The Annals of Applied Statistics
2017, Vol. 11, No. 1, 1–20
DOI: 10.1214/16-AOAS969

RANDOMIZATION INFERENCE FOR STEPPED-WEDGE CLUSTER-RANDOMIZED TRIALS: AN APPLICATION TO COMMUNITY-BASED HEALTH INSURANCE

BY XINYAO JI[*], GUNTHER FINK[†], PAUL JACOB ROBYN[‡] AND DYLAN S. SMALL[*]

University of Pennsylvania[*], Harvard University[†] and The World Bank[‡]

National health insurance schemes are generally impractical in low-income countries due to limited resources and low organizational capacity. In response to such obstacles, community-based health insurance (CBHI) schemes have emerged over the past 20 years. CBHIs are designed to reduce the financial burden generated by unanticipated treatment cost among individuals falling sick, and thus are expected to make health care more affordable. In this paper, we investigate whether CBHI schemes effectively protect individuals against large financial shocks using a stepped-wedge cluster-randomized design on data from a CBHI program rolled out in rural Burkina Faso. We investigate statistical properties of the stepped-wedge design following the parametric mixed model approach proposed by Hussey and Hughes in 2007. We find that testing for the treatment effect is generally sensitive to specification of the parametric model. For instance, if we fail to account for cluster-by-time interactions present in the data, the Type I error rate is severely inflated. We develop a more robust and efficient strategy—randomization inference. We demonstrate how to apply randomization inference to test for constant treatment effects and discuss test statistics suitable for the stepped-wedge design. Randomization inference guarantees a valid Type I error rate; simulation studies show that randomization inference test statistics also have power that is comparable to the currently used procedures that do not guarantee a valid Type I error rate. Finally, we apply our proposed method to the Burkina Faso CBHI dataset. We conclude that the insurance had limited effects on reducing the likelihood of low to moderate levels of catastrophic health expenditure, but substantially reduced the likelihood of extremely high health expenditure that exceeds half of a person's monthly income.

1. Introduction.

1.1. *Community-based health insurance.* The design of adequate health financing systems in low-income countries is a subject of significant debate. Due to low or modest economic growth, limited public tax resources and low organizational capacity, national health insurance schemes are generally impractical. In

Received March 2016; revised July 2016.
Key words and phrases. Randomization inference, stepped-wedge cluster-randomized trials, community-based health insurance.

response to such obstacles, community-based health insurance (CBHI) schemes, which are comparatively easier to set up, have emerged over the past 20 years [Asenso-Okyere et al. (1997), Devadasan et al. (2006), De Allegri et al. (2006), Ekman (2004), Wang et al. (2009)].

CBHI schemes are micro-insurance schemes that are voluntary, not-for-profit health insurance schemes organized at the community level. Under CBHI schemes, members of a community, often defined by geographical proximity or through employment-based relationships, pool resources in order to provide support for covering health expenditure [Robyn et al. (2012)]. CBHI schemes seek to reduce the financial burden generated by unanticipated treatment cost among individuals falling sick, and are thus expected to make health care more affordable. A natural question that emerges then is as follows: do CBHI schemes work as intended and in fact enhance universal financial protection?

We consider a study of a CBHI program in rural Burkina Faso that was implemented by the Ministry of Health and Nouna Health Research Center in collaboration with the University of Heidelberg, Institute of Public Health using a stepped-wedge cluster-randomized trial [De Allegri et al. (2008), Fink et al. (2013)]. We discuss properties of stepped-wedge cluster-randomized trials and problems with the currently used analysis methods for stepped-wedge cluster-randomized trials, present solutions to these problems, and analyze the study of the CBHI program in Burkina Faso.

1.2. *Stepped-wedge cluster-randomized trials*. A stepped-wedge cluster-randomized trial is a one-way crossover trial in which all clusters start out in the control and then clusters are randomized to cross over to the treatment at staggered times [Hall et al. (1987), Hussey and Hughes (2007)]. Figure 1 illustrates the treatment schedule for a stepped-wedge trial; the name "stepped-wedge" refers to the series of steps of the treatment schedule, which results in a wedge shape.

The stepped-wedge design has been gaining popularity in recent years because of a number of attractive features [Mdege et al. (2011)]. First, the stepped-wedge design is useful for settings in which limited resources or geographical constraints make it financially or logistically difficult to start the intervention in many clusters at once [e.g., Brown and Lilford (2006), Hall et al. (1987), Mdege et al. (2011), Moulton et al. (2007)]. For example, in a parallel design (randomize half the clusters to treatment during a single calendar period) or a traditional crossover design (randomize half the clusters to treatment at baseline and then switch these clusters to control and the other clusters to treatment midway through the trial), the intervention must be implemented in half of the clusters simultaneously, while the stepped-wedge design allows researchers to implement the intervention in a smaller fraction of clusters during each calendar period [Hussey and Hughes (2007)]. Second, the stepped-wedge design (as with a traditional crossover design) allows clusters to serve as their own controls, which increases power when there are substantial cluster effects [Woertman et al. (2013)]. The stepped-wedge design

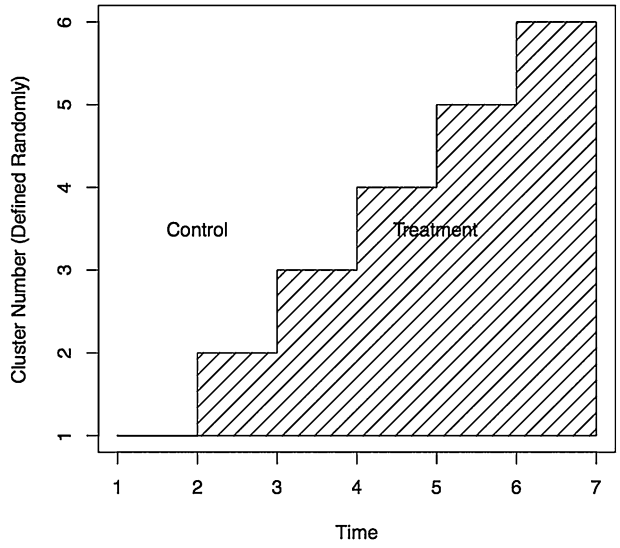


FIG. 1. *Illustration of a stepped-wedge design where different groups of clusters switch from control to treatment during different calendar periods.*

differs from a traditional crossover design, however, in that the crossovers are only in one direction; in particular, the intervention is never removed once it has been implemented (at least over the course of the trial). Third, because all clusters receive the treatment by the end of the trial and a cluster is never withdrawn from receiving the treatment, the stepped-wedge design is particularly useful for settings in which it would not be ethical, healthy or practical to withdraw the treatment or in which it would be difficult for participants to quickly revert to their pretreatment condition quickly after the withdrawal [Rhoda et al. (2011)]. The stepped-wedge design is also useful for evaluating the population-level impact of an intervention that has been shown to be effective in an individually randomized trial or for which there is a majority opinion that the intervention will be effective so that equipoise does not exist [Hussey and Hughes (2007)].

All these features made the stepped-wedge design ideal for studying the benefits of the CBHI program in Burkina Faso. Because the CBHI program was expected to confer benefits, every village in the study area wanted to be enrolled in the program at the early stage. However, it takes time to scale up the program, and so the CBHI management team and the health district had no option but to roll out the program in a progressive manner. The stepped-wedge design allowed the program to be rolled out in a fair manner and the effect of the program to be studied through a randomized trial. The stepped-wedge nature of the trial also helped to alleviate the spillover effect, as the incentive to migrate to a different area just to benefit from the intervention was counterbalanced by the fact that this very same intervention was going to be implemented in the entire study area within the next few years.

1.3. *Analysis methods*. In line with the increasing interest in employing and implementing the stepped-wedge design, a handful of pivotal articles on testing intervention effects, sample size calculations and analytical methods for continuous or dichotomous outcomes have emerged in the literature [e.g., Dimairo, Bradburn and Walters (2011), Hussey and Hughes (2007), Moulton et al. (2007), Woertman et al. (2013)]. Most of them have adopted the linear mixed model approach proposed by Hussey and Hughes (2007).

Hussey and Hughes (2007) considered the linear mixed model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + Z_{ij}\theta + e_{ijk}, \tag{1.1}$$

where Y_{ijk} is the observed response corresponding to individual k during calendar period j from cluster i and Z_{ij} denotes whether cluster i has been assigned the treatment by calendar period j. α_i is a random effect for cluster i such that α_i are i.i.d. $N(0, \tau^2)$, β_j is a fixed effect corresponding to time interval j (j in $1, \ldots, T-1$, $\beta_T = 0$ for identifiability), θ is the treatment effect and e_{ijk} are individual, time-period-specific effects that are assumed to be i.i.d. $N(0, \sigma_e^2)$ and independent of α_i.

One possible violation of assumptions in the linear mixed model (1.1) is the existence of cluster-by-time interactions, which are prevalent in a number of settings. For example, cluster-by-time interactions were a concern in a recent proposal for using the stepped-wedge design to study a vaccine for Ebola while the Ebola epidemic was going on because the Ebola epidemic, like other pandemics, was spreading from place to place over time [Bellan et al. (2015), van der Tweel and van der Graaf (2013)]. In the CBHI study we are considering, cluster-by-time interactions are a concern because the clusters are communities that are affected by different local economic and political developments.

Including all cluster-by-time interactions into the model as fixed effects would make the treatment effect unidentifiable. Hussey and Hughes (2007) proposed one strategy to deal with cluster-by-time interactions and still be able to estimate the treatment effect: create strata of clusters with similar expected time trends and then include stratum-by-time interaction as a factor in the model. This strategy requires some knowledge of the expected time trends before the trial and runs the risk of overfitting if interactions do not exist or are negligible. Without strong a priori knowledge of the pattern of cluster-by-time interactions, a better approach is needed to gauge the treatment effect than either excluding cluster-by-time interactions or including a specific pattern of them.

1.4. *Randomization inference*. In this paper, we develop another approach for the analysis of stepped-wedge cluster-randomized trials that accounts for potential cluster-by-time interactions—randomization inference. In randomization inference as developed by Fisher (1935), hypotheses are tested using only the assumption that the randomization has been properly carried out. Fisher said that randomization inference is "reasoned basis for inference" because it uses only the physical act of randomization as a basis for inference, and is exact and

distribution-free. Tukey (1993) said that randomization inference is the "platinum standard" inference. For discussion and examples of randomization inference, see Welch (1937), Raz (1990), Gail et al. (1996), Braun and Feng (2001), Rosenbaum (2002a, 2002b), Greevy et al. (2004), Ho and Imai (2006), Small, Ten Have and Rosenbaum (2008), Hansen and Bowers (2009).

Randomization inference can be applied to any test statistic of treatment effects. Here we consider Wald test statistics based on model (1.1) or other generalized linear mixed models. Because the randomization procedure adds an extra layer of security to the inference, the Type I error rate is valid even if parametric models for responses are misspecified such as failing to account for cluster-by-time interactions.

We contribute to the literature by applying randomization inference to stepped-wedge cluster-randomized trials. We build a unified framework to develop the randomization distribution for any test statistic, which can be used to calculate p-values and construct confidence intervals. Regarding our specific question, to what extent do CBHI schemes enhance universal financial protection, we use the data from the Burkina Faso study [Fink et al. (2013)] to examine whether CBHI schemes help to reduce the likelihood of catastrophic health expenditure. Our final results show that the insurance had limited effects on reducing the likelihood of low to moderate levels of catastrophic health expenditure, but substantially reduced the likelihood of extremely high health expenditure that exceeds half of a person's monthly income.

The outline of our paper is as follows. In Section 2, we introduce the potential outcomes notation and setup that will be used throughout the paper. In Section 3, we discuss consequences of failing to consider cluster-by-time interactions. In Section 4, we develop our randomization inference approach for the stepped-wedge design. In Section 5, we conduct simulation studies comparing the randomization inference approach to other analytical approaches for the stepped-wedge design. In Section 6, we apply randomization inference for stepped-wedge trials to a study of a community-based insurance program in rural Burkina Faso [Fink et al. (2013), Robyn et al. (2012)]. In Section 7, we provide a summary.

2. Notation and set up. There are I clusters, T calendar periods and n_{ij} individuals sampled from cluster i during calendar period j. $N = \sum_i^I \sum_j^T n_{ij}$ is the total number of observations in the study design. Let ijk index individual k in cluster i during calendar period j. An individual might be sampled at multiple time points; the indices $k = 1, \ldots, n_{ij}$ are time specific so that the same individual might have index k and $k' \neq k$ at different times. During calendar period j, m_j clusters are randomized to start treatment, where $m_1 + \cdots + m_T = I$, so that each cluster eventually starts treatment. $m_1, m_2, \ldots, m_T$ are prespecified before the start of the trial. Let Z_{ij} be the treatment corresponding to cluster i during calendar period j, where $Z_{ij} = 1$ for the active treatment and 0 for the

control. Since the trial is cluster-randomized, we index the treatment status for clusters rather than individuals. Let $\mathbf{Z}$ be the vector of all treatment assignments, $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots, Z_{IT})$. Write Ω for the set containing $|\Omega| = \binom{I}{m_1,\ldots,m_T}$ possible values $\mathbf{z}$ of $\mathbf{Z}$. Let Y_{ijk} be the observed response and $\mathbf{Y}$ be the vector of all observed responses, $\mathbf{Y} = (Y_{111}, Y_{112}, \ldots, Y_{ITn_{IT}})$. In case of a possible lag between the time of treatment assignments and the time that responses are observed, we assume that if individual k in cluster i enters the trial during calendar period j, so is assigned treatment Z_{ij}, then that individual will continue to receive treatment Z_{ij} until response Y_{ijk} is recorded. Each individual has a (row) vector of pretreatment covariates $\mathbf{X}_{ijk}$. $\mathbf{X}$ is the matrix whose rows are $\mathbf{X}_{ijk}$.

We define the causal effect of interest under the potential outcomes framework. We extend the notation of Speed (1990) and Rubin (1974) by representing each potential outcome as a function of the vector of all treatment assignments $\mathbf{z}$ [Rosenbaum (2007)]. Write $Y_{ijk}^{(\mathbf{z})}$ as the response that the kth individual in cluster i during calendar period j would have if the treatment assignment $\mathbf{Z} = \mathbf{z}$ for $\mathbf{z} \in \Omega$. $Y_{ijk}^{(\mathbf{z})}$ indicates that each individual has $|\Omega|$ possible outcomes, only one of which is observed, namely $Y_{ijk}^{(\mathbf{Z})}$. Fisher's sharp null hypothesis of no-treatment effect says that every unit would exhibit the same response under all treatment assignments, $Y_{ijk}^{(\mathbf{z})} = Y_{ijk}^{(\mathbf{z}')}$ for all $\mathbf{z}, \mathbf{z}' \in \Omega$. Under the alternative hypothesis, observed outcomes may exhibit arbitrary dependence.

We let $\mathcal{F} = \langle \mathcal{Y}, \mathbf{X} \rangle$, where $\mathcal{Y}$ is the unobserved array with N rows and $|\Omega|$ columns having entries $Y_{ijk}^{(\mathbf{z})}$. $\mathcal{F}$ does not change as the treatment assignments, $\mathbf{Z}$, change, whereas $\mathbf{Y}$ is a function of $\mathcal{F}$ and $\mathbf{Z}$, and so may change with $\mathbf{Z}$. To employ the cluster-randomized inference, as shown in Section 4, we assume the following assumptions hold for $\mathcal{F}$:

Assumption I: (a) there are no hidden variations of treatments and (b) $Y_{ijk}^{(\mathbf{z})} = Y_{ijk}^{(\mathbf{z}')}$ whenever $z_{ij} = z'_{ij}$. Assumption I(a) is part of the Stable Unit Treatment Value Assumption [Imbens and Rubin (2015), Rubin (1980)] and says that an individual receiving level $\mathbf{z}$ of the treatment cannot receive different forms of the treatment which have different effects. The assumption is implicit in the notation $Y_{ijk}^{(\mathbf{z})}$ which says that there is a single potential outcome for level $\mathbf{z}$ of the treatment. Assumption I(b) asserts that the potential outcomes would not be affected by treatment assignments in other clusters or subjects in different clusters do not interfere. Note that this assumption still allows for the possibility that units within a cluster at a given time interfere with each other. Assumption I(b) can be seen as a relaxation of the usual no interference part of the stable unit treatment value assumption (SUTVA) in the sense that a group of concentrated individuals are allowed to interfere with each other at a given time but interference is not allowed between groups or time points. This assumption also implies no carry-over effect, that is, a previous treatment for one subject does not affect later responses of this

same subject and also treatments for other subjects in the same cluster at previous times do not affect the response of the given subject at this time.

Assumption II: $Pr(\mathbf{Z}=\mathbf{z}|\mathcal{F}) = \frac{1}{|\Omega|} = \frac{1}{\binom{I}{m_1,\ldots,m_T}}$. This assumption says that the clusters are randomly assigned as to when to start treatment according to the stepped-wedge design, and the conditional distribution of treatment assignments given the potential responses and covariates is a fixed known constant. This assumption guarantees that tests derived solely from the randomization have the correct level whether or not potential responses within the same cluster are subject to interference [Fisher (1935), Welch (1937)].

Assumption III: If $\mathbf{z}$ and $\mathbf{z}'$ are the same except that $z_{ij}=1$ while $z'_{ij}=0$, then $Y_{ijk}^{(\mathbf{z})} - Y_{ijk}^{(\mathbf{z}')} = \theta$. This assumption implies that the treatment effect is constant across population and over time. By removing the treatment effect from the whole cluster during a calendar period, the observed responses would be the same as if there were no treatments assigned. This constant effect θ is the causal effect of interest.

3. The importance of cluster-by-time interactions. To motivate the need for accounting for cluster-by-time interactions, we assume that Y_{ijk} is generated by the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + Z_{ij}\theta + e_{ijk}. \tag{3.1}$$

For simplicity, we assume the e_{ijk} are independent, but correlation among the e_{ijk} (as might arise if individuals are observed multiple times) can be accommodated.

Both models (1.1) and (3.1) are observed data models that are consistent with Assumptions I and II. Compared to model (1.1), model (3.1) has an additional term γ_{ij} that accounts for cluster-by-time interactions. γ_{ij}'s are assumed to be i.i.d. $N(0,\eta^2)$ and independent of α and e. Using matrix notation, model (3.1) can be rewritten as

$$Y \sim N\big(M\Gamma, \Sigma = \sigma^2 I + \tau^2 A + \eta^2 B\big), \tag{3.2}$$

where $\mathbf{Y} = (Y_{111},\ldots,Y_{112},\ldots,Y_{ITn_{IT}})$, $\Gamma = (\mu,\beta_1,\ldots,\beta_T,\theta)^T$, and M is the $N\times(T+2)$ design matrix. Let Y_p denote the pth element in the vector $\mathbf{Y}$ which corresponds to a value of ijk. Then $M_{pq}=1$ if (1) $q=1$ or (2) $2\le j\le T+1$ and Y_p is observed during calendar period $q-1$ or (3) $q=T+2$ and Y_p is both observed and treated. $M_{pq}=0$ otherwise. A and B are symmetric positive definite matrices corresponding to cluster and cluster-by-time interactions, respectively:

$$A = \operatorname{diag}\big(\mathbf{1}_{n_{1\cdot}}\mathbf{1}_{n_{1\cdot}}^T,\ldots,\mathbf{1}_{n_{I\cdot}}\mathbf{1}_{n_{I\cdot}}^T\big), \tag{3.3}$$

$$B = \operatorname{diag}\big(\mathbf{1}_{n_{11}}\mathbf{1}_{n_{11}}^T,\mathbf{1}_{n_{12}}\mathbf{1}_{n_{12}}^T,\ldots,\mathbf{1}_{n_{IT}}\mathbf{1}_{n_{IT}}^T\big), \tag{3.4}$$

where $\mathbf{1}_{n1}$ denotes a column vector of 1's with length n_1 and $n_{i\cdot} = \sum_{j=1}^T n_{ij}$ is the size of cluster i.

Given σ^2, τ^2, η^2, the covariance matrix Σ is known. The best linear unbiased estimator of Γ is the Generalized Least Squares (GLS) estimator, which asymptotically has a normal distribution:

$$\hat{\Gamma}_{\mathrm{GLS}} = (M'\Sigma^{-1}M)^{-1}M'\Sigma^{-1}Y, \tag{3.5}$$

$$\hat{\Gamma}_{\mathrm{GLS}} \xrightarrow{d} N(\Gamma, (M'\Sigma^{-1}M)^{-1}). \tag{3.6}$$

If σ^2, τ^2, η^2 are not known, an implementable version of the GLS estimator is the Feasible Generalized Least Squares (FGLS) estimator, which requires a consistent estimate of Σ, say $\hat{\Sigma}$:

$$\hat{\Gamma}_{\mathrm{FGLS}} = (M'\hat{\Sigma}^{-1}M)^{-1}M'\hat{\Sigma}^{-1}Y. \tag{3.7}$$

One common strategy to find a consistent estimate $\hat{\Sigma}$ is to start by finding $\hat{\Gamma}_{\mathrm{OLS}}$ or another consistent (but inefficient) estimator, take the residuals from OLS to build a consistent estimator of the error covariance matrix Σ, update the FGLS estimation, and then apply the same idea iteratively until the estimators vary less than some tolerance. Under regularity conditions, such a FGLS estimator has the same asymptotic distribution as a GLS estimator:

$$\hat{\Gamma}_{\mathrm{FGLS}} \xrightarrow{d} N(\Gamma, (M'\Sigma^{-1}M)^{-1}). \tag{3.8}$$

For finite samples, the estimated covariance matrix of $\hat{\Gamma}_{\mathrm{FGLS}}$ is

$$\widehat{\mathrm{Var}}[\hat{\Gamma}] = (M'\hat{\Sigma}^{-1}M)^{-1}, \tag{3.9}$$

which converges to the asymptotic covariance matrix $(M'\Sigma^{-1}M)^{-1}$ given that $\hat{\Sigma}$ converges to Σ [Greene (2003)].

However, it is not always the case that we can find a consistent estimator of the covariance matrix Σ. The convergence of $\hat{\Sigma}$ to Σ relies on the correct specification of matrix structure and normality assumptions [Jacqmin-Gadda et al. (2007)]. In the process of iteratively computing $\hat{\Sigma}$, any deviation from the correct model would lead to an inconsistent version of $\hat{\Sigma}$. In particular, if we failed to account for cluster-by-time interactions in the case of stepped-wedge cluster-randomized trials, then we would specify the structure of the covariance matrix in a different form from the actual covariance matrix, that is, we would assume the consistent estimate of Σ to be $\hat{\Sigma} = \hat{\sigma}^2 I + \hat{\tau}^2 A$ while the actual covariance matrix is in the form of $\Sigma = \sigma^2 I + \tau^2 A + \eta^2 B$. Since B is a positive definite matrix as defined in (3.4), no values of $\hat{\sigma}^2$ and $\hat{\tau}^2$ would satisfy the equation $\hat{\sigma}^2 I + \hat{\tau}^2 A = \sigma^2 I + \tau^2 A + \eta^2 B$. Consequently, any computed $\hat{\Sigma}$ would be inconsistent, even if it maximizes the likelihood. Therefore, inferences based on $\hat{\Sigma}$ using the asymptotic distribution would be invalid.

We use a simulation study to examine this difference between the estimated variance of the treatment effect, which is the last diagonal element of $\hat{\Sigma}$, and the

Monte Carlo simulation of the true variance, which is the last diagonal element of Σ.

In the simulation, I and T are set to be 30 and 4, respectively. All clusters start with control at T = 1 and during each calendar period starting from $T = 2$, 10 clusters in the control group are randomly selected to be assigned to treatment. All clusters have equal size 100 and the true treatment effect $\theta = 0$. The magnitude of clustering is calibrated by the intracluster correlation coefficient (ICC), which is the proportion of the total variation explained by the respective blocking factor. In particular, the correlation between two randomly selected observations in the same cluster is

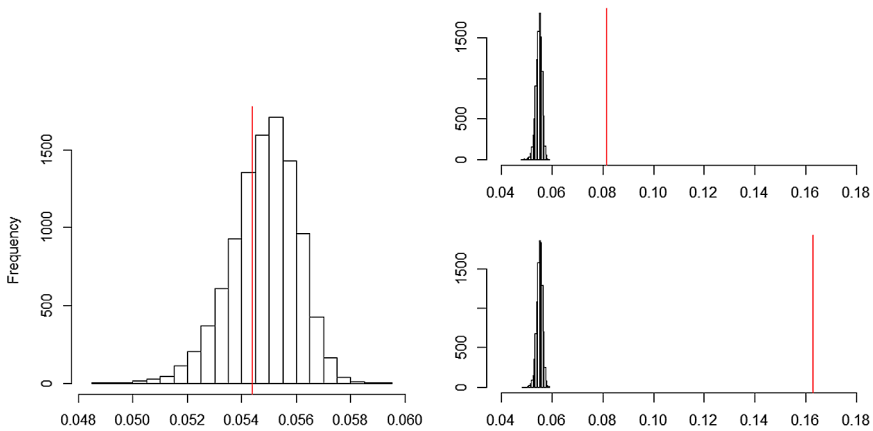
$$\text{ICC}_I = \frac{\tau^2}{\tau^2 + \eta^2 + \sigma^2}. \tag{3.10}$$

The correlation between two randomly selected observations in the same cluster and during the same calendar period is

$$\text{ICC}_{IT} = \frac{\tau^2 + \eta^2}{\tau^2 + \eta^2 + \sigma^2}. \tag{3.11}$$

As a result, the magnitude of interaction can be calibrated by $\text{ICC}_{IT} - \text{ICC}_I = \frac{\eta^2}{\tau^2+\eta^2+\sigma^2}$, which is the extra correlation from the same cluster and calendar period compared to just the cluster.

In Figure 2, we compare the distribution of estimated variances $\widehat{\text{Var}}[\hat{\theta}]$ over 10,000 simulations with the Monte Carlo simulation of the true variance. When



(a) No cluster-by-time interactions (b) With cluster-by-time interactions

FIG. 2. *Comparison of estimated and true variances of the treatment effect in different settings of cluster-by-time interactions. In* (a), *there are no cluster-by-time interactions*, $\text{ICC}_I := \frac{\tau^2}{\tau^2+\sigma^2} = 0.02$. *In* (b), *there are cluster-by-time interactions*, $\text{ICC}_I := \frac{\tau^2}{\tau^2+\sigma^2} = 0.02$, $\text{ICC}_{IT} := \frac{\tau^2+\eta^2}{\tau^2+\eta^2+\sigma^2} = 0.025$ (*upper*) *and* 0.04 (*lower*).

TABLE 1
Properties of the estimated treatment effect given by the Feasible Generalized Least Square estimator

$\dim(M)$	α	γ	ε	$E[\hat{\theta}]$	$\mathrm{Var}[\hat{\theta}]$	$E(\widehat{\mathrm{Var}}[\hat{\theta}])$	$\mathrm{SD}(\widehat{\mathrm{Var}}[\hat{\theta}])$
	$N(0,1)$	Zero	$N(0,49)$	-0.0031	0.0544	0.0548	0.0012
$I=30$	$N(0,1)$	Zero	$7/\sqrt{3}t(3)$	$-3.4e^{-5}$	0.0544	0.0545	0.0080
	$N(0,1)$	$N(0,0.25)$	$N(0,48.75)$	-0.0008	0.0816	0.0549	0.0012
$T=4$	$N(0,1)$	$N(0,0.5)$	$N(0,48.5)$	-0.0008	0.1083	0.0550	0.0012
	$N(0,1)$	$N(0,1)$	$N(0,48)$	-0.0008	0.1626	0.0552	0.0011

there are no cluster-by-time interactions, that is, model (1.1) is correctly specified, the left plot in Figure 2 indicates that the distribution of $\widehat{\mathrm{Var}}[\hat{\theta}]$ is centered around the true variance, marked by the red vertical line. However, when interactions do exist, the estimated variances are far off the true variance. The right plot describes two scenarios with different magnitudes of interactions. Neither of the distributions is close to the true variance.

Table 1 gives a more detailed summary of the estimated treatment effect $\hat{\theta}$ given by the FGLS estimator when the cluster-by-time interactions are not included in the model. As shown by column $E[\hat{\theta}]$, $\hat{\theta}$ is consistent in all settings. When there are no cluster-by-time interactions as shown by the first two rows, the average of the estimated variances $E(\widehat{\mathrm{Var}}[\hat{\theta}])$ is almost the same as the Monte Carlo simulation of the true variance $\mathrm{Var}[\hat{\theta}]$. But this is not the case when the interaction term γ is nonzero. The last column $\mathrm{SD}(\widehat{\mathrm{Var}}[\hat{\theta}])$ describes the dispersion of the estimated variances, which is of a much smaller order than its average.

The above simulation results show that fitting a linear mixed model for the stepped-wedge design while ignoring cluster-by-time interactions can lead to severely wrong standard errors, and this leads to poor control of Type I error rate, as shown by Table 2.

4. Randomization inference for stepped-wedge cluster-randomized trials. We would like to develop a strategy that accounts for cluster-by-time interactions if

TABLE 2
Type I error of linear mixed models not accounting for cluster-by-time interactions

$\dim(M)$	α	γ	ε	ICC_I	ICC_{IT}	Type I error
	$N(0,1)$	Zero	$N(0,49)$	0.02	0.02	0.052
$I=30$	$N(0,1)$	Zero	$7/\sqrt{3}t(3)$	0.02	0.02	0.054
	$N(0,1)$	$N(0,0.25)$	$N(0,48.75)$	0.02	0.025	0.511
$T=4$	$N(0,1)$	$N(0,0.5)$	$N(0,48.5)$	0.02	0.03	0.658
	$N(0,1)$	$N(0,1)$	$N(0,48)$	0.02	0.04	0.756

they exist. We will consider randomization inference. In randomization inference as developed by Fisher (1935), hypotheses are tested using only the assumption that the randomization has been properly carried out and randomization inference provides exact, distribution-free inferences. The significance level is always guaranteed regardless of the underlying mechanism that generates the data.

4.1. *A general setup.* There are I clusters and T calendar periods. At time t, m_t clusters are randomized to start treatment, where $m_1 + \cdots + m_T = I$, so that each cluster eventually starts treatment. Collect all possible values $\mathbf{z}$ of the treatment assignments $\mathbf{Z}$ in a set Ω, $|\Omega| = \binom{I}{m_1,\ldots,m_T}$. Because random numbers are used to assign which clusters start treatment at which times, $P(\mathbf{Z} = \mathbf{z}) = 1/|\Omega|$ for each $\mathbf{z} \in \Omega$.

Let $\mathbf{e}$ be a function of $\mathcal{F} = \langle \mathcal{Y}, \mathbf{X} \rangle$, and let $t(\mathbf{Z}, \mathbf{e}, \mathbf{X})$ be any function of $\mathbf{Z}, \mathbf{e}, \mathbf{X}$. Because $\mathbf{e}$ and $\mathbf{X}$ are functions of $\mathcal{F}$ and randomization ensures $P(\mathbf{Z}|\mathcal{F}) = 1/|\Omega|$, it follows that, for all v,

$$P\big(t(\mathbf{Z}, \mathbf{e}, \mathbf{X}) \geq v | \mathcal{F}\big) = \frac{|\{\mathbf{z} \in \Omega : t(\mathbf{z}, \mathbf{e}, \mathbf{X}) \geq v\}|}{|\Omega|}, \tag{4.1}$$

which is the randomization distribution of $t(\mathbf{Z}, \mathbf{e}, \mathbf{X})$. In words, given $\mathcal{F}$, the chance that $t(\mathbf{Z}, \mathbf{e}, \mathbf{X}) \geq v$ is simply the proportion of treatment assignments $\mathbf{z} \in \Omega$ such that $t(\mathbf{Z}, \mathbf{e}, \mathbf{X}) \geq v$. Moreover, (4.1) is the distribution of $t(\mathbf{Z}, \mathbf{e}, \mathbf{X})$ given $\mathcal{F}$ no matter what process produced $\mathcal{F}$. Fisher's (1935) description of randomization inference as the "reasoned basis for inference" refers to the fact that randomization creates the distribution (4.1) for every function $\mathbf{e}$ of $\mathcal{F}$ without further assumptions.

4.2. *Test of no effect.* The sharp null hypothesis of no effect asserts that the response of each individual is unchanged by receiving the treatment, $H_0 : \forall \mathbf{z}, \mathbf{z}' \in \Omega$, $Y_{ijk}^{(\mathbf{z})} = Y_{ijk}^{(\mathbf{z}')}$ for $i = 1, \ldots, I$, $t = 1, \ldots, T$, $k = 1, \ldots, n_{it}$, that is, $\mathbf{Y}^{(\mathbf{z})} = \mathbf{Y}^{(\mathbf{z}')}$. If H_0 were true, then randomization would label clusters treated or control but the observed outcomes would be unchanged. If H_0 were true, then the observed response $\mathbf{Y}^{(\mathbf{Z})}$ would equal $\mathbf{Y}^{(\mathbf{0})}$, a special case where all clusters are under control. Thus, under the null hypothesis of no treatment effect, the randomization distribution of $t(\mathbf{Z}, \mathbf{Y}, \mathbf{X}) = t(\mathbf{Z}, \mathbf{Y}^{(\mathbf{0})}, \mathbf{X})$ would be given by (4.1) with $\mathbf{e} = \mathbf{Y}^{(\mathbf{0})}$, where both $t(\mathbf{Z}, \mathbf{Y}^{(\mathbf{0})}, \mathbf{X})$ and its null distribution (4.1) would be calculated from the observed data when H_0 were true. For instance, in completely randomized experiments, Welch (1937) tested the null hypothesis of no effect using the randomization distribution of a test statistic suggested by analysis of variance and Raz (1990) used the randomization distribution of a test statistic that adjusted for $\mathbf{X}$ using a data smoother.

4.3. *Test of constant treatment effect.* The above method can be directly extended to test for constant treatment effect

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1.$$

Under the null hypothesis of $\theta = \theta_0$, $\mathbf{Y^{(0)}} = \mathbf{Y} - \mathbf{Z}\theta_0$. If $\mathbf{e} = \mathbf{Y^{(0)}}$, then $t(\mathbf{Z}, \mathbf{e}, \mathbf{X}) = t(\mathbf{Z}, \mathbf{Y} - \mathbf{Z}\theta_0, \mathbf{X}) = t'(\mathbf{Z}, \mathbf{Y}, \mathbf{X})$, where t' is a function on $\mathbf{Z}, \mathbf{Y}, \mathbf{X}$. This is to say $t'(\mathbf{Z}, \mathbf{Y}, \mathbf{X})$ would also have the randomization distribution given by (4.1).

Because of the randomization procedure, any function t on $\mathbf{Z}, \mathbf{Y^{(0)}}, \mathbf{X}$ is a valid test statistic with the Type I error controlled by the prespecified significance level. However, it does not mean that any arbitrarily chosen t is a good test statistic. We need to consider power. In the next section, we will discuss test statistics suitable for stepped-wedge cluster-randomized experiments.

4.4. *Wald randomization test*. A natural choice of t is the Wald statistic based on the maximum likelihood estimation of the treatment effect under model (1.1) or (3.1). Under the null hypothesis $H_0 : \theta = \theta_0$, $\mathbf{Y} - \mathbf{Z}\theta_0 = \mathbf{Y}^{(0)} = \mathbf{e}$. The maximum likelihood estimator of $L(\theta|\mathbf{Z}, \mathbf{Y}, \mathbf{X}) = L(\theta|\mathbf{Z}, \mathbf{e} + \mathbf{Z}\theta_0, \mathbf{X})$ is a function on $\mathbf{Z}$, $\mathbf{e}$ and $\mathbf{X}$. $t(\mathbf{Z}, \mathbf{e}, \mathbf{X})$ can be chosen as the Wald statistic of the null hypothesis $H_0 : \theta = \theta_0$ over the alternate hypothesis $H_1 : \theta \neq \theta_0$:

$$t(\mathbf{Z}, \mathbf{e}, \mathbf{X}) = \frac{(\hat{\theta} - \theta_0)^2}{\widehat{\mathrm{Var}}(\hat{\theta})}. \tag{4.2}$$

Instead of using its asymptotic distribution, which is a χ^2 distribution under the null hypothesis, the level is calculated using the randomization distribution given by (4.1). We can also investigate the power by randomly generating numerous data sets under a pre-specific alternative hypothesis. For each of these data sets, randomization inference is carried out and the evidence for or against the null hypothesis is recorded.

The Wald randomization test is applicable to a wide range of parametric models corresponding to different distributions of observed outcomes and can be implemented using standard functions in R, such as lmer() in the *lme4* package for linear mixed models, glm() for generalized linear models and censReg() in the *censReg* package for censored regression models.

4.5. *Other randomization tests*. Instead of calculating the maximum likelihood estimate and its standard deviation, other test statistics are available for stepped-wedge cluster-randomized trials. For example, because the design is essentially a two-way layout, we can first eliminate row and column effects by estimating their values or using the median polish method if robustness is a concern [Hoaglin, Mosteller and Tukey (2000)]. We then carry out the aligned rank test to compare the adjusted responses between clusters with different interventions [Sen (1968)]. If responses have heavy-tailed distributions, we may consider test statistics involving ranks to avoid bias caused by extreme values.

4.6. *Covariates adjustment*. The discussion in Sections 4.2 and 4.3 make no use of the covariates X, but it is straightforward to incorporate them, with no

change in the logic; see Rosenbaum (2002a). Instead of letting $\mathbf{e} = \mathbf{Y}^{(\mathbf{0})}$, $\mathbf{e}$ could also be a function on X. For example, $\mathbf{e}$ could be residuals when $\mathbf{Y}^{(\mathbf{0})}$is regressed on X by any method of regression. The randomization distribution of $t(\mathbf{Z}, \mathbf{e}, \mathbf{X})$ would still be given by (4.1).

5. Simulation study. We use a simulation study to investigate the level and the power of the Wald test statistic with usual asymptotic inference and with randomization inference in the stepped-wedge design. For demonstration purposes, we assume responses are normal and continuous. In all simulation settings, $I = 30$ and $T = 4$. When $t = 0$, all clusters are in the control group. When $t = 1$, 10 out of 30 clusters are randomly selected to receive treatment. When $t = 2$, 10 out of the remaining 20 untreated clusters are randomly selected to receive treatment. When $t = 3$, all clusters are assigned to treatment. Cluster sizes are randomly sampled between 1000 and 2000 and fixed over time. The true treatment effect θ_0 is set to be 0 and the power is calculated under the alternative $\theta_1 = 0.25, 0.5, 1, 1.5, 2$. $\mathrm{ICC}_I = \frac{\tau^2}{\tau^2+\eta^2+\sigma^2}$ is the intracluster correlation coefficient corresponding to clusters. $\mathrm{ICC}_{IT} = \frac{\tau^2+\eta^2}{\tau^2+\eta^2+\sigma^2}$ is the intracluster correlation coefficient corresponding to both clusters and interactions. All numbers reported are the average over 1000 sets of randomly simulated data set.

We first examine the Type I error rate in several scenarios. $\mathrm{Wald}_{\mathrm{asy}}$ and $\mathrm{Wald}_{\mathrm{rand}}$ are obtained under model (1.1) with usual asymptotic inference and with randomization inference. $\mathrm{Wald}^*_{\mathrm{asy}}$ and $\mathrm{Wald}^*_{\mathrm{rand}}$ are obtained under model (3.1) with usual asymptotic inference and with randomization inference.

It can be seen from Table 3 that both randomization procedures $\mathrm{Wald}_{\mathrm{rand}}$ and $\mathrm{Wald}^*_{\mathrm{rand}}$ guarantee the correct Type I error rate in all settings. When the interaction γ is zero, the Type I error rate is well controlled by both tests with usual asymptotic inference. However, when γ has a stand normal distribution, which leads to a small intracluster correlation coefficient $\mathrm{ICC}_{IT} = 0.04$, the Type I error rate given by $\mathrm{Wald}_{\mathrm{asy}}$ is inflated to 0.315 and 0.342 when e follows a normal and a t distribu-

TABLE 3
Type I error rate of the Wald test statistic based on the asymptotic distribution and the randomization distribution

α	γ	e	$\mathrm{Wald}_{\mathrm{asy}}$	$\mathrm{Wald}_{\mathrm{rand}}$	$\mathrm{Wald}^*_{\mathrm{asy}}$	$\mathrm{Wald}^*_{\mathrm{rand}}$
$N(0, 1)$	Zero	$N(0, 49)$	0.045	0.061	0.044	0.061
$N(0, 1)$	Zero	$7/\sqrt{3}t(3)$	0.042	0.055	0.042	0.054
$N(0, 1)$	Zero	Cauchy	0.055	0.050	0.051	0.053
$N(0, 1)$	$N(0, 1)$	$N(0, 48)$	0.315	0.055	0.069	0.059
$N(0, 1)$	$N(0, 1)$	$4\sqrt{3}t(3)$	0.342	0.051	0.069	0.054
$N(0, 1)$	$N(0, 1)$	Cauchy	0.056	0.054	0.063	0.060

TABLE 4
Power of the Wald test statistic for linear mixed models based on the asymptotic distribution and the randomization distribution

θ_1	α	γ	e	ICC_I	ICC_{IT}	Wald_{asy}	$\text{Wald}_{\text{rand}}$	$\text{Wald}^*_{\text{asy}}$	$\text{Wald}^*_{\text{rand}}$
0.25	$N(0,1)$	Zero	$N(0,49)$	0.02	0.02	0.254	0.275	0.254	0.276
0.5	$N(0,1)$	Zero	$N(0,49)$	0.02	0.02	0.723	0.715	0.721	0.725
1	$N(0,1)$	Zero	$N(0,49)$	0.02	0.02	0.999	0.999	0.999	0.999
1.5	$N(0,1)$	Zero	$N(0,49)$	0.02	0.02	1	1	1	1
2	$N(0,1)$	Zero	$N(0,49)$	0.02	0.02	1	1	1	1
0.25	$N(0,1)$	$N(0,1)$	$N(0,48)$	0.02	0.04	0.427	0.096	0.136	0.108
0.5	$N(0,1)$	$N(0,1)$	$N(0,48)$	0.02	0.04	0.636	0.253	0.335	0.277
1	$N(0,1)$	$N(0,1)$	$N(0,48)$	0.02	0.04	0.941	0.726	0.798	0.752
1.5	$N(0,1)$	$N(0,1)$	$N(0,48)$	0.02	0.04	0.999	0.969	0.982	0.975
2	$N(0,1)$	$N(0,1)$	$N(0,48)$	0.02	0.04	1	1	1	1
0.25	$N(0,1)$	Zero	$7/\sqrt{3}t(3)$	0.02	0.02	0.266	0.272	0.261	0.280
0.5	$N(0,1)$	Zero	$7/\sqrt{3}t(3)$	0.02	0.02	0.751	0.734	0.740	0.744
1	$N(0,1)$	Zero	$7/\sqrt{3}t(3)$	0.02	0.02	0.998	0.999	0.998	0.999
1.5	$N(0,1)$	Zero	$7/\sqrt{3}t(3)$	0.02	0.02	1	1	1	1
2	$N(0,1)$	Zero	$7/\sqrt{3}t(3)$	0.02	0.02	1	1	1	1
0.25	$N(0,1)$	$N(0,1)$	$4\sqrt{3}t(3)$	0.02	0.04	0.416	0.107	0.124	0.115
0.5	$N(0,1)$	$N(0,1)$	$4\sqrt{3}t(3)$	0.02	0.04	0.630	0.240	0.310	0.272
1	$N(0,1)$	$N(0,1)$	$4\sqrt{3}t(3)$	0.02	0.04	0.942	0.718	0.786	0.786
1.5	$N(0,1)$	$N(0,1)$	$4\sqrt{3}t(3)$	0.02	0.04	0.999	0.971	0.999	0.992
2	$N(0,1)$	$N(0,1)$	$4\sqrt{3}t(3)$	0.02	0.04	1	1	1	1

tion, respectively. The $\text{Wald}^*_{\text{asy}}$ test performs better than Wald_{asy} as it incorporates cluster-by-time interactions, but its Type I error rate is still slightly higher than its randomized version. Such a phenomenon disappears when e follows a Cauchy distribution. This might be explained by the fact that the Cauchy distribution is so heavy tailed that it dominates the small interaction term γ.

We next examine power. According to results in Table 4, when there are no cluster-by-time interactions, the randomization tests have comparable power with the tests using the asymptotic distribution. When there are cluster-by-time interactions, we ignore the power calculated from Wald_{asy} and $\text{Wald}^*_{\text{asy}}$ as the level is no longer valid, but only focus on their randomized versions, which give sufficient power to detect wrong values of the treatment effect.

We also carry out a set of simulations for dichotomous outcomes according to the model

$$\text{logit}\big(E(Y_{ijk})\big) = \mu + \alpha_i + \beta_j + \gamma_{ij} + Z_{ij}\theta. \tag{5.1}$$

Results are summarized in Table 5, showing similar advantages of using randomization inference for the stepped-wedge cluster-randomized trials.

TABLE 5
Power of the Wald test statistic for generalized linear mixed models based on the asymptotic distribution and the randomization distribution

θ_1	α	γ	e	ICC_I	ICC_{IT}	Wald_{asy}	$\text{Wald}_{\text{rand}}$	$\text{Wald}^*_{\text{asy}}$	$\text{Wald}^*_{\text{rand}}$
0	$N(0,1)$	Zero	$N(0,49)$	0.02	0.02	0.043	0.051	0.044	0.051
0.5	$N(0,1)$	Zero	$N(0,49)$	0.02	0.02	0.223	0.208	0.216	0.195
1	$N(0,1)$	Zero	$N(0,49)$	0.02	0.02	0.791	0.747	0.773	0.740
1.5	$N(0,1)$	Zero	$N(0,49)$	0.02	0.02	1	0.999	0.998	0.998
0	$N(0,1)$	$N(0,1)$	$N(0,48)$	0.02	0.04	0.217	0.060	0.091	0.048
0.5	$N(0,1)$	$N(0,1)$	$N(0,48)$	0.02	0.04	0.412	0.172	0.318	0.159
1	$N(0,1)$	$N(0,1)$	$N(0,48)$	0.02	0.04	0.726	0.448	0.544	0.377
1.5	$N(0,1)$	$N(0,1)$	$N(0,48)$	0.02	0.04	0.922	0.681	0.837	0.572

6. Application to study of community-based health insurance program.

6.1. *Background*. The Ministry of Health and Nouna Health Research Center in Nouna District, Burkina Faso implemented a CBHI scheme from 2004 to 2006 that aimed to make health care more affordable and to protect local communities from large health expenditure shocks [Fink et al. (2013), Robyn et al. (2012)]. To allow for a proper evaluation, the rollout of the program followed a stepped-wedge cluster-randomized design, enrolling randomly selected communities in three phases. In order to investigate the effect of CBHI schemes on household welfare, we follow Fink et al. (2013) to analyze the effect of CBHI schemes on catastrophic expenditure.

6.2. *Data*. The data we use is the Nouna Health and Demographic Surveillance Site (HDSS) survey data collected from 2003 to 2008. Data from year 2003 are the baseline prior to the intervention and data from years 2007 and 2008 are controls after the final rollout phase. There are 48 areas in the health district and each of them is considered a cluster. Due to residential mobility and migration, the study population is dynamic with an attrition rate of 59% from 2003 to 2008. There are 59,905 records in total and the number of individuals targeted by the insurance program in phase I, II and III are 27,696, 14,292 and 17,917, respectively. Equal mean test indicates that these three rollout groups have balanced covariates of age, gender, years of education, literacy, religion, marital status, household size and wealth index; see Table 4 in Fink et al. (2013).

Since the primary objective of CBHI schemes is to protect individuals against large financial shocks, we investigate the probabilities of facing health expenditure greater than 5%, 10%, 15%, 25% and 50% of monthly income. The catastrophic expenditure is a dichotomous outcome, which is coded as one if the total health expenditure is greater than a certain percentage of the monthly income. For example, the 2003 data suggest that about 10.4% of individuals faced health expenditure

TABLE 6
Distribution of catastrophic expenditure over time, Nouna HDSS Household Survey, 2003–2008

Year	Population size	Expenditure cutoff ≥5%	≥10%	≥15%	≥25%	≥50%
2003	7796	814	610	460	347	207
2004	8619	1037	716	577	361	191
2005	6875	1402	977	742	519	311
2006	10,712	925	576	481	306	224
2007	13,784	1316	939	690	377	211
2008	12,118	950	663	452	291	141

larger than 5% of their monthly income in the sample, and 2.7% of individuals had to cover health expenditure of more than half their monthly income. See Table 6 for a detailed year-by-year summary of the data.

6.3. *Model*. The models (1.1) and (3.1) assume the continuity of observed responses and the normality of random components. In our data, catastrophic health expenditure is binary, and so we use the generalized linear mixed model and then apply the Wald randomization test. In particular, we use P_{ijk} to denote the probability of facing catastrophic expenditure for individual k during calendar period j from cluster i; the observed response Y_{ijk} follows the model

$$\begin{aligned} Y_{ijk} &\sim \text{Bernoulli}(P_{ijk}), \\ \text{logit}(P_{ijk}) &= \mu + \alpha_i + \beta_j + Z_{ij}\theta + X_{ijk}^T\gamma + e_{ijk}, \end{aligned} \tag{6.1}$$

where α_i, β_j, Z_{ij} and θ are defined the same as in model (1.1) and X_{ijk} is a vector of covariates that we adjust for, which are age, gender, years of education, literacy, religion, marital status, household size and wealth index. Because we have repeated observations on people and there might be unmeasured covariates not included in X_{ijk}, e_{ijk} could be correlated for $j \in \{1, 2, \ldots, T\}$. As a result, we include person-level random effects to allow for correlation between e_{ijk} and $e_{ij'k}$.

6.4. *Results*. We first investigate catastrophic expenditure that is greater than 5% of monthly income. We use the function lmer() from the package *lme4* to solve for the maximum likelihood estimate of θ in (6.1), which has mean value -0.3966 and standard deviation 0.0554. Hence, the Wald test statistic for the actual insurance rollout is 51.093 with p-value < 0.001, indicating that there is significant evidence that the CBHI insurance program helped to reduce the likelihood of facing health expenditure greater than 5% of monthly income. We then carry out the Wald randomization test by assuming that there was no such effect. The p-value given by (4.1) is 0.117, indicating that there is no strong evidence that insurance

TABLE 7
CBHI impact on catastrophic health expenditure based on generalized linear mixed models

Expenditure cutoff	p-value			
	Wald_{asy}	$\text{Wald}_{\text{rand}}$	$\text{Wald}^{*}_{\text{asy}}$	$\text{Wald}^{*}_{\text{rand}}$
≥5%	<0.001	0.117	0.135	0.115
≥10%	<0.001	0.339	0.351	0.331
≥15%	<0.001	0.431	0.463	0.427
≥25%	<0.001	0.442	0.422	0.410
≥50%	0.009	0.041	0.014	0.038

had an effect on the catastrophic expenditure. We also consider an expanded version of model (6.1) that includes cluster-by-time interactions:

$$
\begin{aligned}
Y_{ijk} &\sim \text{Bernoulli}(P_{ijk}),\\
\text{logit}(P_{ijk}) &= \mu + \alpha_i + \beta_j + \gamma_{ij} + Z_{ij}\theta + X_{ijk}^{T}\gamma + e_{ijk}.
\end{aligned}
\tag{6.2}
$$

The Wald statistic based on this model for the actual insurance rollout is 2.229 with p-value 0.135 and the Wald randomization test gives p-value 0.115.

We repeat the same analysis for expenditure cutoffs 10%, 15%, 25% and 50% and summarize results in Table 7. P-values in columns Wald_{asy} and $\text{Wald}_{\text{rand}}$ are obtained under model (6.1) with usual asymptotic inference and with randomization inference. P-values in columns $\text{Wald}^{*}_{\text{asy}}$ and $\text{Wald}^{*}_{\text{rand}}$ are obtained under model (6.2) with usual asymptotic inference and with randomization inference.

6.5. *Conclusion*. Based on randomization inference that controls the Type I error rate properly, there is no strong evidence that the CBHI program carried out in Nouna District, Burkina Faso affected catastrophic expenditure that are defined to be greater than 5%, 10%, 15% and 25% of monthly income. The CBHI program, however, conferred a large benefit to people facing extremely high health expenditure that exceeds half of their monthly income. We see discrepancy between results from model (6.1) and model (6.2) using asymptotic inference. The model (6.1) would conclude that the CBHI program substantially reduced the likelihood of all levels of catastrophic health expenditure, but model (6.2) would conclude so only for the 50% cutoff.

Table 7 suggests that conclusions given by the asymptotic inference and the randomization inference are consistent only for model (6.2), which is an indication of the presence of cluster-by-time interactions. If we failed to consider the cluster-by-time interactions, the standard asymptotic inference is likely to greatly overestimate the protective effects of the insurance program.

7. Summary. There is a lack of literature on the theoretical aspects of analyzing the stepped-wedge cluster-randomized trials. We focus on statistical properties of the stepped-wedge design following the linear mixed model approach proposed by Hussey and Hughes [Hussey and Hughes (2007)]. Our simulations raise a red flag about using model-based inference for stepped-wedge trials. Specifically, the results can be very sensitive to model misspecification. As a result, bias can be introduced by cluster-by-time interactions and any other violations of assumptions.

We thus propose a new approach to the analysis of stepped-wedge cluster-randomized trials—using randomization inference to test for constant interventions. We introduce a unified framework to develop the randomization distribution for any test statistic, which can be used to calculate p-values and construct confidence intervals. Simulations based on linear mixed models show that randomization inference always guarantees the valid Type I error rate and has power comparable to the usual asymptotic inference.

We demonstrate our method using the Burkina Faso CBHI dataset to investigate whether CBHI schemes protect individuals against large financial shocks. We conclude that the insurance had limited effects on reducing the likelihood of low to moderate levels of catastrophic health expenditure in the target areas, but substantially benefited people facing extremely high health expenditure that exceeds half of their monthly income.

We hope that this paper serves as a valuable contribution to the literature on statistical properties of stepped-wedge cluster-randomized trials and its practical implementation in health economics, education, public health and other fields in which cluster-randomized trials are of interest. Our goal in this paper is to emphasize the value of randomization inference for stepped-wedge cluster-randomized trials and provide methods for implementing such randomization inference. With a strong belief in a parametric model, one can make inferences and calculate power and sample size based on asymptotic distributions, but these inferences can be sensitive to the model; randomization inference can deliver similar power while the inferences remain valid regardless of whether the parametric model holds or not.

REFERENCES

ASENSO-OKYERE, W. K., OSEI-AKOTO, I., ANUM, A. and APPIAH, E. N. (1997). Willingness to pay for health insurance in a developing economy. A pilot study of the informal sector of Ghana using contingent valuation. *Health Policy* **42** 223–237.

BELLAN, S. E., PULLIAM, J. R., PEARSON, C. A., CHAMPREDON, D., FOX, S. J., SKRIP, L., GALVANI, A. P., GAMBHIR, M., LOPMAN, B. A., PORCO, T. C. et al. (2015). Statistical power and validity of Ebola vaccine trials in Sierra Leone: A simulation study of trial design and analysis. *Lancet*, *Infect. Dis.* **15** 703–710.

BRAUN, T. M. and FENG, Z. (2001). Optimal permutation tests for the analysis of group randomized trials. *J. Amer. Statist. Assoc.* **96** 1424–1432. MR1946587

BROWN, C. A. and LILFORD, R. J. (2006). The stepped wedge trial design: A systematic review. *BMC Med. Res. Methodol.* **6** 54.

DEVADASAN, N., RANSON, K., DAMME, W. V., ACHARYA, A. and CRIEL, B. (2006). The landscape of community health insurance in India: An overview based on 10 case studies. *Health Policy* **78** 224–234.

DE ALLEGRI, M. D., KOUYATÉ, B., BECHER, H., GBANGOU, A., POKHREL, S., SANON, M. and SAUERBORN, R. (2006). Understanding enrolment in community health insurance in sub-Saharan Africa: A population-based case-control study in rural Burkina Faso. *Bull. World Health Organ.* **84** 852–858.

DE ALLEGRI, M. D., POKHREL, S., BECHER, H., DONG, H., MANSMANN, U., KOUYATÉ, B., KYNAST-WOLF, G., GBANGOU, A., SANON, M., BRIDGES, J. and SAUERBORN, R. (2008). Step-wedge cluster-randomised community-based trials: An application to the study of the impact of community health insurance. *Health Res. Policy Syst.* **6** 10.

DIMAIRO, M., BRADBURN, M. and WALTERS, S. J. (2011). Sample size determination through power simulation; practical lessons from a stepped wedge cluster randomised trial (SW CRT). *Trials* **12 (Suppl 1)** A26.

EKMAN, B. (2004). Community-based health insurance in low-income countries: A systematic review of the evidence. *Health Policy Plan.* **19** 249–270.

FINK, G., ROBYN, P. J., SIÉ, A. and SAUERBORN, R. (2013). Does health insurance improve health? Evidence from a randomized community-based insurance rollout in rural Burkina Faso. *J. Health Econ.* **32** 1043–1056.

FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

GAIL, M. H., MARK, S. D., CARROLL, R. J., GREEN, S. B. and PEE, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Stat. Med.* **15** 1069–1092.

GREENE, W. H. (2003). *Econometric Analysis*. Pearson Education India.

GREEVY, R., SILBER, J. H., CNAAN, A. and ROSENBAUM, P. R. (2004). Randomization inference with imperfect compliance in the ACE-inhibitor after anthracycline randomized trial. *J. Amer. Statist. Assoc.* **99** 7–15. MR2061884

HALL, A. J., INSKIP, H. M., LOIK, F., DAY, N. E., O'CONOR, G., BOSCH, X. and MUIR, C. S. (1987). The Gambia hepatitis intervention study. *Cancer Res.* **47** 5782–5787.

HANSEN, B. B. and BOWERS, J. (2009). Attributing effects to a cluster-randomized get-out-the-vote campaign. *J. Amer. Statist. Assoc.* **104** 873–885. MR2562000

HO, D. E. and IMAI, K. (2006). Randomization inference with natural experiments: An analysis of ballot effects in the 2003 California recall election. *J. Amer. Statist. Assoc.* **101** 888–900. MR2324090

HOAGLIN, D. C., MOSTELLER, F. and TUKEY, J. W. (2000). *Understanding Robust and Exploratory Data Analysis*. Wiley, New York. MR1800901

HUSSEY, M. A. and HUGHES, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemp. Clin. Trials* **28** 182–191.

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951

JACQMIN-GADDA, H., SIBILLOT, S., PROUST, C., MOLINA, J.-M. and THIÉBAUT, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Comput. Statist. Data Anal.* **51** 5142–5154. MR2370713

MDEGE, N. D., MAN, M. S., TAYLOR, C. A. and TORGERSON, D. J. (2011). Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *J. Clin. Epidemiol.* **64** 936–948.

MOULTON, L. H., GOLUB, J. E., DUROVNI, B., CAVALCANTE, S. C., PACHECO, A. G., SARACENI, V., KING, B. and CHAISSON, R. E. (2007). Statistical design of THRio: A phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. *Clin. Trials* **4** 190–199.

NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. *Statist. Sci.* **5** 463–464.

RAZ, J. (1990). Testing for no effect when estimating a smooth function by nonparametric regression: A randomization approach. *J. Amer. Statist. Assoc.* **85** 132–138. MR1137359

RHODA, D. A., MURRAY, D. M., ANDRIDGE, R. R., PENNELL, M. L. and HADE, E. M. (2011). Studies with staggered starts: Multiple baseline designs and group-randomized trials. *Am. J. Publ. Health* **101** 2164–2169.

ROBYN, P. J., FINK, G., SIÉ, A. and SAUERBORN, R. (2012). Health insurance and health-seeking behavior: Evidence from a randomized community-based insurance rollout in rural Burkina Faso. *Soc. Sci. Med.* **75** 595–603.

ROSENBAUM, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. MR1962487

ROSENBAUM, P. R. (2002b). *Observational Studies*. Springer, New York.

ROSENBAUM, P. R. (2007). Interference between units in randomized experiments. *J. Amer. Statist. Assoc.* **102** 191–200. MR2345537

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *J. Educ. Psychol.* **66** 688–701.

RUBIN, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.

SEN, P. K. (1968). On a class of aligned rank order tests in two-way layouts. *Ann. Math. Stat.* **39** 1115–1124. MR0226774

SMALL, D. S., TEN HAVE, T. R. and ROSENBAUM, P. R. (2008). Randomization inference in a group-randomized trial of treatments for depression: Covariate adjustment, noncompliance, and quantile effects. *J. Amer. Statist. Assoc.* **103** 271–279. MR2420232

TUKEY, J. W. (1993). Tightening the clinical trial. *Control. Clin. Trials* **14** 266–285.

VAN DER TWEEL, I. and VAN DER GRAAF, R. (2013). Issues in the use of stepped wedge cluster and alternative designs in the case of pandemics. *Am. J. Bioeth.* **13** 23–24.

WANG, H., YIP, W., ZHANG, L. and HSIAO, W. C. (2009). The impact of rural mutual health care on health status: Evaluation of a social experiment in rural China. *Health Econ.* **18** S65–S82.

WELCH, B. L. (1937). On the z-test in randomized blocks and latin squares. *Biometrika* **29** 21–52.

WOERTMAN, W., DE HOOP, E., MOERBEEK, M., ZUIDEMA, S. U., GERRITSEN, D. L. and TEERENSTRA, S. (2013). Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J. Clin. Epidemiol.* **66** 752–758.

X. JI
D. S. SMALL
DEPARTMENT OF STATISTICS
UNIVERSITY OF PENNSYLVANIA
3730 WALNUT STREET, SUITE 400
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: xinyaoji@wharton.upenn.edu
dsmall@wharton.upenn.edu

G. FINK
DEPARTMENT OF GLOBAL HEALTH AND POPULATION
HARVARD UNIVERSITY
665 HUNTINGTON AVENUE
BUILDING 1, ROOM 1110
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: gfink@hsph.harvard.edu

P. J. ROBYN
THE WORLD BANK
1818 H STREET NW
WASHINGTON, DC 20433
USA
E-MAIL: probyn@worldbank.org