}


```

```
# Set the % of subject with delay to 50%, change the % of time interval that
subjects delay
ggplot(df1, aes(x = possible.q*100, y =
powers25,colour="25%",linetype="25%"),lwd=1)+  geom_line(se=FALSE)+
geom_point()+

geom_line(aes(x=possible.q*100,y=powers50,colour="50%",linetype="50%"),se=FAL
SE,lwd=1)+
  geom_point(aes(x=possible.q*100,y=powers50,colour="50%"))+

geom_line(aes(x=possible.q*100,y=powers75,colour="75%",linetype="75%"),se=FAL
SE,lwd=1)+
  geom_point(aes(x=possible.q*100,y=powers75,colour="75%"))+
  geom_segment(aes(x = 0, y =powers25[1] , xend = 100, yend
=powers25[1] ),linetype=2)+
  scale_colour_manual("subject delay",
                      breaks = c("25%", "50%", "75%"),
                      values = c("blue", "green", "red")) +
  scale_linetype_manual("subject delay",
                      breaks = c("25%", "50%", "75%"),
                      values=c(1,2,3))+
  xlab("Time delay(%)")+ylab("Power")+
  ggtitle(bquote(atop(.("Power vs Time Delay"), atop(italic(.("18 practices,
15 patients per practice")), ""))))+
  ylim(0,1) +xlim(0,100)+ theme_bw()

```

Graph 3&4 the number of subject in each practice vs. power under different levels of delayed treatment

- j from 5 to 20, k set to 18
- continuous case without delay

```{r}
# continuous case without delay, j from 5 to 50
possible.ns<-seq(from=5,to=20,by=5)
powers1<-rep(NA,length(possible.ns))
iter<-5000
for(Z in 1: length(possible.ns)){
  j=possible.ns[Z]
  powers1[Z]<-sim_con(iter=iter,j=j,o=0.5,q=0)

}
powers1
```

- continuous case with moderate-mild delay(50% patients with 25% delay of the time interval)

```{r}
# continuous case with delay, j from 5 to 50
possible.ns<-seq(from=5,to=20,by=5)
powers2<-rep(NA,length(possible.ns))
iter<-5000
for(Z in 1: length(possible.ns)){
  j=possible.ns[Z]
  powers2[Z]<-sim_con(iter=iter,j=j,o=0.5,q=0.25)

}
powers2
```

```
- continuous case with moderate-moderate delay(50% patients with 50% delay of the time interval)
```{r}
# continuous case with delay, j from 5 to 50
possible.ns<-seq(from=5,to=20,by=5)
powers3<-rep(NA,length(possible.ns))
iter<-5000
for(Z in 1: length(possible.ns)){
  j=possible.ns[Z]
  powers3[Z]<-sim_con(iter=iter,j=j,o=0.5,q=0.5)

}
powers3
```

- continuous case with moderate-major delay(50% patients with 75% delay of the time interval)
```{r}
# continuous case with delay, j from 5 to 50
possible.ns<-seq(from=5,to=20,by=5)
powers4<-rep(NA,length(possible.ns))
iter<-5000
for(Z in 1: length(possible.ns)){
  j=possible.ns[Z]
  powers4[Z]<-sim_con(iter=iter,j=j,o=0.5,q=0.75)

}
powers4
```

- continuous case with mild-moderate delay(25% patients with 50% delay of the time interval)
```{r}
# continuous case with delay, j from 5 to 50
possible.ns<-seq(from=5,to=20,by=5)
powers5<-rep(NA,length(possible.ns))
iter<-5000
for(Z in 1: length(possible.ns)){
  j=possible.ns[Z]
  powers5[Z]<-sim_con(iter=iter,j=j,o=0.25,q=0.5)

}
powers5
```

- continuous case with major-moderate delay(75% patients with 50% delay of the time interval)
```{r}
# continuous case with delay, j from 5 to 50
possible.ns<-seq(from=5,to=20,by=5)
powers6<-rep(NA,length(possible.ns))
iter<-5000
for(Z in 1: length(possible.ns)){
  j=possible.ns[Z]
  powers6[Z]<-sim_con(iter=iter,j=j,o=0.75,q=0.5)
```
```

```
}
powers6
```


```{r results='hide'}
# combine all powers to one dataframe
df2 = data.frame(possible.ns,
powers1,powers2,powers3,powers4,powers5,powers6)
```

- plot the powers of different levels of delayed treatment
```{r}
# Set the % of subject with delay to 50%, change the % of time interval that
subjects delay
ggplot(df2, aes(x = possible.ns, y = powers1, colour="no delay",linetype="no
delay"),lwd=1)+  geom_line(se=FALSE,span=1)+geom_point()+
  geom_smooth(aes(x=possible.ns,y=powers2,
colour="mild",linetype="mild"),se=FALSE,span=1,lwd=1)+
  geom_point(aes(x=possible.ns,y=powers2, colour="mild"))+

geom_smooth(aes(x=possible.ns,y=powers3,colour="moderate",linetype="moderate"
),se=FALSE,span=1,lwd=1)+
  geom_point(aes(x=possible.ns,y=powers3,colour="moderate"))+

geom_smooth(aes(x=possible.ns,y=powers4,colour="major",linetype="major"),se=F
ALSE,span=1,lwd=1)+
  geom_point(aes(x=possible.ns,y=powers4,colour="major"))+
  geom_segment(aes(x = 5, y = 0.8, xend = 20, yend = 0.8),linetype=2)+
  geom_text(aes(x=20, label="80% power",
y=0.8),hjust=1,colour="black",size=4)+
  scale_colour_manual("time delay",
                      breaks = c("no delay", "mild", "moderate","major"),
                      values = c("blue", "green", "red","purple")) +
  scale_linetype_manual("time delay",
                      breaks = c("no delay", "mild", "moderate","major"),
                      values=c(4,3,2,1))+
  labs(x="Number of patients in each practice",y="Power")+
  ggtitle(bquote(atop(.("Power vs Patients"), atop(italic(.("18 practices,
50% patients with delay")), ""))))+
  ylim(0,1) +xlim(5,20)+ theme_bw()

# Set the % of time subjects delay to 50%, change the % of subjects that
delay
ggplot(df2, aes(x = possible.ns, y = powers1,colour="no delay",linetype="no
delay"),lwd=1)+  geom_line(se=FALSE,span=1)+geom_point()+

geom_smooth(aes(x=possible.ns,y=powers5,colour="mild",linetype="mild"),se=FAL
SE,span=1,lwd=1)+
  geom_point(aes(x=possible.ns,y=powers5,colour="mild"))+

geom_smooth(aes(x=possible.ns,y=powers3,colour="moderate",linetype="moderate"
),se=FALSE,span=1,lwd=1)+
  geom_point(aes(x=possible.ns,y=powers3,colour="moderate"))+

geom_smooth(aes(x=possible.ns,y=powers6,colour="major",linetype="major"),se=F
ALSE,span=1,lwd=1)+
```

```
  geom_point(aes(x=possible.ns,y=powers6,colour="major"))+
  geom_segment(aes(x = 5, y = 0.8, xend = 20, yend = 0.8),linetype=2)+
  geom_text(aes(x=20, label="80% power",
y=0.8),hjust=1.2,vjust=1.2,colour="black",size=4)+
  scale_colour_manual("subject delay",
                      breaks = c("no delay", "mild", "moderate","major"),
                      values = c("blue", "green", "red","purple")) +
  scale_linetype_manual("subject delay",
                      breaks = c("no delay", "mild", "moderate","major"),
                      values=c(4,3,2,1))+
  xlab("Number of patients in each practice")+ylab("Power")+
  ggtitle(bquote(atop(.("Power vs Patients"), atop(italic(.("18 practices,
50% time delay")), ""))))+
  ylim(0,1) + xlim(5,20)+ theme_bw()

```

BIBLIOGRAPHY

1. Saran R, Li Y, Robinson B, Abbott KC, Agodoa LY, Ayanian J, Bragg-Gresham J, Balkrishnan R, Chen JL, Cope E, Eggers PW. US Renal Data System 2015 Annual Data Report: Epidemiology of Kidney Disease in the United States. American journal of kidney diseases: the official journal of the National Kidney Foundation. 2016 Mar;67(3 Suppl 1):A7.

2. National Kidney Foundation. About Chronic Kidney Disease. Available from: https://www.kidney.org/kidneydisease/aboutckd

3. Levey AS, Becker C, Inker LA. Glomerular filtration rate and albuminuria for detection and staging of acute and chronic kidney disease in adults: a systematic review. JAMA. 2015 Feb 24;313(8):837-46.

4. Levey AS, Inker LA, Matsushita K, Greene T, Willis K, Lewis E, De Zeeuw D, Cheung AK, Coresh J. GFR decline as an end point for clinical trials in CKD: a scientific workshop sponsored by the National Kidney Foundation and the US Food and Drug Administration. American Journal of Kidney Diseases. 2014 Dec 31;64(6):821-35.

5. Narva AS. Decision support and CKD: not there yet. Clinical Journal of the American Society of Nephrology. 2012 Apr 1;7(4):525-6.

6. Smart NA, Titus T, Dooley L. Early referral to specialist nephrology services for preventing the progression to end‐stage kidney disease. The Cochrane Library. 2008.

7. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemporary clinical trials. 2007 Feb 28;28(2):182-91.

8. Hemming K, Lilford R, Girling AJ. Stepped‐wedge cluster randomised controlled trials: a generic framework including parallel and multiple‐level designs. Statistics in medicine. 2015 Jan 30;34(2):181-96.

9. Hargreaves JR, Copas AJ, Beard E, Osrin D, Lewis JJ, Davey C, Thompson JA, Baio G, Fielding KL, Prost A. Five questions to consider before conducting a stepped wedge trial. Trials. 2015 Aug 17;16(1):350.

10. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. bmj. 2015 Feb 6;350:h391.

11. Baio G, Copas A, Ambler G, Hargreaves J, Beard E, Omar RZ. Sample size calculation for a stepped wedge trial. Trials. 2015 Aug 17;16(1):354.

12. Copas AJ, Lewis JJ, Thompson JA, Davey C, Baio G, Hargreaves JR. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. Trials. 2015 Aug 17;16(1):352.

13. Dimairo M, Bradburn M, Walters SJ. Sample size determination through power simulation; practical lessons from a stepped wedge cluster randomised trial (SW CRT). Trials. 2011 Dec 13;12(Suppl 1):A26.

14. Girling AJ, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. Statistics in medicine. 2016 Jan 1.

15. Feldman HA, McKinlay SM. Cohort versus cross‐sectional design in large field trials: precision, sample size, and a unifying model. Statistics in medicine. 1994 Jan 15;13(1):61-78.

16. Woertman W, de Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL, Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. Journal of clinical epidemiology. 2013 Jul 31;66(7):752-8.

17. Kotz D, Spigt M, Arts IC, Crutzen R, Viechtbauer W. Use of the stepped wedge design cannot be recommended: a critical appraisal and comparison with the classic cluster randomized controlled trial design. Journal of clinical epidemiology. 2012 Dec 31;65(12):1249-52.

18. Moerbeek M. The consequence of ignoring a level of nesting in multilevel analysis. Multivariate Behavioral Research. 2004 Jan 1;39(1):129-49.

19. Davey C, Hargreaves J, Thompson JA, Copas AJ, Beard E, Lewis JJ, Fielding KL. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. Trials. 2015 Aug 17;16(1):358.

20. Kotz D, Spigt M, Arts IC, Crutzen R, Viechtbauer W. The stepped wedge design does not inherently have more power than a cluster randomized controlled trial. Journal of clinical epidemiology. 2013 Jan 9;66(9):1059-60.

21. Donner A, Klar N. Design and analysis of cluster randomization trials in health research. 2000.

22. Meurer WJ, Lewis RJ. Cluster randomized trials: evaluating treatments applied to groups. JAMA. 2015 May 26;313(20):2068-9.

23. Sur D, Ochiai RL, Bhattacharya SK, Ganguly NK, Ali M, Manna B, Dutta S, Donner A, Kanungo S, Park JK, Puri MK. A cluster-randomized effectiveness trial of Vi typhoid vaccine in India. New England Journal of Medicine. 2009 Jul 23;361(4):335-44.

24. Porter J, Haines T, Truby H. Implementation of protected mealtimes in the subacute setting: stepped wedge cluster trial protocol. Journal of advanced nursing. 2016 Feb 1.

25. Pearse R. Enhanced peri-operative care for high-risk patients (EPOCH) trial: a stepped wedge cluster randomised trial of a quality improvement intervention for patients undergoing emergency laparotomy. Lancet. 2014:1-28.