Xia *et al. Trials* (2021) 22:598
https://doi.org/10.1186/s13063-021-05542-9

Trials

METHODOLOGY Open Access

# Power and sample size calculation for stepped-wedge designs with discrete outcomes

Fan Xia[1*], James P. Hughes[2], Emily C. Voldal[2] and Patrick J. Heagerty[2]

## Abstract

**Background:** Stepped-wedge designs (SWD) are increasingly used to evaluate the impact of changes to the process of care within health care systems. However, to generate definitive evidence, a correct sample size calculation is crucial to ensure such studies are properly powered. The seminal work of Hussey and Hughes (Contemp Clin Trials 28(2):182–91, 2004) provides an analytical formula for power calculations with normal outcomes using a linear model and simple random effects. However, minimal development and evaluation have been done for power calculation with non-normal outcomes on their natural scale (e.g., logit, log). For example, binary endpoints are common, and logistic regression is the natural multilevel model for such clustered data.

**Methods:** We propose a power calculation formula for SWD with either normal or non-normal outcomes in the context of generalized linear mixed models by adopting the Laplace approximation detailed in Breslow and Clayton (J Am Stat Assoc 88(421):9–25, 1993) to obtain the covariance matrix of the estimated parameters.

**Results:** We compare the performance of our proposed method with simulation-based sample size calculation and demonstrate its use on a study of patient-delivered partner therapy for STI treatment and a study that assesses the impact of providing additional benchmark prevalence information in a radiologic imaging report. To facilitate adoption of our methods we also provide a function embedded in the R package "swCRTdesign" for sample size and power calculation for multilevel stepped-wedge designs.

**Conclusions:** Our method requires minimal computational power. Therefore, the proposed procedure facilitates rapid dynamic updates of sample size calculations and can be used to explore a wide range of design options or assumptions.

**Keywords:** Stepped-wedge designs, Power calculation, Non-normal outcomes, Minimal computational power

## Background

Great progress has been made in public health and medical care during the past century through immunization, food safety, improvements in maternal and infant health, and advances in drugs, devices, and strategies to treat disease. However, continuing efforts to improve the quality and efficiency of care require rigorous evaluation to further guide decision-making. To evaluate novel strategies within health care delivery systems, cluster randomized trials (CRT) represent a key experimental design that may be used when individual randomization is not feasible due to administrative, financial or ethical reasons [1, 2].

Stepped-wedge designs (SWD) are a type of contemporary and novel CRT that have been used to evaluate new interventions and programs deployed in the context

*Correspondence: fanxia@uw.edu
[1]National Alzheimer's Coordinating Center, University of Washington, Seattle, WA, USA
Full list of author information is available at the end of the article

of routine implementation [3, 4]. A SWD is unique in that it combines key elements of cluster-randomized designs with a crossover component commonly used in longitudinal designs. Specifically, in SWD, all clusters (typically) start in the control group, cross over to the intervention group at different time points, and stay on intervention until the end of the trial. The time at which each cluster starts the intervention is randomized. Either different individuals (cross-sectional design) from each cluster may be measured at different time points or the same individuals may be repeatedly assessed (cohort design).

In the past 10 years, an increasing number of studies have used the SWD in health-related research within a broad range of domains, including HIV treatment, infection prevention, nutrition, asthma, cancer, and trauma. Recent SWD trials have been conducted in various global settings including America, Europe, Africa, Asia, and Australia. Compared to a standard parallel design CRT, the SWD is preferred in some circumstances due to practical, ethical, or methodological concerns [5–7].

Linear mixed models and generalized linear mixed models [8] are commonly used for the analysis of SWD data. A linear mixed model (LMM) is a type of regression model that includes random effects in addition to the standard fixed effects used in a linear model to account for dependence among observations from the same cluster. The use of random effects is a natural way to represent the heterogeneity among clusters under study and these methods produce valid inference when assumptions are satisfied. Generalized linear mixed models (GLMM) extend the LMM framework to non-normal data and non-identity links such as logistic regression for binary outcomes or Poisson regression for count data.

The increasing adoption of the SWD necessitates the development of flexible and valid sample size calculations. Hussey and Hughes [9] provided analytical formulae for power calculations based on repeated cross-sectional samples using a weighted least squares approach. The Hussey and Hughes [9] power calculations were based on a linear mixed model with a random cluster effect only. Woertman et al. [10] proposed a design effect that accounts for the inflation caused by the within-cluster correlation based on the Hussey and Hughes [9] formulation. Hooper et al. [11] reviewed designs for cluster randomized trials with repeated cross-sections and included random effects for time within clusters for stepped wedge designs. Hooper et al. [12] introduced sample size calculation for longitudinal CRTs including SWD, in which they include random effects for time within clusters for closed cohort designs. Hemming and Taljaard [13] summarized the design effects for SWD and CRT and provided a unified approach for their sample size calculation. Power calculations based on mixed models with random intervention effects may also be important to consider [14].

As an alternative to using the analytical expressions based on weighted least squares or maximum likelihood, Baio et al. [15] proposed simulation-based power calculations. The strategy is to specify a complete model that represents the data generating procedure with flexible choices for random effects and then calculate power using data generated by the model coupled with the planned primary analysis strategy. Simulation methods are computationally intense yet totally flexible and may be used with both cross-sectional and cohort study designs.

However, little research has been done on power calculation for non-normal responses such as binary and count outcomes when these are modeled on their natural regression scale such as logit or log, respectively. Rather, such data are commonly treated using linear model methods for the SWD which implicitly is either moment-based or assumes approximately normally distributed data. As a result, SWD power calculations for binary data are typically conducted in terms of risk differences or rate differences [11, 13]. However, it may be preferable to model the outcomes on the natural and unconstrained scale of interest particularly for the adoption of flexible multilevel models which can characterize multiple sources of heterogeneity. In the presence of fixed time effects (which are considered necessary for SWDs) or other fixed effects, contrasts such as risk differences generally cannot be translated to simple overall odds ratios associated with intervention due to the change of model scale. Existing methods for non-normal outcomes are limited to simulation-based power calculation strategies [15] and the exact maximum likelihood-based power calculation strategy [16]. Both of these approaches are computationally intensive and inhibit the exploration of a wide range of design configurations for a proposed study.

When the outcome is non-normal, a full maximum likelihood analysis for a GLMM based on the marginal distribution of the outcome in the observed data requires numerical or stochastic integration for the calculation of the log-likelihood. Breslow and Clayton [17] proposed a rigorous approximate inference method based on penalized quasi-likelihood (PQL). Dang et al. [18] use PQL approximations to propose sample size and power calculation based on GLMM with correlated binary outcomes. Similarly, Kapur et al. [19] considered sample size determination for longitudinal designs with binary response data using a two-level mixed effect logistic regression model. Amatya and Bhaumik [20] proposed a general methodology for sample size determination with hierarchical designs and their approach involves complex expressions that have to be solved iteratively using estimates of variance components. To facilitate the use of

PQL-based sample size calculation in SWDs with non-normal outcomes, we propose a sample size and power calculation formula for SWDs with normal or non-normal outcomes by simplifying the Laplace approximation of the covariance matrix of the estimated parameter of interest. The method is intuitive, requires minimal computational power, and allows for rapid dynamic updates of sample size calculations when different parameters or design options are of interest. To facilitate adoption of our methods, we also provide a function embedded in the R package "swCRTdesign" [21] for sample size and power calculation for multilevel stepped wedge designs.

This paper is structured as follows. In "Method" section, we introduce our proposed method and provide an analytical formula for power/sample size calculation. In "Results" section, we use simulation experiments to compare the variance and power calculated by the proposed method with those given by the computationally intensive MLE-based method through repeated simulations. In "Discussion" section, we apply the proposed method to two studies, a public health patient-delivered partner therapy study for STI treatment and prevention with one level of clustering, and a health care delivery study with two levels of clustering that assesses the impact of providing additional benchmark prevalence information with a spine imaging report. Finally, we discuss the scope of application of the proposed method in "Appendix" section.

## Method

Denote the outcome for $n_j$ observations from a given cluster $j$ as $\boldsymbol{Y}_j^{n_j\times 1} = (Y_{1j},\ldots,Y_{n_jj})$, the design matrix for fixed effects as $\boldsymbol{X}_j^{n_j\times p}$, the coefficient vector for the fixed effects is denoted as $\boldsymbol{\beta}$, random effects as $\boldsymbol{b}_j^{q\times 1}$ with a design matrix $\boldsymbol{Z}_j^{n_j\times q}$, and $g = h^{-1}$ as the link function. Suppose the mean and variance of the outcome take the following flexible GLMM form:

$$E\left[\boldsymbol{Y}_j \mid \boldsymbol{b}_j\right] \equiv \boldsymbol{\mu}_j^{\boldsymbol{b}_j} = h\left(\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{b}_j\right) \equiv h\left(\boldsymbol{\eta}_j^{\boldsymbol{b}_j}\right),$$

$$Var[\,Y_{ij} \mid \boldsymbol{b}_j] = \phi a_{ij} v(\mu_{ij}^{\boldsymbol{b}_j}), \tag{1}$$

where $\boldsymbol{b}_j \sim Normal(0,\boldsymbol{D})$, $\phi$ is a dispersion parameter, $a_{ij}$ is a known constant for each observation, and $v()$ is a variance function. $\phi$, $a_{ij}$, and $v()$ depend on the distribution of $\boldsymbol{Y}_j$ (see Table 1). The outcomes are conditionally independent given the random effects. We will assume that the $p^{th}$ column of $\boldsymbol{X}_j$ corresponds to the intervention effect. Thus, the parameter of interest is $\beta_p$.

Typically, in a stepped-wedge design, the fixed effects consist of (at least) fixed time effect(s) and a fixed intervention effect. Random effects may consist of a random intercept, a random time effect(s), and/or a random intervention effect [14]. Additional random effects may be included in a cohort design to further characterize repeated measures on individuals within a cluster.

**Table 1** Variance function values for selected distributions and links ([22])

| | $\boldsymbol{\phi}$ | $\boldsymbol{a}$ | $\boldsymbol{v(\mu)}$ | $\boldsymbol{g(\mu)}$ | $\boldsymbol{g'(\mu)}$ |
|---|---|---|---|---|---|
| Normal | $\sigma^2$ | 1 | 1 | $\mu$ | 1 |
| Bernoulli | 1 | 1 | $\mu(1-\mu)$ | $log(\frac{\mu}{1-\mu})$ | $\frac{1}{\mu(1-\mu)}$ |
| Poisson | 1 | $\frac{1}{m_i}$ | $\mu$ | $log(\mu)$ | $\frac{1}{\mu}$ |
| Binomial | 1 | $\frac{1}{m_i}$ | $\mu(1-\mu)$ | $log(\frac{\mu}{1-\mu})$ | $\frac{1}{\mu(1-\mu)}$ |

*The $\frac{1}{m_i}$ indicates that the $i^{th}$ count is based on $m_i$ intervals or units; typically, $m_i = 1$.

### Variance approximation

Breslow and Clayton [17] use Laplace's method of integral approximation for marginalizing over the random effects in (1) to approximate the covariance matrix of the estimated parameter $\hat{\beta}$ by:

$$Var(\hat{\beta}) = \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}, \tag{2}$$

where

$$\boldsymbol{V} = \boldsymbol{W}^{\boldsymbol{b}} + \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T,$$

and $\boldsymbol{W}^{\boldsymbol{b}}$ denotes a diagonal matrix with entries $w_i^{\boldsymbol{b}} = \phi a_i v\left(\mu_i^{\boldsymbol{b}}\right)\left[g'\left(\mu_i^{\boldsymbol{b}}\right)\right]^2$, which depend on random effects $\boldsymbol{b}$. Here $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices for the fixed effect and the random effect for all observations across clusters. Breslow and Clayton [17] refer to their procedure as penalized quasi-likelihood or PQL. Note that PQL is an estimation strategy that focuses primarily on the regression parameters and the variance components; however, as part of the overall PQL approximation, individual random effects estimates, $\boldsymbol{b}$, are also available. To simplify the power calculation procedure with specified regression and variance component parameters, we propose the use of (2) with $\boldsymbol{b}$ set to their prior mean/mode of 0, that is, setting $\boldsymbol{W}^{\boldsymbol{b}} = \boldsymbol{W}^{\boldsymbol{0}}$.

For cluster designs, assuming clusters are independent, (2) may be rewritten as $Var(\hat{\beta}) = \left(\sum_j \boldsymbol{X}_j^T\boldsymbol{V}_j^{-1}\boldsymbol{X}_j\right)^{-1}$, where $j$ is the index for cluster, and $\boldsymbol{X}_j$ is a $n_j \times p$ design matrix for cluster $j$ with $n_j$ observations. The terms on the right-hand side of (2) are computed separately for each cluster and the sum is over clusters. Our proposed variance estimator is theoretically well-justified ([17]) but relies on the essential PQL approximation and the plug-in value for random effects. The PQL approximation is exact when the outcome is normal with an identity link. For non-linear outcomes that we consider, extensive simulation evaluation is conducted below to detail operating properties of this strategy.

### Sample size and power calculation

For stepped-wedge clustered designs, the sample size is a combination of the number of clusters, the number of sequences, the number of time periods, and the number of individuals per cluster period. Power can be calculated given the sample size, or sample size may be computed given power by satisfying the equation ([23]):

$$Power = \Phi\left(\frac{|\beta_p| - Z_{1-\frac{\alpha}{2}}\sqrt{V_0(\hat{\beta}_p)}}{\sqrt{V_a(\hat{\beta}_p)}}\right), \quad (3)$$

where $\alpha$ is the (two-tailed) significance level, $\beta_p$ is the intervention effect under the alternative hypothesis, and $V_0\left(\hat{\beta}_p\right)$ and $V_a\left(\hat{\beta}_p\right)$ are the variances of the estimated parameter under the null and alternative hypotheses, respectively.

## Results

### Simulation

In this section we use simulation experiments to compare the variance and power calculated by the proposed method with those given by simulation-based variance and power computations.

#### *Simulation settings*

We simulate binary outcomes with a Bernoulli distribution and a logit link since this scenario is biomedically important and known to be a situation for which PQL estimation may not perform well. We use a standard cross-sectional SWD in which the number of sequences is one less than the total number of time points. Specifically, we generate data from a SWD with four time periods and three sequences. For simplicity, each treatment sequence is set to have the same number of clusters (see below), and each cluster has the same number of individuals. The methodology can be applied to general settings with unequal number of clusters/individuals. We consider two outcome models with different random effects. Outcome Model I consists of a random intercept and a random effect for the treatment at the cluster level. Outcome Model II consists of a random intercept and a random time effect at the cluster level. Fixed effects include time effects and the treatment effect. Denote the number of time periods as $N_t$. For cluster $j$, individual $i$, the data generating model takes the following form:

$$logit\left(Pr\left[Y_{ij} = 1 \mid \boldsymbol{b}_j\right]\right) = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{b}_j,$$

where

$$\boldsymbol{X}_{ij} = \left(\mathbf{1}\ \boldsymbol{X}_{time,j}\ X_{treatment,j}\right), \boldsymbol{\beta} = \left(\beta_0\ \boldsymbol{\beta}_{time}\ \beta_p\right),$$

where $X_{time,j}$ is a $N_t - 1$ vector reparametrizing time as dummy variables. For Outcome Model I, $\boldsymbol{Z}_{ij} = \left(\mathbf{1}\ X_{treatment,j}\right)$, $\boldsymbol{b}_j = \left(b_{cluster,j}, b_{treatment,j}\right) \sim Normal\left(\mathbf{0}, \boldsymbol{D}_1\right)$, $\boldsymbol{D}_1$ is a diagonal matrix with elements $(\sigma^2_{cluster}, \sigma^2_{treatment})$, assuming the two random effects are uncorrelated (the proposed method could also be implemented with correlated random effects). Similarly for Outcome Model II, $\boldsymbol{Z}_{ij} = \left(\mathbf{1}\ \boldsymbol{X}^*_{time,j}\right)$, $\boldsymbol{b}_j = \left(b_{cluster,j}, \boldsymbol{b}_{time,j}\right) \sim Normal\left(\mathbf{0}, \boldsymbol{D}_2\right)$, where $\boldsymbol{X}^*_{time,j}$ is a $N_t$ vector whose $n^{th}$ elements is 1 and the rest of the elements are 0 where $n_t$ is the time of observation for individual $i$, $\boldsymbol{b}_{time,j}$ is a vector of length $N_t$, and $\boldsymbol{D}_2$ is a diagonal matrix with elements $\left(\sigma^2_{cluster}, \boldsymbol{\sigma}^2_{time}\right)$, assuming the two random effects are uncorrelated.

Each simulated dataset is analyzed using a mixed model regression to generate maximum likelihood (ML) estimates (using function glmer() from "lme4" package). We compute the variance of the ML estimates over many simulations and compare this to the variance predicted by Eq. (2). We compute the ML-based power by simulating data under the alternative hypothesis and calculating the frequency of rejection, then compare it to the power computed using the predicted variance.

These comparisons are made under a range of scenarios, including ones for which the PQL approximation may perform poorly such as a small number of clusters, a high variance for the random treatment effect, small sample size within each cluster, and low prevalence.

Specifically, we consider 4, 8, or 12 clusters per sequence (so 12, 24, or 36 clusters in total, respectively), a low (0.03), moderate (0.12), or high (0.43) prevalence in the null effect group (log odds approximately equal to $-3.5$, $-2.0$, and $-0.28$, respectively), a cluster size of 20, 50, or 100, and an effect size of 0.2 (log odds ratio). On the log odds scale, the standard deviation of the random cluster effect is set to 0.05 and the standard deviation of the random treatment and time effects is set to be 0.05 or 0.1. The exact scenarios are given in Table 2.

For each scenario, the variance of the estimated treatment effect coefficient $\hat{\beta}_p$ is computed across 2000 simulated datasets for the ML estimate and compared to Eq. (2) for the proposed method.

#### *Simulation results*

The relative variance of the treatment effect coefficient $\beta_p$ for the proposed method compared to the MLE (as measured across simulations) is displayed in Fig. 1, under both the null and alternative hypotheses.

Figure 1 shows that in the presence of a random treatment effect, the relative variance between the ML-based method and the proposed method, under the alternative and the null hypotheses, is close to 1 in most scenarios except in the extreme case where the number of clusters is small (with a total of 12 independent clusters), the cluster size is small (with 20 subjects per cluster), and the prevalence is 0.03. In this extreme case, the variance estimate

**Table 2** Parameter setting for binomial outcomes (log odds scale)

| | Simulation | EPT Trial | LIRE Trial |
|---|---|---|---|
| # of sequence | 3 | 4 | 5 |
| # of time period ($N_t$) | 4 | 5 | 6 |
| # of cluster per sequence | 4, 8, 12 | 6 | 20 |
| # of pcp per cluster | NA | NA | 35 |
| Cluster size | 20, 50, 100 | ? | ? |
| Prevalence (roughly) | 0.03, 0.12, 0.43 | 0.08 | 0.19 |
| Fixed time effect $\beta_{time}$ | (0.1, 0.1, 0.1) | (−0.008, −0.08, -0.17, −0.11) | −0.124 |
| Effect size for the intervention $\beta_p$ | 0.2 | −0.3 | −0.055 |
| SD of random cluster effect | 0.05 | 0.2 | 0.011 |
| SD of random treatment effect | 0.05, 0.1 | NA | 0.0054 |
| SD of random time effect | 0.05, 0.1 | 0.12 | NA |
| SD of random cluster (pcp) effect | NA | NA | 0.0015 |

*The cluster for LIRE Trial represents clinic, the random treatment effect is at clinic level, and time is modeled as a continuous variable
**SD stands for standard deviation

given by the proposed method is smaller than the actual variance of the MLE which could lead to an overestimate of power.

Relative estimates of power, comparing the proposed method and the MLE for situations when the power according to MLE is around 80% (achieved by varying the effect size), are displayed in Fig. 2. Here we target comparison at the common benchmark power of 80%. A relative power close to 1 is favorable for it indicates that the proposed method does not over- or under-estimate the power.

Note that the relative variance plots (Fig. 1) are not directly comparable to the relative power plot (Fig. 2) because the effect sizes (all greater than 1) in the latter are chosen such that the estimated power is approximately 80% for the MLE. As a result, the prevalence may not be low under the alternative hypothesis. Therefore, the poor performance of the proposed method in the extreme scenarios in Fig. 1 does not carry over to Fig. 2.

Figure 2 shows that when the estimated power calculated by MLE is around 80%, in the presence of random treatment effect or random time effect, the relative power between the ML-based method and the proposed method is close to 1 implying validity for use in sample size and power planning exercises.

## Applications

### *Application with a simple data structure: partner notification*

Patient delivered partner therapy (PDPT) is a partner notification strategy for individuals with sexually transmitted infections (STIs). Drugs or drug vouchers are given to patients with STIs to give to their sex partners.

The effectiveness of a PDPT-based partner notification strategy dubbed EPT (expedited partner therapy) was established by an individually randomized trial conducted in King County, Washington, between 1998 and 2003 for chlamydia and/or gonorrhea infection treatment.

EPT was then implemented in all counties in Washington between 2007 and 2009 through a cluster randomized trial using a stepped wedge design. County-based health districts in Washington state were randomized to EPT at one of four possible time periods with 5–6 districts at a time. Each time interval was 6–8 months. The prevalence of chlamydia was measured using cross-sectional sampling among women tested in family planning clinics for each county in each time interval.

The proposed model (see appendix 6.1) includes random cluster and time effects and we use the coefficient estimates from the final analysis of chlamydia to demonstrate the use of the proposed method in sample size calculation (Table 2). With an effect size of −0.3 (log odds ratio), prevalence of chlamydia must be measured in approximately 140 women in each cluster period in 24 counties to achieve a power of 80%.

### *Application with a complex data structure: Lumbar Imaging with Reporting of Epidemiology (LIRE)*

Incidental anatomic spine findings given by diagnostic imaging may lead to unnecessary additional tests and treatments among pain-free individuals. However, research has suggested that primary care patients were less likely to receive subsequent tests or medical interventions if the radiology report provides addition information on the prevalence of imaging finding among patients without back pain. Thus, Roland and van Tulder [24] proposed providing reference prevalence of various degenerative findings among patients without back pain in the spine imaging report to help reduce unnecessary medical attention.

A large, prospective stepped wedge cluster randomized control trial was designed to assess the impact of providing additional benchmark prevalence information in the imaging report.

A total of 100 primary care clinics from four large health systems were randomized to initiate the intervention at one of five possible times, each 6 months apart. The number of sequences, time periods, and the number of clusters per sequence are given by the study protocol ([25]). The

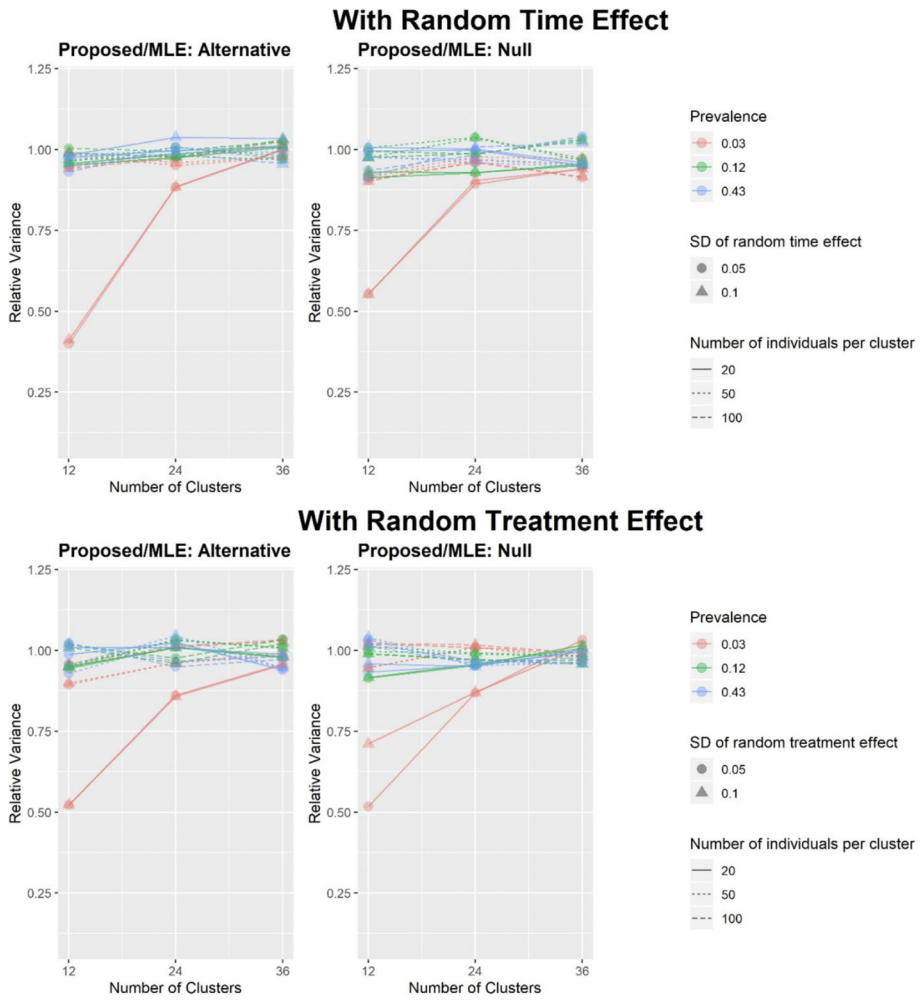


**Fig. 1** The relative estimated variance of $\hat{\beta}_p$ for the proposed method versus the MLE (variance measured across simulations), under both the alternative hypothesis ($\beta_p = 0.2$) and the null hypothesis ($\beta_p = 0$). The standard deviation of the treatment effect is on log odds scale

randomization was stratified by the clinic size. The size of the clinic was determined by the number of primary care providers (PCPs) and the site of the clinic. We examine the power calculation for the secondary outcome, the indicator of any opioid prescription within 90 days of the index imaging study. The random effect structure considered in our power calculation parallels those considered in the original power calculation for the primary outcome. The effect sizes and standard deviations of random effects for the power calculations we present are given by a model fit to the actual data from the trial (Table 2). The model is included in appendix 6.2.

We consider power calculation for LIRE to demonstrate the use of the proposed method for a problem with two-levels of clustering (clinic level and PCP level), which is computationally intensive for simulation-based methods and effectively impossible for existing exact methods. With an effect size of −0.055 (log odds ratio), Fig. 3 shows the relationship between the total number of clinics and the power. When outcomes are measured in 140, 175, or 210 patients per clinic-period, approximately 135, 160, and 200 clinics are needed to achieve a power of 80%. Figure 3 shows the value of high-fidelity approximation methods since we can explore a wide range of design alternatives with a computationally feasible strategy. To illustrate the difference in computing burden between our method and the simulation-based method, we have calculated the time required for the two methods to explore the 60 different scenarios presented in Fig. 3 for the LIRE trial. The results are included in appendix 6.3.

## Discussion

Power/sample size calculations are difficult for clustered data when the outcome is non-normal. Specifically, use of a normal approximation may be poor and there are no analytical formulae for power under non-identity links.

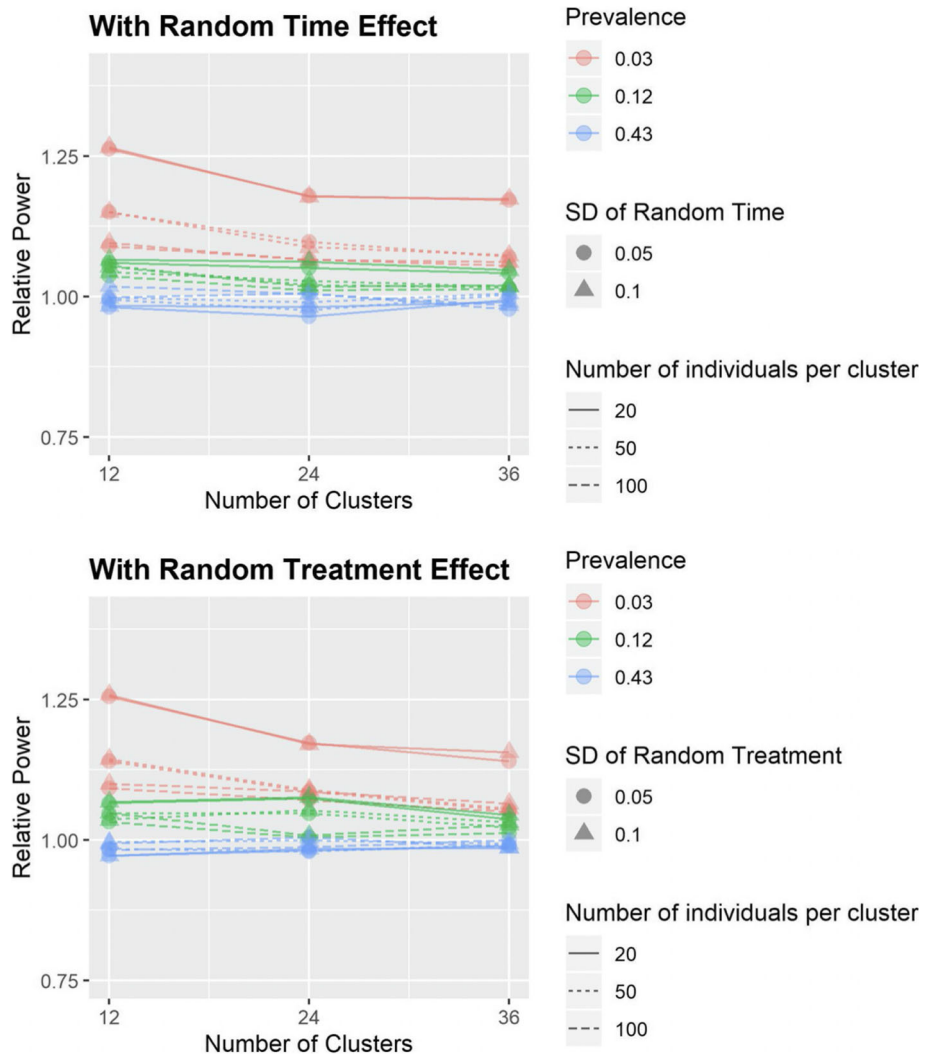


**Fig. 2** The relative estimated power (around 80%) of $\hat{\beta}_p$ for the proposed method versus the MLE (variance measured across simulations), under effect size of 0.2 (log odds ratio)

For stepped-wedge designs, the assumed time trend can affect power when the outcome model has a nonlinear link. Existing methods for non-normal outcomes, including simulation-based or exact ML-based power calculations, are computationally intensive.

In this paper, we propose a sample size and power calculation formulae for SWD with normal or non-normal outcomes using a Laplace approximation of the covariance matrix of the estimated parameter of interest. The method is fast computationally and has a good performance under non-extreme cases with a reasonable number of clusters, cluster sizes, and prevalence. This approach can be extended to any design as long as the outcome model can be specified in the form of a generalized linear mixed model (model 1), which includes cohort designs, incomplete designs, or exponential decay models [26].

By allowing one to compute power on the same scale as the intended analysis, the proposed approach can provide a more accurate estimate of study power. In addition, boundary issues (i.e., proportions outside the range 0–1) are avoided by working on canonical scales such as the logit or log. Of course, this also means that the variances of random effects must be specified on those same scales even though, at present, most published values for variance components are given for an identity link scale. Also, as noted previously, the assumed time trend may affect power when computed on nonlinear scales so greater attention must be given to this component during the design phase.

Implementing the proposed method involves taking the inverse of matrices $V_j$'s, which has dimension equal to the cluster sizes. This can take extended computing time

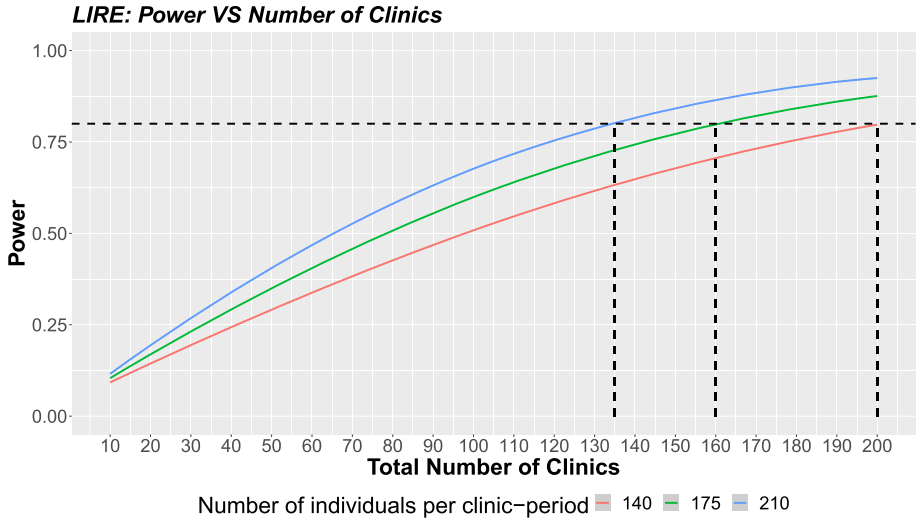


**Fig. 3** The relationship between the total number of clinics and the power with different number of individuals per clinic-period

with large clusters. One way to speed up the inversion is to use the Woodbury matrix identity [27], which requires inversion of a matrix of size equal to the dimension of the random effect vector $b_j$ instead. As long as the dimension of $b_j$ is smaller than the cluster size, the inversion is faster.

The extreme cases where the variance given by the proposed method does not reflect the actual cross-simulation variance of the MLE are characterized by settings with a combination of a small number of clusters, small sample sizes, and extreme prevalence. Indeed, in simulations, we also find that the mean of the variance estimates from ML estimation often does not match the true cross-simulation variance in these situations. These are likely to be cases in which other approximate power calculations (such as the existing closed-form formulas that approximate non-normal outcomes using the normal distribution) also perform poorly [16].

In summary, the proposed method provides a unified procedure for sample size calculations for stepped wedge design trials based on linear and generalized linear mixed models. The method is computationally fast and so allows for easy exploration of a variety of designs and parameter values. To make the greatest use of these methods, it is important that researchers reporting results from completed trials publish variance component values on the natural analytic scales (e.g., logit and log) for binary and count data.

## Appendix

### Model for the EPT Trial

$cluster : j = 1,\ldots,J; individual : i = 1,\ldots,I; time : t = 1,\ldots,T$

$E(\boldsymbol{Y}_j \mid \boldsymbol{b}_k) = h(\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{b}_j), \boldsymbol{\beta} = (\beta_0, \beta_{time}, \beta_{tx}), \boldsymbol{b}_j = (b_{j,0}, b_{1j}, \ldots, b_{Tj}), \boldsymbol{Y}_j = (Y_{1j}, \ldots, Y_{Ij})$,

$\boldsymbol{X}_j = (\mathbf{1}_I, (\boldsymbol{time}_{1j}, \ldots, \boldsymbol{time}_{Ij})^T, (tx_{1j}, \ldots, tx_{Ij})^T)$, $\boldsymbol{time}$ are $(T-1)$ vectors with 1 at exactly 1 place (reparametrizing time as dummy variables, -1 degrees of freedom).

$\boldsymbol{Z}_j = (\mathbf{1}_I, \boldsymbol{A})$, $\boldsymbol{A}$ is a $TI \times T$ matrix consists of time dummy variables.

### Model for the LIRE Trial

$clinic : k = 1,\ldots,K; pcp : j = 1,\ldots,J; individual : i = 1,\ldots,I; time : t = 1,\ldots,T$

$E(\boldsymbol{Y}_j \mid \boldsymbol{b}_k) = h(\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{b}_j), \boldsymbol{\beta} = (\beta_0, \beta_{time}, \beta_{tx}), \boldsymbol{b}_k = (b_{k,0}, b_{k,1}, b_{1k,0}, \ldots, b_{Jk,0})$,

$\boldsymbol{Y}_k = (Y_{11k}, Y_{21k}, \ldots, Y_{I1k}, Y_{12k}, \ldots, Y_{I1k}, \ldots, Y_{1Jk}, \ldots, Y_{IJk})$,

$\boldsymbol{X}_k = (\mathbf{1}_{IJ}, (time_{11k}, time_{21k}, \ldots, time_{I1k}, \ldots, time_{1Jk}, \ldots, time_{IJk})^T,$
$(tx_{11k}, tx_{21k}, \ldots, tx_{I1k}, \ldots, tx_{1Jk}, \ldots, tx_{IJk})^T)$,

$\boldsymbol{Z}_k = (\mathbf{1}_{IJ}, (tx_{11k}, tx_{21k}, \ldots, tx_{I1k}, \ldots, tx_{1Jk}, \ldots, tx_{IJk})^T, \boldsymbol{B})$,

$\boldsymbol{B} = diag(\mathbf{1}_I, \ldots, \mathbf{1}_I)_{IJ \times J}$.

### Computation time comparison

To better illustrate the difference in computing burden between our method and the simulation-based method, we have calculated the time required for the two methods to explore the 60 different scenarios presented in Fig. 3 for the LIRE trial in which multilevel clustering is present for non-normal outcomes (see Table 3).

To summarize, the table compares the calculation time for the simulation-based method with 1000 replicates and the proposed analytic method for a total of 60 scenarios using a single core of a 2.6 GHz Intel Core i7 processor. As expected, the simulation-based method's calculation time increases drastically as the number of individuals considered increases, but the calculation time of the analytic method is consistently short. The table shows the total

**Table 3** Calculation time for the 60 scenarios considered in the paper for the LIRE trial

| Number of individuals per clinic-period | Methods | Computing time (in days) |
|---|---|---|
| 140 | Analytic | 0.02 |
| | Simulation-based | 12.43 |
| 175 | Analytic | 0.04 |
| | Simulation-based | 16.11 |
| 210 | Analytic | 0.07 |
| | Simulation-based | 19.88 |

number of days taken for both methods to explore the 60 scenarios. It takes weeks for a simulation-based method to explore all scenarios, but the analytic method takes less than 2 h. In fact, exploring just one single scenario using the simulation-based method can take up to 2 days in this example.

Moreover, since the simulation-based methods rely on replicates that successfully converge without parameter estimates on the boundary, the actual time taken to get 1000 successful replicates could be much longer.

#### Abbreviations

CRT: cluster randomized trials; SWD: stepped wedge designs; LMM: linear mixed models; GLMM: generalized linear mixed models; PQL: penalized quasi-likelihood; MLE: maximum likelihood estimation; PDPT: patient delivered partner therapy; STI: sexually transmitted infections; EPT: expedited partner therapy; LIRE: lumbar imaging with reporting of epidemiology; PCP: primary care providers

#### Acknowledgements

Not applicable.

#### Authors' contributions

JPH and PJH conceived the original idea of the project. All of the authors designed the simulation experiments. FX performed the numerical simulations and took the lead in writing the manuscript. FX and ECV incorporated the methods into R. All authors provided critical feedback and contributed to the final version of the manuscript. All authors read and approved the final manuscript.

#### Funding

We gratefully acknowledge grants from National Institutes of Health (NIH) grant AI29168 (PI: James P Hughes), and Patient-Centered Outcomes Research Institute (PCORI) contract ME-1507-31750 (PI: Patrick J Heagerty). These funding sources had no role in the analyses and interpretation of the results.

#### Availability of data and materials

Not applicable.

## Declarations

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

#### Author details

[1]National Alzheimer's Coordinating Center, University of Washington, Seattle, WA, USA. [2]Department of Biostatistics, University of Washington, Seattle, WA, USA.

Received: 9 March 2021 Accepted: 13 August 2021
Published online: 06 September 2021

#### References

1. Gail MH, Mark SD, Carroll RJ, Green SB, Pee D. On design considerations and randomization-based inference for community intervention trials. Stat Med. 1996;15(11):1069–92.
2. Donner A, Klar N. Design and analysis of cluster randomization trials in health research. London: Arnold; 2000.
3. Mdege ND, Man M-S, Taylor CA, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. J Clin Epidemiol. 2011;64(9):936–48.
4. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. Bmj. 2015;350:h391.
5. Beard E, Lewis JJ, Copas A, Davey C, Osrin D, Baio G, Thompson JA, Fielding KL, Omar RZ, Ononge S, et al. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. Trials. 2015;16(1):353.
6. Martin J, Taljaard M, Girling A, Hemming K. Systematic review finds major deficiencies in sample size methodology and reporting for stepped-wedge cluster randomised trials. BMJ Open. 2016;6:2e010166.
7. Grayling MJ, Wason JM, Mander AP. Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. Trials. 2017;18(1):33.
8. Diggle P, Diggle PJ, Heagerty P, Liang K-Y, Heagerty PJ, Zeger S, et al. Analysis of longitudinal data. Oxford: Oxford University Press; 2002.
9. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials. 2007;28(2):182–91.
10. 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. J Clin Epidemiol. 2013;66(7):752–8.
11. Hooper R, Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. Bmj. 2015;350:h2925.
12. Hooper R, Teerenstra S, de Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Stat Med. 2016;35(26):4718–28.
13. Hemming K, Taljaard M. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. J Clin Epidemiol. 2016;69: 137–46.
14. Hughes JP, Granston TS, Heagerty PJ. Current issues in the design and analysis of stepped wedge trials. Contemp Clin Trials. 2015;45:55–60.
15. Baio G, Copas A, Ambler G, Hargreaves J, Beard E, Omar RZ. Sample size calculation for a stepped wedge trial. Trials. 2015;16(1):354.
16. Zhou X, Liao X, Kunz LM, Normand S-LT, Wang M, Spiegelman D. A maximum likelihood approach to power calculations for stepped wedge designs of binary outcomes. Biostatistics. 2020;21(1):102–21.
17. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc. 1993;88(421):9–25.
18. Dang Q, Mazumdar S, Houck PR. Sample size and power calculations based on generalized linear mixed models with correlated binary outcomes. Comput Methods Prog Biomed. 2008;91(2):122–7.
19. Kapur K, Bhaumik R, Charlene Tang X, Hur K, Reda DJ, Bhaumik DK. Sample size determination for longitudinal designs with binary response. Stat Med. 2014;33(22):3781–800.
20. Amatya A, Bhaumik DK. Sample size determination for multilevel hierarchical designs using generalized linear mixed models. Biometrics. 2018;74(2):673–84.
21. Voldal EC, Hakhu NR, Xia F, Heagerty PJ, Hughes JP. swCRTdesign: An R package for stepped wedge trial design and analysis. Comput Methods Prog Biomed. 2020;196:105514.
22. Nelder JA, Wedderburn RW. Generalized linear models. J R Stat Soc Ser A (General). 1972;135(3):370–84.
23. Rosner B. Fundamentals of biostatistics. Pacific Grove: Cengage learning; 2015.
24. Roland M, van Tulder M. Should radiologists change the way they report plain radiography of the spine? Lancet (British edition). 1998;352(9123): 229–30.
25. Jarvik JG, Comstock BA, James KT, Avins AL, Bresnahan BW, Deyo RA, Luetmer PH, Friedly JL, Meier EN, Cherkin DC, et al. Lumbar imaging with reporting of epidemiology (lire)–protocol for a pragmatic cluster randomized trial. Contemp Clin Trials. 2015;45:157–63.

26. Kasza J, Hemming K, Hooper R, Matthews J, Forbes A. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. Stat Methods Med Res. 2019;28(3):703–16.
27. Woodbury MA. Inverting modified matrices. Princeton: Memorandum Report 42, Statistical Research Group; 1950.

## Publisher's Note