Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Bias from cluster-specific time trajectories in cluster-randomized stepped wedge trials

Kevin Dobbin

kdobbin@augusta.edu

Augusta University

Research Article

Keywords:

Posted Date: April 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8978583/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Bias from cluster-specific time trajectories in cluster-randomized stepped wedge trials

Kevin K. Dobbin, Ph.D.
Assoc. Prof. of Biostatistics
Dept. of Biostatistics, Data Science and Epidemiology
Augusta University/Medical College of Georgia, Augusta GA, USA
kdobbin@augusta.edu

Abstract

Cluster-randomized stepped-wedge trials (SWT's) randomize each cluster to the time at which it crosses over from control to intervention. Commonly used models for SWT's assume that all clusters share a common time trajectory, meaning that the effect of time is the same across all clusters. But, the time effects may vary by cluster, resulting in cluster-specific time trajectories (CSTT's). This paper explores the bias induced in treatment effect estimatees is CSTT's exist but are ignored. A formula for estimating the bias is derived. To correct for the bias, two models with less restrictive assumptions are proposed that allow time trajectories to be cluster-specific, estimated using polynomials. It is shown using simulations that the bias formula is accurate, and that the proposed CSTT models reduce the bias. The bias seems particularly problematic when CSTT patterns are nonlinear. These methods are applied to two cluster-randomized SWT datasets.

1 Background and Motivation

The cluster-randomized stepped wedge trial (SWT) design can be visualized as a two-way grid (e.g., see Figure 1), where the columns represent time periods and the rows represent clusters (clinics, practices, etc.). A randomized intervention is rolled out across clusters over time. A common assumption of SWT models is that time effects are shared across clusters, so that each cluster has the same time pattern. But a more flexible model would allow each cluster to have its own time pattern. In other words, cluster-specific time trajectories (CSTT's) may be nested within cluster. This paper presents an analytic method to quantify the bias induced in intervention effect estimates when CSTT's are present, but not included in the model. Methods for detecting and adjusting for CSTT's that remove the bias are presented and evaluated. The paper focuses on cross-sectional SWT setting.

CSTT's can be caused by cluster-specific demographic shifts, seasonal trends, practice or policy changes, personnel changes, or changes in a clinic's catchment area over time. A well-known example of CSTT's is presented in the Prospective Randomized Amlodipine Survival Evaluation (PRAISE) trial, where survival outcomes in the placebo arm from PRAISE I and the later PRAISE 2 differed signficantly for the same target population (Friedman et al., 2015, Figure 5.1). Concern over the potential for CSTT's to introduce bias into intervention effect estimates motivated the development of permuted block designs (Zelen, 1974; Peto et al., 1976; Pocock, 1979). Permuted block designs ensure CSTT's cancel out of intervention effect estimates by balancing intervention arms within each time block of each cluster. This type of design is not possible in SWT cluster-randomized trials, because each time block in each cluster receives a uniform treatment assignment. Therefore, for SWT's, potential CSTT-induced bias must be addressed with analysis, rather than design, adjustments.

Common analysis methods for SWT's are based on fitting mixed models (Hussey and Hughes, 2007; Hemming et al., 2014; Li et al., 2021), combined with robust inference using generalized estimating equations, small-sample corrections, or Bayesian adjustment (Tong et al., 2024). For the bias studies in this paper, the Hussey/Hughes (H/H) and Hooper/Girling (H/G) models (Li et al., 2021) will be the focus. Both have a random cluster effect and a fixed time effect. The H/G model includes a random time-by-cluster interaction effect. Neither model permits systematic CSTT's.

This paper builds on previous work that suggested that CSTT's may be

problematic in SWT's. Nickless et al. (2018) graphed data from a longitudinal SWT that showed cluster-specific changes in means over time. More recently, Wu et al. (2025) were the first of which we are aware to propose a model that explicitly adjusts for CSTT's. Wu et al. (2025) presented a Bayesian approach to fitting SWT data, and used quadratic CSTT patterns in their simulations. Their model (their Equation (1)) uses cluster-specific spline curves to adjust for CSTT's. Subsequently, the impact of an intervention effect that varies over time (their Equation (2)) is added and a full Bayesian hierarchical penalized spline model for SWT data presented. However, they restricted attention to Bayesian methods, and did not derive results for bias in treatment effect estimates as a function of varying CSTT patterns.

Systematic reviews and guidelines for SWT have been published. Systematic reviews have identified hundreds of SWT studies (Barker et al., 2016; Grayling et al., 2017), despite early controversy about their advisability (Kotz et al., 2012). Literature reviews of SWT's have found their use in a wide range of intervention settings (Brown and Lilford, 2006; Barker et al., 2016; Grayling et al., 2017), but also note variation in data analysis methods and reporting quality (Brown and Lilford, 2006; Mdege et al., 2011; Davey et al., 2015; Barker et al., 2016; Twisk et al., 2016). General advice (Hemming et al., 2015) and CONSORT guidelines for SWT have been developed (Hemming et al., 2018; Hemming et al., 2019b), and adherence to them evaluated (Varghese et al., 2025; Hemming et al., 2019a).

Extensions of the basic H/H and H/G models have been explored, including permutation approaches. A modification of H/H and H/G models allow for exposure-time-dependent intervention effects (Kenny et al., 2022; Wu et al., 2025; Nickless et al., 2018). Other approaches use synthetic controls (Kennedy-Shaffer et al., 2020). Theoretical properties of various SWT estimators and estimation methods have also been evaluated (Wang et al., 2024). Applications to non-standard contexts such as mediation analysis are being developed (Cao and Li, 2025). Aside from mixed model approaches, permutation models and randomization-based inference have been developed (Wang and de Gruttola, 2017; Hughes et al., 2020; Thompson et al., 2016; Rabideau and Wang, 2021).

Study design and sample size for cluster-randomized SWT's are fairly well developed (Woertman et al., 2013; Girling and Hemming, 2016; Hooper et al., 2016; Matthews and Forbes, 2017). Kasza et al. (2019) used information calculations to evaluate SWT designs. Design guidance is also provided

(Copas et al., 2015; Taljaard et al. 2016; Hemming and Taljaard, 2020).

The central claim of this paper is that systematic CSTT's will induce bias in treatment effect estimates in models that do not ajust for CSTT's. The CSTT-induced bias is due to there being a relatively small number of clusters used in typical SWT designs, so that the arrangement of those clusters in the design (Figure 1) may induce bias if CSTT's are present. Models that explicitly adjust for systematic CSTT's can therefore reduce this potential source of bias. As shown in our formulas below, the bias is a function of both (1) the individual CSTT patterns and (2) their arrangement in the design.

The novel materials in this paper are (1) the focus on CSTT's as a potential source of bias in SWT designs, (2) the mathematical results for estimating bias in intervention effect estimates from cluster-randomized SWT's with CSTT's fit under the H/H model, (3) the presentation and evaluation of models that adjust for CSTT-induced bias, and (4) the empirical application of the methods to two SWT datasets.

This paper is structured as follows. In Section 2, common models for SWT data are reviewed, two models that allow for CSTT's introduced, and the bias formula and examples are presented. In Section 3, simulation studies are used to examine the adequacy of the formulas for approximating the treatment effect bias in Section 2, and to evaluate the capacity of the CSTT models to eliminate the bias. In Section 4, these methods are applied to two datasets, and results and application issues discussed. In Section 5, results are summarized.

cluster	Time					
1	C	T	T	T	T	T
2	C	T	T	T	T	T
3	C	T	T	T	T	T
4	C	T	T	T	T	T
5	C	C	T	T	T	T
6	C	C	T	T	T	T
7	C	C	T	T	T	T
8	C	C	T	T	T	T
9	C	C	C	T	T	T
10	C	C	C	T	T	T
11	C	C	C	T	T	T
12	C	C	C	T	T	T
13	C	C	C	C	T	T
14	C	C	C	C	T	T
15	C	C	C	C	T	T
16	C	C	C	C	T	T
17	C	C	C	C	C	T
18	C	C	C	C	C	T
19	C	C	C	C	C	T
20	C	C	C	C	C	T
Time	1	2	3	4	5	6

(a)

cluster	Time										
1	C	T	T	T	T	T	T	T	T	T	T
2	C	C	T	T	T	T	T	T	T	T	T
3	C	C	C	T	T	T	T	T	T	T	T
4	C	C	C	C	T	T	T	T	T	T	T
5	C	C	C	C	C	T	T	T	T	T	T
6	C	C	C	C	C	C	T	T	T	T	T
7	C	C	C	C	C	C	C	T	T	T	T
8	C	C	C	C	C	C	C	C	T	T	T
9	C	C	C	C	C	C	C	C	C	T	T
10	C	C	C	C	C	C	C	C	C	C	T
Time	1	2	3	4	5	6	7	8	9	10	11

(b)

Figure 1: Examples of cross-sectional stepped wedge designs. "C" indicates control condition and "T" active intervention condition. Left (a) is example from Hemming and Taljaard (2020). Right (b) is example from Copas et al. (2015).

2 Models and bias results

Table 1 presents models that will be discussed throughout the paper. Row 1 (CSTT-general model) of Table 1 is a flexible two-way model, with full main effects and interaction effects for the clusters and time blocks. This model allows for the widest range of systematic patterns for the CSTT's, and is used to generate simulation data. But, this model cannot be used to fit data from a SWT design. Rows 2 and 3 of the table present the H/H and H/G models, respectively. Both these models assume CSTT's either do not exist (H/H) or are non-systematic normal noise (H/G). Rows 4 and 5 of the table present two models being proposed for handling systematic CSTT's in data analysis, and are discussed more below. A key difference between the models on rows 2 and 3, versus the models on rows 4 and 5, is that the former treat time as categorical and the latter treat time as numerical.

2.1 Bias formula

Various theoretical results are shown in the supplemental material. In Theorem 1 it is shown that model "CSTT-general" on row 1 of Table 1 is not identifiable for data from an SWT design. In Theorem 2 it is shown that the same model is identifiable when an alternative design is used, so that the non-identifiability is a result of using the SWT design.

Theorem 3 of the Supplement addresses the GLM with identity link and continuous response. In Theorem 3, data are fit with the model on row 2 ("H/H model") of Table 1, that is,

$$Y_{ijk} = \mu + (\alpha)_i + \beta_j + \delta X_{ij} + \epsilon_{ijk}$$

where $(\alpha_i) \sim Normal(0, \tau^2)$ and $\epsilon_{ijk} \sim Normal(0, \sigma^2)$. The theorem shows that, if "CSTT-general" is the truth, then the generalized least squares (GLS) estimator of δ from the "H/H" model is biased. The derivation is outlined here. Suppose $B^{red.}$ is the vector of fixed effects for the model (where "red." superscript indicates the reduced dimension vector of fixed effects so the corresponding $X_{red.}$ is full rank (see, e.g., Hocking (1996)). Then the GLS estimator of $B^{red.}$ for the "H/H model" has the form:

$$\hat{B}^{red.}_{H/H,GLS} \quad = \quad (X^T_{red.} V^{-1} X_{red.})^{-1} X^T_{red.} V^{-1} \overline{Y}$$

where $\overline{Y}$ is a vector of cell means, and $V = Cov(\overline{Y})$. Specifically, the GLS

Model Name	Model Equation	Identifiable?
CSTT-general	$g(E[Y_{ijk}]) = \mu + (\alpha)_i + \beta_j + \alpha\beta_{ij} + \delta X_{ij}$	No
Hussey/Hughes (H/H)	$g(E[Y_{ijk}]) = \mu + (\alpha)_i + \beta_j + \delta X_{ij}$	Yes
Hooper/Girling (H/G)	$g(E[Y_{ijk}]) = \mu + (\alpha)_i + \beta_j + (\alpha\beta)_{ij} + \delta X_{ij}$	Yes
CSTT Random Poly.	$g(E[Y_{ijk}]) = \sum_{k=0}^{p} \eta_{0k} t_j^k + \sum_{k=0}^{p} (\eta)_{ik} t_j^k + \delta X_{ij}$	If $(p+1) \leq b$
CSTT Fixed Poly.	$g(E[Y_{ijk}]) = \mu + (\alpha)_i + \sum_{k=1}^{p} \eta_{ik} t_j^k + \delta X_{ij}$	If $(p+1) \leq b$

Table 1: Some generalized linear models (GLM's) for SWT data. g is the GLM link function. Random effects are in parentheses. Y_{ijk} is the response for individual k from cluster i at time point j, where $k = 1, ..., n$, $i = 1, ..., a$ and $j = 1, ..., b$. The (random) cluster effects are represented by $(\alpha)_i$. On row 1, β_j are fixed effects of cluster-specific time, and the $\alpha\beta_{ij}$ general interaction terms for CSTT's (with the usual constraints for identifiability). On rows 2 and 3, β_j are fixed common (across clusters) time effects. On row 3, the random interaction effect is $(\alpha\beta)_{ij}$. On all rows, X_{ij} is an indicator of intervention at time j in cluster i and the intervention effect is δ. On Row 4 and 5, the numerical values of the times are $t_j, j = 1, ..., b$. On row 4, $\eta_{0k}, k = 0, ..., p$ are fixed effects coefficients, and $(\eta)_{ik}$ are the cluster-specific random effect coefficients. On row 5, μ is the population intercept, $(\alpha)_i$ the random cluster intercept, and η_{ik} the fixed effects coefficients for each cluster.

estimator of $\hat{\delta}$ turns out to have the form

$$\hat{\delta}_{H/H,GLS} \quad = \quad \sum_{i=1}^{a} \sum_{j=1}^{b} w_{ij}(\tau^2, \sigma^2) \overline{Y}_{ij}$$

where $w_{ij}(\tau^2, \sigma^2)$ is a function, and $\overline{Y}_{ij}$ is the mean for cluster i at time j. The function $w_{ij}(\tau^2, \sigma^2)$ depends on the variance parameters and the design used (e.g., the function would differ for Figure 1(a) compared to Figure 1(b), even if the variance parameters were the same).

Next, note that if we know the CSTT-general parameters (model truth), then we know

$$\mu_{ij} \quad = \quad E\left[\overline{Y}_{ij}\right],$$

which allows us to subsequently calculate the bias

$$Bias\left(\hat{\delta}_{H/H,GLS}\right) \quad = \quad \left[\sum_{i=1}^{a} \sum_{j=1}^{b} w_{ij}(\tau^2, \sigma^2) \mu_{ij}\right] - \delta. \tag{1}$$

Moreover, since the GLS estimator is approximated by the restricted maximum likelihood estimator (REML), we propose the bias estimator,

$$\widehat{Bias}\left(\hat{\delta}_{H/H,REML}\right) \;\hat{=}\; Bias\left(\hat{\delta}_{H/H,GLS}\right)$$

which provides a means of estimating the bias for the fitted H/H model. Similar reasoning could be used to derive an estimate for the bias from the fitted "H/G model" of row 3 of Table 1. But since the bias from the "H/H model" and "H/G model" are generally quite similar, we omit this result.

2.1.1 Formula method examples

Figure 2 shows CSTT patterns along with the resulting bias from Equation 1. For (A)-(D) of the figure, the design in Figure 1 (a) is assumed. For (E)-(F) of Figure 2, designs and estimates are from the real datasets analyzed below. As expected, when the CSTT patterns are uniformly zero (corresponding to the H/H model), then the H/H bias is zero (Figure 2 A). Interestingly, the design is also unbiased when the CSTT's are each linear as in Figure 2 (B); it will generally be the case that the bias will tend to be small if the cluster-specific patterns are linear due to symmetry in the weight matrix used for estimating $\hat{\delta}_{H/H,GLS}$. Interested readers are referred to the Supplement. However, for nonlinear CSTT's the bias is non-zero (Figure 2 C and D). For the real datasets, CSTT patterns are necessarily estimated imperfectly using cell means because the true patterns are confounded with the treatment effect; however, based on this simplistic method of estimation, the CSTT's produce bias in both the SMARTHealth data (E) and the UBT data (F), which are described below in Section 4.

2.2 SWT Models that include CSTT adjustment

Table 1 presents models for SWT data that include CSTT adjustment. Row 4 of the table uses a separate random polynomial CSTT for each cluster, and row 5 of the table a separate fixed polynomials for each cluster with a random intercept. These both fit into the paradigm of generalized linear mixed models (GLMM's). GLMM models allow for a natural adjustment for systematic underlying cluster-specific time patterns that are unrelated to the treatment. The resulting models are polynomials with a single "jump" discontinuity when crossing from control to treatment. These models will be identifiable under the restrictions given in column 3 of the table. These two models assume the treatment effect is constant over exposure time, although

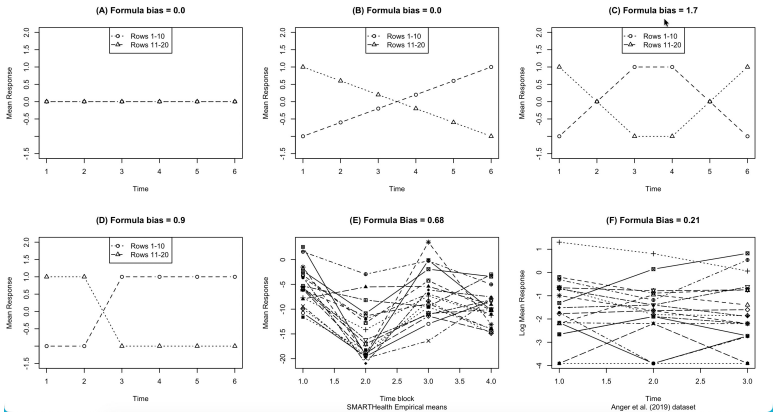


Figure 2: CSTT patterns and resulting bias estimates using formula described in paper. Figures (A)-(D) are simulation settings using the design in Figure 1(a), and (E) and (F) are estimates based on cell means from the two applied datasets studied in the paper. For figure (F), the y-axis corresponds to $Ln(x + 0.02)$ where x is the cell mean, in order to avoid taking the log of zero and because Poisson regression uses a log transformation.

one could also fold in changing exposure-time effects as, for example, in Wu et al. (2025) and Kenny et al. (2022).

2.2.1 Testing for CSTT effects

For the fixed polynomial model, one can test for CSTT's using the following nested models

$$H_0 \quad : \quad g(E[Y_{ijk}]) = \mu + (\alpha)_i + \sum_{k=1}^{p} \eta_k t_j^k + \delta X_{ij}$$

$$H_1 \quad : \quad g(E[Y_{ijk}]) = \mu + (\alpha)_i + \sum_{k=1}^{p} \eta_{ik} t_j^k + \delta X_{ij}$$

where the null hypothesis model assumes a common polynomial curve for all clusters and the alternative hypothesis model is from Row 5 of Table 1, with a separate polynomial for each cluster. The models can be compared with asymptotic chi-square likelihood tests, Akaike Information Criteria (AIC), or Bayesian Information Criteraia (BIC). A test statistic for comparing the models will be

$$-2 \cdot Ln\left(\frac{L(Model|H_0)}{L(Model|H_A)}\right) \quad \sim \quad \chi^2_{p\cdot(a-1)} \tag{2}$$

which is asymptotically chi-squared with $(a-1)\cdot p$ degrees of freedom. One rejects the null for large values of the statistic. We have assumed $b \geq (p+1)$.

Similarly, for the random polynomial model, one can test for CSTT's using the models

$$\begin{aligned} H_0 &: g(E[Y_{ijk}]) = \mu + (\alpha)_i + \sum_{k=1}^{p} \eta_k t_j^k + \delta X_{ij} \\ H_1 &: g(E[Y_{ijk}]) = \sum_{k=0}^{p} \eta_{0k} t_j^k + \sum_{k=0}^{p} (\eta)_{ik} t_j^k + \delta X_{ij} \end{aligned}$$

where the null hypothesis model assumes a common polynomial for the CSTT's (with random intercepts) and the alternative a separate set of random effect polynomial coefficients for each cluster. The models can be compared using AIC or BIC criteria.

3 Simulation Evaluation

Simulation studies were conducted to evaluate the bias formula in Equation 1, with respect to the H/H and H/G models of Rows 2 and 3 of Table 1. Also the performance of the polynomial methods of Table 1, Rows 4 and 5 with respect to reducing bias, are evaluated.

For the simulations, data were generated using the model of Table 1, Row 1. Simulated data were fit using models of Rows 2-5 of Table 1. Mixed models were fit using R Version 4.4.1, and the lme4 package (Version 1.1-34; Bates et al. 2015). To simplify presentation, R notation conventions are used in the tables in this section; specifically, "$a : b$" represents the sequence of integers from a to b (e.g., "1:4" is "1,2,3,4"), and "$rep(a, b)$" represents repeating the number a a total of b times (e.g., "rep(2,3)" is "2,2,2").

Table 2 Monte Carlo refers to Figure 1 (a), with 6 time-points and 20 clusters. Throughout this range of simulations, which includes different values for δ, different time patterns of the clusters, and different arrangements of the time patterns, the bias formula estimate is close to the true observed bias for both the Hussey/Hughes and Hooper/Girling models. This is seen by comparing the "Bias $\hat{\delta}$ Formula" column to the "Bias $\hat{\delta}$ H/H" column and the "Bias $\hat{\delta}$ H/G" column. One can also see that if the true $\delta = 0$ but our formula indicates bias is present, then the p-values for the intervention effects ("H/H $\hat{\delta}$ pval" column and "H/G $\hat{\delta}$ pval" column) tend to be (wrongly) signficant. This indicates that the bias can cause false positive findings about intervention effects. Similarly, the bias could also mask effects that are really present (not shown). The total sample size in these simulations is $20 * 6 * 10 = 1200$. Other simulation parameter values are given in the table caption.

Table 3 shows results for the fixed effect polynomial and random-effect polymomial models applied to data from the layout in Figure 1 (a). The CSTT patterns are given in columns 1 and 2, the row/cluster arrangement in column 3, the true δ in column 4, and the empirical bias of the $\hat{\delta}$ estimates from the H/H model in coulmn 5. The mean p-values from the chi-square tests of Equation 2 for CSTT's appear in the "p CSTT" column. For this column, the first two rows of the table have p-values that are correctly large, since simulation settings correspond to no interaction present. For all the other rows of the table, "p CSTT" values are close to zero, suggesting that the test has good power to detect the CSTT's in these cases. The $\hat{\delta}$ columns give the mean and standard deviation of the estimated intervention effects

Pattern 1	Pattern 2	Cluster Arrange	True δ	Bias $\hat{\delta}$ Formula	Bias $\hat{\delta}$ H/H	H/H $\hat{\delta}$ mean(SD)	H/G $\hat{\delta}$ mean(SD)	H/H $\hat{\delta}$ pval	H/G $\hat{\delta}$ pval
rep(0,6)	rep(0,6)	Any	0	0.0	5.4e-3	5.4e-3 (0.11)	5.4e-3 (0.11)	0.49	0.50
rep(0,6)	rep(0,6)	Any	2	0.0	0.00	2.00 (0.11)	2.00 (0.11)	<1e-6	<1e-6
0:2,2:0	2:0,0:2	1A	0	1.67	1.70	1.70 (0.10)	1.68 (0.10)	<1e-6	<1e-6
0:2,2:0	2:0,0:2	1A	2	1.67	1.71	3.71 (0.10)	3.69 (0.11)	<1e-6	<1e-6
0:2,2:0	2:0,0:2	1B	0	1.25	1.29	1.29 (0.10)	1.26 (0.11)	<1e-6	8e-5
0:2,2:0	2:0,0:2	1C	0	0.42	0.42	0.42 (0.10)	0.41 (0.12)	1.3e-2	0.22
0:2,2:0	2:0,0:2	1D	0	-0.42	-0.43	-0.43 (0.11)	-0.42 (0.12)	1.1e-2	0.21
0:2,2:0	2:0,0:2	1E	0	-1.25	-1.28	-1.28 (0.11)	-1.26 (0.11)	<1e-6	8e-5
1:4,4,4	4:1,1,1	1A	0	0.88	0.86	0.86 (0.11)	0.91 (0.14)	1e-5	0.04
1,1,rep(2,4)	2,2,rep(1,4)	1A	0	0.43	0.43	0.43 (0.10)	0.43 (0.10)	4e-3	0.06
1,1,rep(2,4)	2,2,rep(1,4)	1A	2	0.43	0.43	2.43 (0.10)	2.43 (0.11)	<1e-6	<1e-6
1,1,rep(2,4)	rep(2,4),1,1	1A	0	0.00	-2e-3	-2e-3 (0.10)	-1e-3 (0.10)	0.53	0.67

Table 2: Simulations to evaluate bias formula. Design layout is shown in Figure 1 (a). Pattern 1 and Pattern 2 are the two patterns of CSTT's. "Row/Cluster Arrange" is the row arrangement[a] (from top=1 to bottom=6). Column 4 δ is the intervention effect size truth. "Bias $\hat{\delta}$ Formula" is the bias approximation calculated from the formula. "H/H" is the mean(SD) of the Hussey/Hughes estimates of δ. "H/G" is the mean(SD) of the Hooper/Girling estimates of δ. "H/H $\hat{\delta}$pval" is the mean p-value for testing two-sided $H_0 : \delta = 0$ from the Hussey/Hughes model. "H/G $\hat{\delta}$pval" is the mean p-value for testing two-sided $H_0 : \delta = 0$ from the Hooper/Girling model. Settings are 1,000 Monte Carlo, 10 replicates per cell, 20 clusters, 4 clusters per cross-over, cluster main effect SD of 2, noise SD of 1, time main effects SD of 1.

[a] Arrange 1A is c(rep(1,10),rep(2,10)). Arrange 1B is c(rep(1,8),rep(2,10),rep(1,2)). Arrange 1C is c(rep(1,6),rep(2,10),rep(1,4)). Arrange 1D is c(rep(1,4),rep(2,10),rep(1,6)). Arrange 1E is c(rep(1,2),rep(2,10),rep(1,8)).

		Cluster	True	Bias $\hat{\delta}$	Polynomial Fixed				Polynomial Random		
Pattern 1	Pattern 2	Arrange	δ	H/H	p CSTT	$\hat{\delta}$	Bias	p $\hat{\delta}$	$\hat{\delta}$	Bias	p $\hat{\delta}$
rep(0,6)	rep(0,6)	Any	0	-1.2e-2	0.48	-5e-3 (0.15)	-5e-3	0.50	0.01 (0.10)	0.01	0.55
rep(0,6)	rep(0,6)	Any	2	0.00	0.87	2.00 (0.15)	0.00	<1e-6	2.00 (0.11)	0.00	<1e-6
0:2,2:0	2:0,0:2	1A	0	1.71	<1e-6	0.19 (0.15)	0.19	0.36	-0.01 (0.10)	-0.01	0.58
0:2,2:0	2:0,0:2	1A	2	1.71	<1e-6	2.18 (0.14)	0.18	<1e-6	1.99 (0.10)	-0.01	<1e-6
0:2,2:0	2:0,0:2	1B	0	1.28	<1e-6	0.22 (0.15)	0.22	0.29	0.25 (0.13)	0.25	0.30
0:2,2:0	2:0,0:2	1C	0	0.43	<1e-6	0.11 (0.14)	0.11	0.49	0.13 (0.14)	0.13	0.44
0:2,2:0	2:0,0:2	1D	0	-0.42	<1e-6	-0.14 (0.14)	-0.14	0.40	-0.17 (0.14)	-0.17	0.37
0:2,2:0	2:0,0:2	1E	0	-1.28	<1e-6	-0.24 (0.14)	-0.24	0.28	-0.26 (0.14)	-0.26	0.14
1:4,4,4	4:1,1,1	1A	0	0.86	<1e-6	0.13 (0.15)	0.13	0.49	0.11 (0.14)	0.11	0.37
1,1,rep(2,4)	2,2,rep(1,4)	1A	0	0.43	<1e-6	0.17 (0.15)	0.17	0.38	0.21 (0.14)	0.21	0.24
1,1,rep(2,4)	2,2,rep(1,4)	1A	2	0.43	<1e-6	2.17 (0.15)	0.17	<1e-6	2.20 (0.13)	0.20	<1e-6
1,1,rep(2,4)	rep(2,4),1,1	1A	0	2.3e-3	<1e-6	0.03 (0.15)	0.03	0.53	0.02 (0.13)	0.02	0.49

Table 3: Simulations to evaluate polynomial methods. Design is Figure 1(a). Pattern 1 and Pattern 2 are the two patterns of CSTT's (from left to right in Figure 1 (a)). "Row/Cluster Arrange" is the vertical row arrangement[a] for Figure 1 (a) (from top=1 to bottom=20). Column 4 δ is the intervention effect size truth. "Bias $\hat{\delta}$Formula" is the bias approximation calculated from the formula. "H/H Est." is the mean(SD) of the Hussey/Hughes estimates of δ. "Polynomial Fixed" are results of cluster-specific quadratic polynomials to quantitative time values; "p interact" is the mean interaction p-value test for CSTT's in polynomial models; "$\hat{\delta}$" is the mean and standard deviation of the intervention effect estiamtes, and "p $\hat{\delta}$" is the mean p-value for intervention. Similarly, 'Polynomial Random" are results for the random effect polynomial models; "p interact" is the mean interaction p-value test for CSTT's in spline models; "$\hat{\delta}$" is the mean and standard deviation of the intervention effect estimates, and "p $\hat{\delta}$" is the mean p-value for intervention in spline models. Settings are 500 Monte Carlo, 10 replicates per cell, 20 clusters, 4 clusters per cross-over, cluster main effect SD of 2, noise SD of 1, time main effects SD of 1.

[a] Arrange 1A is c(rep(1,10),rep(2,10)). Arrange 1B is c(rep(1,8),rep(2,10),rep(1,2)). Arrange 1C is c(rep(1,6),rep(2,10),rep(1,4)). Arrange 1D is c(rep(1,4),rep(2,10),rep(1,6)). Arrange 1E is c(rep(1,2),rep(2,10),rep(1,8)).

for the "Polynomial Fixed" and "Polynomial Random" models in rows 4 and 5 of Table 1. The corresponding biases of the $\hat{\delta}$ are given in the "Bias" columns. Comparing these biases to the H/H biases in column 5, it appears that both polynomials reduce the bias in all rows except the first two and the last (where the formula indicates no bias for H/H). The mean p-values for testing for a zero intervention effect are shown in the "p $\hat{\delta}$" columns. Here the values are very close to zero when the true effect size is 2, and relatively large when the true effect size is zero; ideally, these average p-values should be close to 0.50 when the true effect size is zero, but one sees that this is not always the case and sometimes they are shrunk towards zero. It isn't clear whether the polynomial fixed or random models are doing better.

Table 4 shows the simulation results corresponding to the design of Figure 1 (b). This design has more timepoints but fewer clusters per cross-over time than Table 3. In general, the performance of the polynomial methods are similar. On row 4 of the table, one can see that the polynomial fixed mean estimate of δ is somewhat biased (Bias -0.41) compared to the polynomial random model (Bias 0.00), although for both the bias is still much less than the Hussey/Hughes estimator (Bias -3.09). Finally, the mean p-values for the intervention effect for both the polynomial random and the polynomial fixed models are correctly small in the 3 rows where $\delta = 2$ is the truth (second, fourth and second from bottom row). The mean p-values in the other rows are relatively large although not uniformly centered at 0.50, as would be ideal under these null ($\delta = 0$) scenarios. This is perhaps not surprising since the patterns in columns 1 and 2 do not correspond to any functional polynomial form, so some residual bias may be expected.

4 Application and evaluation with real datasets

The methods and issues discussed in this paper are applied to data from two studies that used the SWT design.

4.1 Descriptions of the datasets

The SMARThealth India study (Peiris et al., 2019) was a SWT of a multi-faceted primary health care intervention for people assessed at high cardio-vascular disease risk in the state of Andhra Pradesh in India. This example focuses on the secondary endpoint of change in systolic blood pressure (SBP) from baseline to six month follow-up. This study utilized 18 clusters, consisting of primary health centers, randomized to three cross-over times.

		Row/Cluster	True	Bias $\hat{\delta}$	Emp.	Polynomial Fixed				Polynomial Random		
Pattern 1	Pattern 2	Arrange	δ	Formula	H/H $\hat{\delta}$	p CSTT	$\hat{\delta}$	Bias	p $\hat{\delta}$	$\hat{\delta}$	Bias	p $\hat{\delta}$
rep(0,11)	rep(0,11)	Any	0	0	0.017	0.90	-0.014 (0.14)	-0.014	0.62	0.01 (0.11)	0.01	0.64
rep(0,11)	rep(0,11)	Any	2	0	0.002	0.87	2.01 (0.15)	0.01	<1e-6	2.00 (0.11)	0.00	<1e-6
c(5:0,1:5)	c(0:5,4:0)	1A	0	-3.06	-3.09	<1e-6	-0.42 (0.15)	-0.42	0.09	-0.41 (0.14)	-0.41	0.07
c(5:0,1:5)	c(0:5,4:0)	1A	2	-3.06	-1.11	<1e-6	1.59 (0.15)	-0.41	<1e-6	2.00 (0.11)	0.00	3e-4
c(5:0,1:5)	c(0:5,4:0)	1B	0	-1.62	-1.66	<1e-6	-0.36 (0.15)	-0.36	0.14	-0.35 (0.13)	-0.35	0.09
c(5:0,1:5)	c(0:5,4:0)	1C	0	-0.83	-0.82	<1e-6	-0.13 (0.15)	-0.13	0.48	-0.11 (0.13)	-0.11	0.52
c(5:0,1:5)	c(0:5,4:0)	1D	0	0.83	0.84	<1e-6	0.15 (0.14)	0.15	0.40	0.19 (0.14)	0.19	0.37
c(5:0,1:5)	c(0:5,4:0)	1E	0	2.27	2.31	<1e-6	0.37 (0.14)	0.37	0.11	0.40 (0.14)	0.40	0.08
1:6,rep(6,5)	6:1,rep(1,5)	1A	0	1.61	1.57	<1e-6	0.22 (0.14)	0.22	0.36	0.22 (0.15)	0.22	0.23
1,1,1rep(2,8)	2,2,2,rep(1,8)	1A	0	0.43	0.43	<1e-6	0.08 (0.14)	0.08	0.55	0.13 (0.17)	0.13	0.41
1,1,1rep(2,8)	2,2,2,rep(1,8)	1A	2	0.43	0.43	<1e-6	2.10 (0.14)	0.10	<1e-6	2.15 (0.18)	0.15	2e-5
1,1,1rep(2,8)	rep(2,8),1,1,1	1A	0	0	0.014	<1e-6	4e-3 (0.14)	4e-3	0.61	0.03 (0.14)	0.03	0.47

Table 4: Simulations to evaluate polynomial methods. Design is Figure 1 (b). Pattern 1 and Pattern 2 are the two patterns of CSTT's. "Row/Cluster Arrange" is the row arrangement[a] . Column 4 δ is the intervention effect size. Formula is the bias approximation calculated from the formula. "H/H $\hat{\delta}$Est." is the mean(SD) of the Hussey/Hughes estimates of δ. "'Polynomial fixed" are results of cluster-specific quadratic polynomials to quantitative time values; "p interact" is the mean interaction pvalue test for CSTT's in polynomial models; "$\hat{\delta}$" is the mean and standard deviation of the intervention effect estiamtes, and "p $\hat{\delta}$" is the mean p-value for intervention. Similarly, "Polynomial Random"' are results of cluster-specific random-effect polymials ; "p interact" is the mean interaction pvalue test for CSTT's in polynomial models; "$\hat{\delta}$" is the mean and standard deviation of the intervention effect estimates, and "p $\hat{\delta}$" is the mean p-value for intervention in polynomial models. Settings are 500 Monte Carlo, 10 replicates per cell, 10 clusters, 1 clusters per cross-over, cluster main effect SD of 2, noise SD of 1, time main effects SD of 1.

[a] Arrange 1A is c(rep(1,5),rep(2,5)) where the numbers are pattern. Arrange 1B is c(rep(1,4),rep(2,5),rep(1,1)). Arrange 1C is c(rep(1,3),rep(2,5),rep(1,2)). Arrange 1D is c(rep(1,2),rep(2,5),rep(1,3)). Arrange 1E is c(rep(1,1),rep(2,5),rep(1,4)).

The Anger et al. (2019) dataset was a SWT of uterine balloon tamponade (UBT) intervention for women with postpartum haemorrhage in Uganda, Egypt and Senegal. There were 18 clusters. This study used two cross-over times. At each cross-over time, 9 randomized clusters crossed from control to intervention. Focus is on the secondary endpoint of number of blood transfusions. Because the response data are counts, data were fit using Poisson regression.

4.2 Computation and model details

Adding CSTT's to the H/H model has some risk because of increased model complexity. CSTT models require more care in fitting than H/H or H/G models. The H/H models are fit with the "lme4" package using the "lmer" and "glmer" functions for continuous and Poisson responses, respectively. To make inference robust, generalized estimating equations (GEE's) are fit using the "geeglm" function in the "geepack" package, with correlation specified as exchangeable.

For the CSTT models, we use generalized linear mixed models (GLMM's) and robust GLMM's for fitting. To ensure random effect coefficients models were stable, we fit using uncorrelated polynomial coefficients. For the mixed model the "lmer" function in "lme4" was used. For the robust fit, the "rlmer" function in the "robustlmm" package for the continuous response, and the "robmixglm" function in that same package for the Poisson response.

4.3 Testing for CSTT's

Model	Params	AIC	BIC	LL	-2 Log(L)	χ^2	DF	p-val
SMARTHealth								
Cubic No CSTT	7	82306	82355	-41146	82292			
Cubic CSTT	58	82271	82683	-41078	82155	136	51	1.1e-9
UBT								
Cubic No CSTT	5	1728	1754	-859	1718			
Cubic CSTT	39	1674	1871	-798	1596	123	34	5.7e-12

Table 5: Testing the nested models for both the SMARTHealth dataset an the UBT dataset. For both datasets, cubic polynomials for all clusters (No CSTT) were compared to a model with separate polynomials for each clusters (CSTT).

Table 5 shows the results of testing for cluster-specific time trends (CSTT's)

Model	Trt Eff Est	p-value
SMARTHealth dataset		
H/H lmer	1.188	0.218
H/H GEE	1.295	0.469
Cubic No CSTT	1.188	0.218
Cubic No CSTT rLMM	1.396	0.150
Cubic CSTT LMM	6.630	0.075
Cubic CSTT rLMM	6.390	0.091
UBT dataset		
H/H lmer	0.365	0.087
H/H GEE	0.324	0.223
Quad. No CSTT	0.365	0.089
Quad. No CSTT rLMM	0.332	0.176
Quad. CSTT LMM	0.363	0.220
Quad. CSTT rLMM	0.115	N/A

Table 6: Results of model fits for the SMARTHealth and UBT datasets. Column "Model" denotes the different models fitted. Column "Trt Eff Est" indicates the $\hat{\delta}$ treatment/intervention effect estimates. The "p-value" is the p-value corresponding to the $\hat{\delta}$ from the model output. For the SMARTHealth dataset, positive numbers indicate improvement in SBP outcomes in the treatment group compared to the control group.

in both datasets, using fixed effects polynomials. For both datasets, both the chi-square statistics reject the null hypothesis, indicating that CSTT's are present. Also, the Akaike Information Criteria (AIC's) find the CSTT models preferable in each case. The BIC criterion finds the "No CSTT" model optimal for both datasets. Overall, these results suggest that CSTT's may exist in these datasets. Visual depiction of the CSTT's appears in Figure 2.

4.4 Comparing results from H/H, H/G, and polynomial CSTT models

Table 6 compares the results of different model fits to these datasets. For the SMARTHealth datasets, the H/H model, the Cubic No CSTT model, and the Cubic No CSTT rLMM model all result in similar treatment effect estimates and p-values for the intervention effect. The CSTT models (lmer, GEE, and rLMM) result in larger effect size estimates and slightly smaller p-

values, but are overall consistent with one another. This may suggest bias in the H/H estimates. For the UBT dataset, here only quadratic polynomials could be fit because of the limited number of timepoints, and constraints in Table 1. The results of the H/H and no CSTT models are similar in terms of effect sizes and p-values. The CSTT models are similar in terms of effect size and p-values to the "No CSTT" models. This may be attributed to the linearity of the effects noted in the data plots and the observation that bias is minimal when CSTT's are linear.

5 Conclusion

This is the first paper to focus on the bias induced on treatment effect estimate if cluster-specific time trends (CSTT's) are present but not included in a model for cluster-randomized SWT data. The paper presented a closed form approximation to the bias based on a generalized least squares approximation to restricted maximum likelihood estimates. Extensive simulations showed that the bias exists and is well estimated by the approximation formula. To address the bias, we proposed polynomial CSTT models and showed that these reduced the bias in simulated datasets. Although the CSTT models are somewhat more complicated than traditional models, the advantage in terms of reducing potential bias seems to justify the complexity, particularly if the CSTT's effects are nonlinear.

There is a tradeoff between the increased model complexity when CSTT's are included in the model, and the reduction in bias. In very limited datasets or incomplete SWT designs where observations are not available in every cluster-by-time-block combination, the risk of model fitting errors may not be worth the potential reduction in bias of estimation. However, for complete SWT cluster-randomized trials with a reasonable number of observations in each cluster-by-time-block combination, the CSTT models potentially result in better treatment effect estimates. The CSTT models can be challenging to fit. For experienced data analysts, a reasonable course might be to fit both the H/H model and a CSTT model and, if the two differ, try to find the reason for the difference in terms of the CSTT patterns which may be visualized with interaction plots available in common software.

We examined both fixed effects polynomial coefficient CSTT's and random effects polynomial coefficient CSTT's for fitting the data. The fixed-effects models made testing for any CSTT's more straightforward, since the models are properly nested. But the random effect coefficients CSTT's fit somewhat

better into standard best practices. For random effect coefficients, it is recommended that the correlation of coefficients be set to zero to make the fitted model estimates more stable and reduce complexity. The two models (fixed and random effects) peformed very similarly in our simulations.

Another approach to fitting CSTT's is to fit separate splines to each cluster's time pattern. This was done in Wu et al. (2025). Splines can be more flexible than polynomials, and are less likely to veer off from the data. However, fitting splines falls less into the standard statistical paradigm. That, combined with the fact that SWT's generally have very few distinct timepoints for fitting, may make splines challenging to implement correctly in practice. More research is needed in evaluating the relative merits of splines and polynomials for removing bias.

Finally, there are some additional limitations to this work. First, this work has only studied cross-sectional SWT designs, and not longitudinal designs where some individual patients cross-over from the control to the intervention. Also, this work has focused primarily on the setting of a continuous response. If the response is binary or survival, it is not clear to what extent these results would apply. This is an area of future research. Also, several models in the literature allow for changing effects of the intervention based on length of exposure to the intervention condition; our models could be expanded to allow for this but this is beyond the scope of this paper, and would potentially further complicate model fitting. Finally, although the models proposed here show improvement on currently used models which ignore CSTT's, further refinement may lead to improvements in intervention effect estimation and inference.

6 Declarations

Acknowledgements
Not applicable.

Authors' contributions
KD wrote the manuscript, performed simulations, and analyzed the datasets.

Funding
Not applicable.

Data availability
No new data was collected and all data used in the paper are available in the Harvard Dataverse respository. Analysis code is available upon request.

Ethics approval and consent to participate
All judged in accordance with the Declaration of Helsinki.

Competing interests
The authors declare no competing interests.

7 References

1. Anger HA, Dabash R, Durocher J, Hassanein N, Ononge S, Frye LJ, Diop A, Beye SB, Burkhardt G, Darwish E, Ramadan MC, Kayaga J, Charles D, Gaye A, Eckardt M, Winikoff B. (2019) The effectiveness and safety of introducing condom-catheter uterine balloon tamponade for postpartum haemorrhage at secondary level hospitals in Uganda, Egypt and Senegal: a stepped wedge, cluster-randomised trial. BJOG. 126(13):1612-1621.

2. Barker D, McElduff P, D'Este C, Campbell MJ. (2016) Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. BMC Med Res Methodol. 16:69.

3. Bates D, Machler M, Bolker B and Walker S. (2015) Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software; 67: 1-48.

4. Brown CA and Lilford RJ (2006). The stepped wedge trial design: a systematic review. BMC Medical Research Methodology; 6: 54.

5. Cao Z and Li F (2025) Assessing mediation in cross-sectional stepped wedge cluster randomized designs. Statistics in Medicine, 44: e70175.

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

7. Davey C, Hargreaves J, Thompson JA, et al. (2015) Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. Trials, 16: 358.

8. Friedman LM, Furberg CD, DeMets DL, Reboussin DM and Granger CB (2015) Fundamentals of Clinical Trials, Fifth Edition. Springer, New York.

9. Girling AJ and Hemming K (2016) Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. Statistics in Medicine; 35(15): 2149-2166.

10. Grayling MJ, Wason JMS, Mander AP (2017) Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. Trials, 18:33.

11. Hemming K, Lilford R and Girling AJ (2014) Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. Statistics in Medicine, 34: 181-196.

12. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ (2015) The stepped wedge cluster randomised trial: Rationale, design, analysis and reporting. British Medical Journal, 351: h391.

13. Hemming K, et al. (2018) Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. British Medical Journal, 363: k1614.

14. Hemming K, et al. (2019a) Quality of stepped-wedge trial reporting can be reliably assessed using an updated CONSORT: crowd-sourcing systematic review. Journal of Clinical Epidemiology, 107: 77-88.

15. Hemming K, Taljaard M, Grimshaw J (2019b) Introducing the new CONSORT extension for stepped-wedge cluster randomised trials. Trials 20: 68.

16. Hemming K and Taljaard M. (2020) Reflection on modern methods: when is a stepped-wedge cluster randomized trial a good study design choice? International Journal of Epidemiology; 49: 1043-1052.

17. Hocking RR (1996) Methods and Applications of Linear Models: Regression and the Analysis of Variance. Wiley, New York.

18. Hooper R, Teerenstra S, de Hoopo E, et al. (2016). Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Statistics in Medicine; 35(26): 4718-4728.

19. Hughes JP, Heagerty PJ, Xia F, Ren Y (2020) Robust inference for the stepped wedge design. Biometrics, 76(1): 119-130.

20. Hussey MA and Hughes JP (2007). Design and analysis of stepped wedge cluster randomized trials. Contemporary Clinical Trials; 28: 182-191.

21. Kasza J, Taljaard M, Forbes AB (2019) Information content of cluster-period cells in stepped wedge trials. Biometrics, 75(1): 144-152.

22. Kenny A, Voldal EC, Xia F, Heagerty PJ, Hughes JP (2022) Analysis of stepped wedge cluster randomized trials in the presence of a time-varying treatment effect. Statistics in Medicine, 41: 4311-4339.

23. Kennedy-Shaffer L, de Gruttola V, Lipsitch M (2020) Novel methods for the analysis of stepped wedge cluster randomized trials. Statistics in Medicine, 39: 815-844.

24. Kotz D, Spigt M, Arts ICW, Crutzen R, Viechtbauer W (2012) Use of the stepped wedge design cannot be recommended: A critical appraisal and comparison with the classic cluster randomized controlled trial design. Journal of Clinical Epidemiology, 65: 1249-1252.

25. Li F, Hughes JP, Hemming K, Taljaard M, Melnick ER, Heagerty PJ (2021) Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: An overview. Statistical Methods in Medical Research, 30(2): 612-639.

26. Matthews JNS and Forbes AB (2017) Stepped wedge designs: insights from a design of experiments perspective. Statistics in Medicine, 36: 3772-3790.

27. Mdege ND, Man M, Taylor CA et al. (2011). Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. Journal of Clinical Epidemiology; 64: 936-948.

28. Nickless A, Voysey M, Geddes J, Yu LM, Fanshawe TR. (2018) Mixed effects approach to the analysis of the stepped wedge cluster randomised trial-Investigating the confounding effect of time through simulation. PLoS One, 13(12):e0208876.

29. Peiris, D (2018) Replication Data for: Smarthealth India. Harvard Dataverse, V1. https: //doi.org/10.7910/DVN/NSKFK2.

30. Peiris D, Praveen D, Mogulluru K, Ameer MA, Raghu A, Li Q, Heritier S, MacMahon S, Prabhakaran D, Clifford GD, Josshi R, Maulik P, Jan S, Tarassenko L and Patel A (2019) SMARThealth India: A stepped-wedge, cluster randomised controlled trial of a community health worker managed mobile health intervention for people assessed at high cardiovascular disease risk in rural India. PLOS One, 14(3): e0213708.

31. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG (1976) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer. 34(6):585-612.

32. Pocock SJ (1979) Allocation of patients to treatment in clinical trials. Biometrics, 35(1):183-97.

33. Rabideau DJ and Wang R (2021) Randomization-based inference for a marginal treatment effect in stepped wedge cluster randomized trials. Statistics in Medicine, 40(20), 4442-4456.

34. Taljaard M, Teerenstra S, Ivers NM, Fergusson DA (2016) Substantial risks associated with few clusters in cluster randomized stepped and stepped wedge designs. Clinical trials, 13(4): 459-463.

35. Thompson JA, Davey C, Fielding K, Hargreaves JR, Hayes RJ (2018) Robust analysis of stepped wedge trials using cluster-level summaries within periods. Statistics in Medicine, 38: 2487-2500.

36. Tong G, Nevins P, Taljaard M (2024) A review of current practice in the design and analysis of extremely small stepped-wedge cluster randomized trials. Clinical Trials, 22: 1.

37. Twisk JWR, Hoogendijk EO, Zwijsen SA, de Boer MR (2016) Different methods to analyze stepped wedge trial designs revealed different aspects of intervention effects. Journal of Clinical Epidemiology, 72: 75-83.

38. Varghese E, Briola A, Kennel T, Pooley A, Parker RA (2025) A systematic review of stepped wedge cluster randomized trial in high impact journals: assessing the design, rationale and analysis. Journal of Clinical Epidemiology, 178: 111622.

39. Wang R and de Gruttola V (2017) The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials. Statistics in Medicine, 36(18): 2831-2843.

40. Wang B, Wang X, Li F (2024) How to achieve model-robust inference in stepped wedge trials with model-based methods. Biometrics, 80(4): ujae123.

41. Woertman W, de Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL, Teerenstra S (2013) Stepped wedge designs could reduce the required sample size in cluster randomized trials. Journal of Clinical Epidemiology, 66: 752-758.

42. Wu D, Park HG, Grudzen CR, Goldfeld KS (2025) Bayesian Hierarchical Penalized Spline Models for Immediate and Time-Varying In-

tervention Effects in Stepped Wedge Cluster Randomized Trials. Stat Med. 2025 Feb 28;44(5):e10304.

43. Zelen M (1974) The randomization and stratification of patients to clinical trials. J Chronic Dis. Sep;27(7-8):365-75.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementdraft.pdf