Supporting information for "Assessing exposure-time treatment effect heterogeneity in stepped wedge cluster randomized trials" by

Lara Maleyeff[1,*], **Fan Li**[2,3], **Sebastien Haneuse**[1], and **Rui Wang**[1,4]

[1]Department of Biostatistics, Harvard T. H. Chan School of Public Health

[2]Department of Biostatistics, Yale School of Public Health

[3]Center for Methods in Implementation and Prevention Science, Yale School of Public Health

[4]Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School

[*]*email*: lmaleyeff@g.harvard.edu

Web Appendix A. Schematic illustration of stepped wedge cluster randomized trial

The stepped wedge cluster randomized trial (CRT) is a design variation in which treatment is rolled out in different clusters at randomly-assigned time points, until all clusters are exposed under the treatment condition; see Figure S1 for a graphical illustration of this design with 8 clusters and 5 time periods.

Figure S1: A schematic illustration of stepped wedge cluster randomized trial. “O” represents control intervention and “A” represents treatment intervention.

	Time				
Cluster	1	2	3	4	5
1	O	A	A	A	A
2	O	A	A	A	A
3	O	O	A	A	A
4	O	O	A	A	A
5	O	O	O	A	A
6	O	O	O	A	A
7	O	O	O	O	A
8	O	O	O	O	A

Web Appendix B. Causal interpretation of treatment effect parameters

In this section we provide a perspective on the model-based causal interpretation of the parameters in Models 1-5. Recall the general model formulation is

$$h\{\mathbb{E}(Y_{kti} \mid E_{kt}, \alpha_k)\} = \mu + \beta_t + \theta(E_{kt}) + \alpha_k, \tag{S1}$$

where μ is the global mean on the link function scale, β_t $(t = 1, \ldots, T)$ is the fixed effect for time ($\beta_1 = 0$ for identifiability), $\theta(E_{kt})$ is the treatment effect for each exposure time, and

$\alpha_k \sim \mathcal{N}(0, \sigma_\alpha^2)$ is a random cluster intercept. We assume that for each cluster k, there is a common underlying source population of size N_k, from which a random sample of size n_{kt} ($n_{kt} << N_k$) is obtained during each period t. This is typical assumption made in cross-sectional stepped wedge CRTs. We then write Y_{kti}^e as the potential outcome for participant i ($i = 1, \ldots, N_k$) in cluster k at period t that would have been observed if exposure time was set to e. For the purpose of this presentation, we focus on a standard stepped wedge design with one baseline period such that $E = T - 1$. Due to the study design, the potential outcomes Y_{kti}^e's are well defined only when $t \geq e + 1$, because the maximum exposure time never exceeds the period count minus one. To connect the potential outcome with the observed outcome, we then assume that there are not multiple versions of the exposure, meaning that all individuals receiving treatment for exposure time e cannot receive different forms of treatment which have different effects. We also assume that the potential outcomes are not affected by treatment assignments in other clusters and there is no interference between subjects in different clusters. These conditions lead to the consistency assumption such that $Y_{kti} = Y_{kti}^e$ if the exposure-time treatment level is set to be $E_{kt} = e$ in a specific period t with $e \leq t - 1$ (VanderWeele, 2009).

Under the above set up, for each participant i ($i = 1, \ldots, N_k$) in the underlying source population during calendar period t in cluster k, and for any $e \leq t - 1$, we have

$$\mathbb{E}(Y_{kti}^e \mid \alpha_k) = \mathbb{E}(Y_{kti}^e \mid E_{kt} = e, \alpha_k) = \mathbb{E}(Y_{kti} \mid E_{kt} = e, \alpha_k). \tag{S2}$$

The first equation holds because the randomized treatment sequence (and hence the exposure time E_{kt} which is fully determined by the sequence) is independent of any potential outcomes in each cluster; the second equation holds due to the consistency assumption.

From equation (S2), Model (S1) can be viewed as an example of a longitudinal structural mixed model (Sitlani *et al.*, 2012; Li *et al.*, 2021). Then, the parameters $\theta(e)$ in the general model (S1) can be written as a participant-level contrast:

$$\theta(e) = h\left\{\mathbb{E}(Y_{kti}^{e} \mid \alpha_k)\right\} - h\left\{\mathbb{E}(Y_{kti}^{0} \mid \alpha_k)\right\} \quad \text{for } i = 1, \ldots, N_k,\ 0 < e \leq t-1. \tag{S3}$$

In other words, $\theta(e)$ measures the difference between transformed mean potential outcomes under exposure time e and usual care, for a typical participant in the source population. This is a well-defined individual-level causal parameter once we reformulate the potential outcome as $\widetilde{Y}_{ki}^{e} = h\left(\mathbb{E}(Y_{ki}^{e} \mid \alpha_k)\right)$, and provides a plausible perspective on the causal interpretation of the exposure-time specific treatment effect parameter $\theta(e)$. Since Model S1 does not involve additional covariates to account for between-individual heterogeneity, the parameter $\theta(e)$ can be alternatively interpreted as a population-level causal effect parameter, by noticing

$$\theta(e) = \frac{\sum_{k=1}^{K}\left[h\left\{\mathbb{E}(Y_{kti}^{e} \mid \alpha_k)\right\} - h\left\{\mathbb{E}(Y_{kti}^{0} \mid \alpha_k)\right\}\right]}{\sum_{k=1}^{K} N_k} \quad \text{for } 0 < e \leq t-1. \tag{S4}$$

This is a well-defined causal effect parameter because we are averaging contrasts in transformed potential outcomes over a common population combined over all clusters.

Furthermore, we can average the exposure-time specific causal effects (comparing exposure level e and usual care) over levels of exposure times to obtain

$$\begin{aligned}&\text{(Exposure-time) average treatment effect}\\ =&\frac{1}{E}\sum_{e=1}^{E}\theta(e) = \frac{\sum_{e=1}^{E}\left[h\left\{\mathbb{E}(Y_{k,e+1,i}^{e} \mid \alpha_k)\right\} - h\left\{\mathbb{E}(Y_{k,e+1,i}^{0} \mid \alpha_k)\right\}\right]}{E}.\end{aligned} \tag{S5}$$

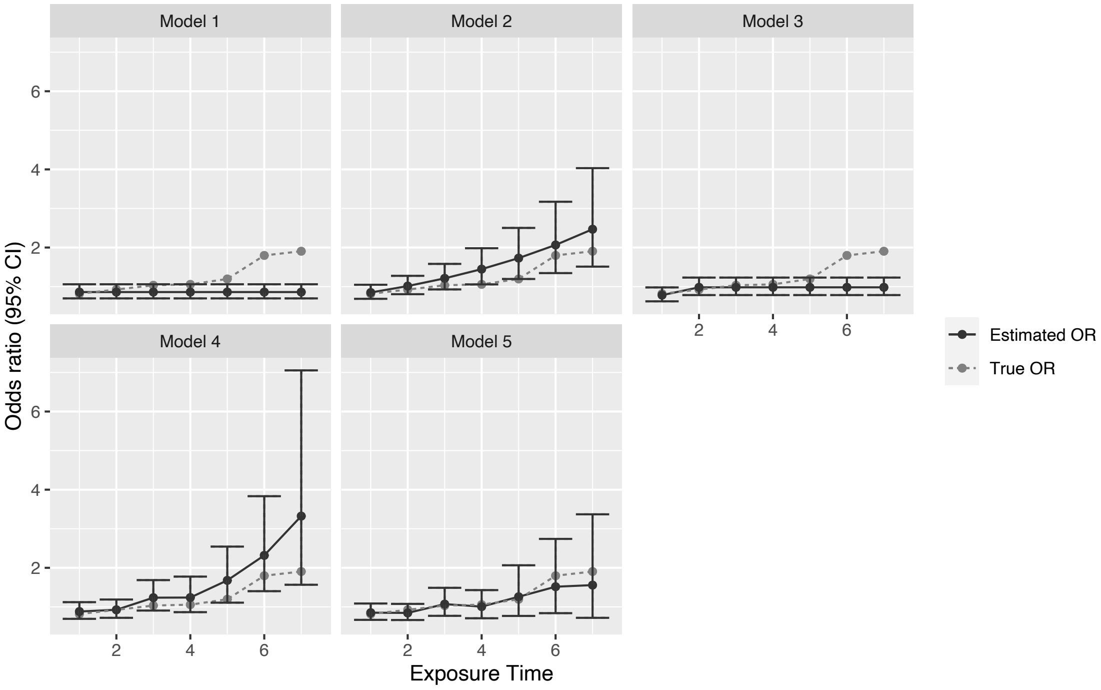
This average treatment effect is a well-defined participant-level causal effect parameter, because it is a well-defined summary of exposure-time specific causal effect parameter $\theta(e)$ for a typical participant in the underlying source population from cluster k. We can similarly interpret (S5) as a population-level causal parameter by averaging over the entire population over clusters. We acknowledge that the above discussion is only one possible framework under which the general model (S1) can be causally interpreted. Future work is needed is to develop estimands and their interpretations under the nonparametric causal model with fewer assumptions (and possibly including covariates).

Web Appendix C. Visual representation and comparison of exposure-time specific treatment effects estimates for the data example

To illustrate the use of our methods in settings where treatment effect varies by exposure time, we conducted an analysis of synthetic data where heterogeneity was induced in the Trajman *et al.* (2015) data described in Section 3 of the main paper. The procedure for inducing heterogeneity is also described in Section 3 of the main paper.

The exposure-time specific treatment effects from the simulated synthetic data are heterogeneous with increasing magnitude over exposure time (i.e. effect at exposure time 1 is less than effect at exposure time 2, and so on), with an average treatment effect of 1.19. Figure S2 shows the odds ratio estimates and their 95% confidence interval for each model's estimate of the exposure-time specific effect, compared with the true odds ratio which was used to generate the synthetic data. The 95% confidence interval was constructed using the within-cluster bootstrap standard error for each exposure-time effect estimate. In Model 1, we see a constant treatment effect over time with narrow confidence intervals. Model 1 is severely biased in our synthetic data, estimating the average treatment effect to be 0.86 (compared with 1.19). Model 2, which assumes a linearly exposure-time trend, slightly overestimates the odds ratios, with wider confidence intervals as the exposure-time period increases. Model 3 underestimates the later exposure-time effects, with the most severe underestimation being at month 7. Model 4 also over-estimates the later exposure-time effects, with increasing variance as exposure time increases. Model 5 had more stable confidence

Figure S2: Estimated odds ratio (OR), with 95% confidence interval (CI), as a function of exposure time for the synthetic data (with induced treatment effect heterogeneity) created from the XpertMTB/RIF Tuberculosis stepped wedge cluster randomized trial. Confidence intervals are constructed using the within-cluster bootstrap standard error. True OR is represented with a grey dotted line, estimated OR is represented with a solid black line.



interval widths over time, with effect estimates closest to the truth. The confidence intervals of Models 2, 4, and 5 all cover the true effect.

Web Appendix D. Additional continuous outcome simulation studies

We conducted two additional sets of simulations varying sample sizes and level of treatment effect heterogeneity across exposure time. For the first set, data were generated similarly to the simulation study in Section 4.2.1 except that the number of steps was increased from 8 to 30. For the 2nd set, we considered $n_{kt} \in \{30, 100\}$ and varying $\sigma_\delta^2 \in \{0.0, 0.2, 2.0\}$. Results are presented in Tables S1 and S2.

Table S1: Estimation of the average treatment effects with varying types of treatment effect heterogeneity across exposure time in 1000 simulated datasets with continuous outcomes. Each simulated dataset represents a stepped wedge CRT with $E = 29$ exposure-time periods, $T = 30$ steps, $K = 29$ clusters, and $n = 100$ individuals per cluster per time period. The Monte Carlo error associated with 95% coverage for 1000 simulation iterations is 0.7%. We use Δ to denote the average treatment effect, $\widehat{\Delta}$ to denote its estimator, $\widehat{\mathbb{E}}(\cdot)$ and $\widehat{SD}(\cdot)$ to denote sample average and sample standard deviation across simulated experiments, and $\widehat{SE}(\cdot)$ to denote the standard error estimator.

						Bootstrap			
					Model-based	Within cluster		Within cluster-period	
Model	$\widehat{\mathbb{E}}(\widehat{\Delta})$	$\widehat{\mathbb{E}}(\widehat{\sigma}_\alpha)$	$\widehat{\mathbb{E}}(\widehat{\sigma}_\delta)$	$\widehat{SD}(\widehat{\Delta})$	$\widehat{\mathbb{E}}(\widehat{SE}(\widehat{\Delta}))$	$\widehat{\mathbb{E}}(\widehat{SE}(\widehat{\Delta}))$	Coverage (%)	$\widehat{\mathbb{E}}(\widehat{SE}(\widehat{\Delta}))$	Coverage (%)
$g(e) = 2$, $\Delta = 2$, $\sigma_\delta = 0$, $\sigma_\alpha = 0.141$									
1	2.000	0.140	-	0.012	0.012	0.012	94.6	0.011	94.5
2	2.000	0.140	-	0.019	0.019	0.019	94.6	0.019	94.7
3	2.000	0.140	-	0.012	0.012	0.012	94.6	0.012	94.8
4	2.000	0.140	-	0.020	0.020	0.020	94.0	0.019	94.5
5	2.000	0.140	0.005	0.012	0.012	0.012	95.7	0.012	96.0
$g(e) = 2 + \delta_e, \delta_e \sim \mathcal{N}(0, 2^2)$, $\Delta = 2$, $\sigma_\delta = 2$, $\sigma_\alpha = 0.141$									
1	1.630	0.226	-	0.012	0.020	0.022	0.0	0.012	0.0
2	2.107	0.168	-	0.020	0.033	0.023	0.0	0.018	0.0
3	1.893	0.182	-	0.012	0.021	0.023	0.0	0.012	0.0
4	2.000	0.140	-	0.020	0.020	0.020	94.1	0.019	93.4
5	2.000	0.141	2.000	0.020	0.372	0.020	93.8	0.019	93.3
$g(e) = (e - 1.840)/1.080$, $\Delta = 2$, $\sigma_\delta = 2$, $\sigma_\alpha = 0.141$									
1	-1.405	1.040	-	0.011	0.015	0.015	0.0	0.012	0.0
2	1.999	0.139	-	0.019	0.019	0.019	94.1	0.019	94.6
3	-1.208	0.975	-	0.012	0.015	0.016	0.0	0.012	0.0
4	1.999	0.139	-	0.020	0.020	0.020	94.2	0.019	94.4
5	1.997	0.139	1.999	0.019	0.372	0.020	94.6	0.019	94.6
$g(e) = -2.536I(e = 1) + 2.756I(e > 1)$, $\Delta = 2$, $\sigma_\delta = 2$, $\sigma_\alpha = 0.141$									
1	0.294	0.605	-	0.012	0.025	0.037	0.0	0.012	0.0
2	2.734	0.260	-	0.021	0.039	0.028	0.0	0.019	0.0
3	2.000	0.142	-	0.012	0.012	0.012	94.8	0.012	94.6
4	2.000	0.142	-	0.020	0.020	0.020	94.2	0.019	93.6
5	1.999	0.143	2.001	0.021	0.372	0.020	93.8	0.019	93.5

Table S2: Estimation of the average treatment effects with varying degrees of exposure-time treatment effect heterogeneity in 1000 simulated datasets with continuous outcomes. Each simulated dataset represents a stepped wedge CRT with $E = 29$ exposure-time periods, $T = 30$ steps, $K = 29$ clusters, and n individuals per cluster per time period. The Monte Carlo error associated with 95% coverage for 1000 simulation iterations is 0.7%. We use Δ to denote the average treatment effect, $\widehat{\Delta}$ to denote its estimator, $\widehat{\mathbb{E}}(\cdot)$ and $\widehat{SD}(\cdot)$ to denote sample average and sample standard deviation across simulated experiments, and $\widehat{SE}(\cdot)$ to denote the standard error estimator.

								Bootstrap			
							Model-based	Within cluster		Within cluster-period	
n	σ_δ	Model	$\widehat{\mathbb{E}}(\widehat{\Delta})$	$\widehat{\mathbb{E}}(\widehat{\sigma}_\alpha)$	$\widehat{\mathbb{E}}(\widehat{\sigma}_\delta)$	$\widehat{SD}(\widehat{\Delta})$	$\widehat{\mathbb{E}}(\widehat{SE}(\widehat{\Delta}))$	$\widehat{\mathbb{E}}(\widehat{SE}(\widehat{\Delta}))$	Coverage (%)	$\widehat{\mathbb{E}}(\widehat{SE}(\widehat{\Delta}))$	Coverage (%)
30	0.0	1	2.000	0.140	-	0.021	0.021	0.021	92.4	0.020	92.2
30	0.0	2	2.001	0.140	-	0.036	0.033	0.032	91.7	0.031	90.9
30	0.0	3	2.000	0.140	-	0.022	0.022	0.021	92.7	0.021	91.2
30	0.0	4	2.000	0.140	-	0.037	0.035	0.033	93.2	0.032	91.7
30	0.0	5	2.000	0.140	0.009	0.022	0.022	0.022	94.0	0.021	93.7
30	0.2	1	1.965	0.141	-	0.022	0.021	0.021	60.0	0.020	58.8
30	0.2	2	2.011	0.140	-	0.034	0.034	0.032	92.5	0.031	91.9
30	0.2	3	1.990	0.140	-	0.023	0.022	0.021	90.7	0.021	89.8
30	0.2	4	2.000	0.140	-	0.035	0.035	0.033	92.7	0.033	92.8
30	0.2	5	2.000	0.140	0.199	0.031	0.049	0.029	93.3	0.029	93.2
30	2.0	1	1.653	0.217	-	0.022	0.036	0.040	0.0	0.021	0.0
30	2.0	2	2.112	0.161	-	0.038	0.054	0.037	14.4	0.028	5.9
30	2.0	3	1.906	0.175	-	0.023	0.036	0.040	24.6	0.020	0.8
30	2.0	4	2.002	0.139	-	0.036	0.035	0.033	93.0	0.032	91.6
30	2.0	5	2.002	0.139	2.000	0.036	0.373	0.033	92.8	0.032	91.6
100	0.0	1	2.000	0.140	-	0.012	0.012	0.012	94.6	0.011	94.5
100	0.0	2	2.000	0.140	-	0.019	0.019	0.019	94.6	0.019	94.7
100	0.0	3	2.000	0.140	-	0.012	0.012	0.012	94.6	0.012	94.8
100	0.0	4	2.000	0.140	-	0.020	0.020	0.020	94.0	0.019	94.5
100	0.0	5	2.000	0.140	0.005	0.012	0.012	0.012	95.7	0.012	96.0
100	0.2	1	1.963	0.141	-	0.012	0.012	0.012	11.8	0.011	10.9
100	0.2	2	2.010	0.140	-	0.019	0.019	0.019	91.1	0.019	90.5
100	0.2	3	1.989	0.140	-	0.012	0.012	0.012	83.9	0.012	83.1
100	0.2	4	2.000	0.140	-	0.020	0.020	0.020	94.0	0.019	94.5
100	0.2	5	1.999	0.140	0.199	0.019	0.042	0.019	94.4	0.018	94.2
100	2.0	1	1.630	0.226	-	0.012	0.020	0.022	0.0	0.012	0.0
100	2.0	2	2.107	0.168	-	0.020	0.033	0.023	0.0	0.018	0.0
100	2.0	3	1.893	0.182	-	0.012	0.021	0.023	0.0	0.012	0.0
100	2.0	4	2.000	0.140	-	0.020	0.020	0.020	94.1	0.019	93.4
100	2.0	5	2.000	0.141	2.000	0.020	0.372	0.020	93.8	0.019	93.3

Web Appendix E. Additional simulation study to assess estimation of exposure-time specific effects

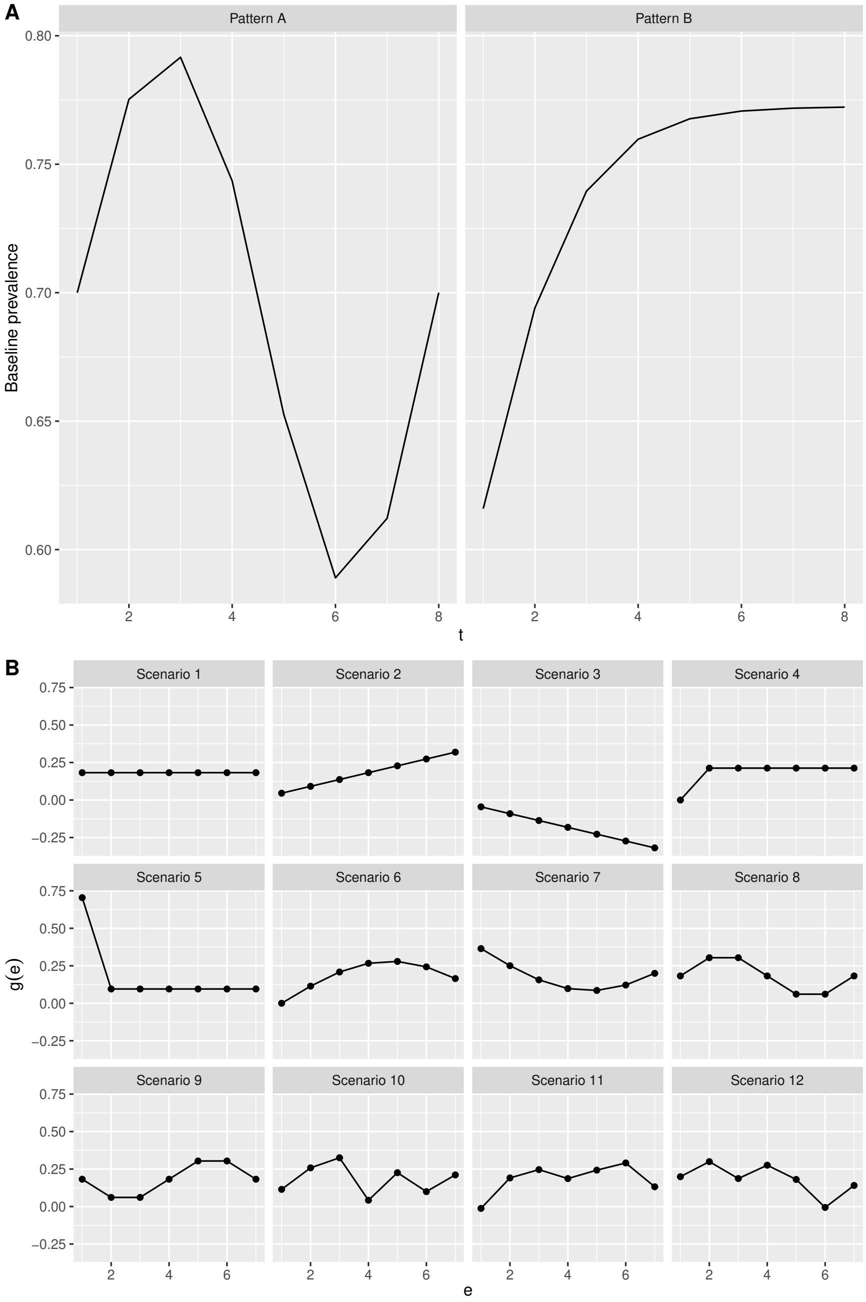
To assess estimation of the exposure-time specific effects in each of the five models, we conducted simulations A-B, in which we considered a study conducted over 8 months. The corresponding 7 clusters were labelled as such: cluster 1 receives their first unit of exposure time at month 2, cluster 2 receives their first unit of exposure time at month 3, and so on. Each cluster k had an equal number of subjects ($n_{kt} = 30$) sampled at each time point t. Data were generated from the logistic generalized linear mixed model: $\text{logit}\{\mathbb{P}(Y_{kti} = 1 \mid \alpha_k, E_{kt})\} = \text{logit}(0.7) + f(t) + g(E_{kt}) + \alpha_k$, where $f(t)$ corresponds to a background calendar time trend, $g(E_{kt})$ models the treatment effect as a function of exposure time, and α_k is drawn from $\mathcal{N}(0, 0.05)$. Figure S3 shows the functional forms of $f(t)$ and $g(E_{kt})$, respectively, that we considered. Table S3 describes these functional forms in equations.

Figure S4 shows the 95% CI coverage and width for each exposure-time specific treatment effect estimate in scenario 10 (A and B), where data were generated from Model 5 with $\delta_{E_{kt}} \sim \mathcal{N}(0, \sigma_\delta^2)$. Here, Models 1, 2, and 3 are all misspecified. Model 1 assumes a constant effect over exposure time, implying the same estimate for each exposure time effect. As such, the coverage is constantly below nominal (average coverage is 89.7%). Models 4 and 5 have similar coverage for earlier exposure months. In exposure months 1-4, coverage for Model 4 ranges from 92.7% to 94.7% and coverage for Model 5 ranges from 92.5% to 94.7%. For months 5 and 6, Model 5 has higher coverage than model 4, with coverage ranging from 95.8% to 97.1% in Model 5 and 92.3% to 93.9% in Model 4. We infer that this is because coverage based on Model 5 is generally higher in exposure-time periods which had an exposure-time specific effect closer to the average treatment effect as Model 5 shrinks the estimates of the exposure-time treatment effects towards the overall mean. At month 7, both Models 4 and 5 have similar coverage ranging from 96.2% to 97.5%, with Model 4 having a much wider confidence interval (2.2 vs. 0.8). The CI width for estimates based on

Table S3: Simulation scenarios for different exposure-time treatment effect patterns (both A and B). Data were generated from the logistic generalized linear mixed model, $\text{logit}\{\mathbb{P}(Y_{kti} = 1|\alpha_k, E_{kt})\} = \text{logit}(0.7) + f(t) + g(E_{kt}) + \alpha_k$, where $f(t)$ corresponds to a background calendar time trend, $g(e)$ models the treatment effect as a function of exposure time e, and α_k is drawn from $\mathcal{N}(0, 0.05)$. Simulation scenarios A and B differ only in the background secular trend. Scenario A sets $f(t) = 0.5 \sin\{2\pi(t-1)/7\}$ and scenario B sets $f(t) = 1.5e^{t-1}/(1+e^{t-1}) - 0.75$. AR-1: first order autoregressive

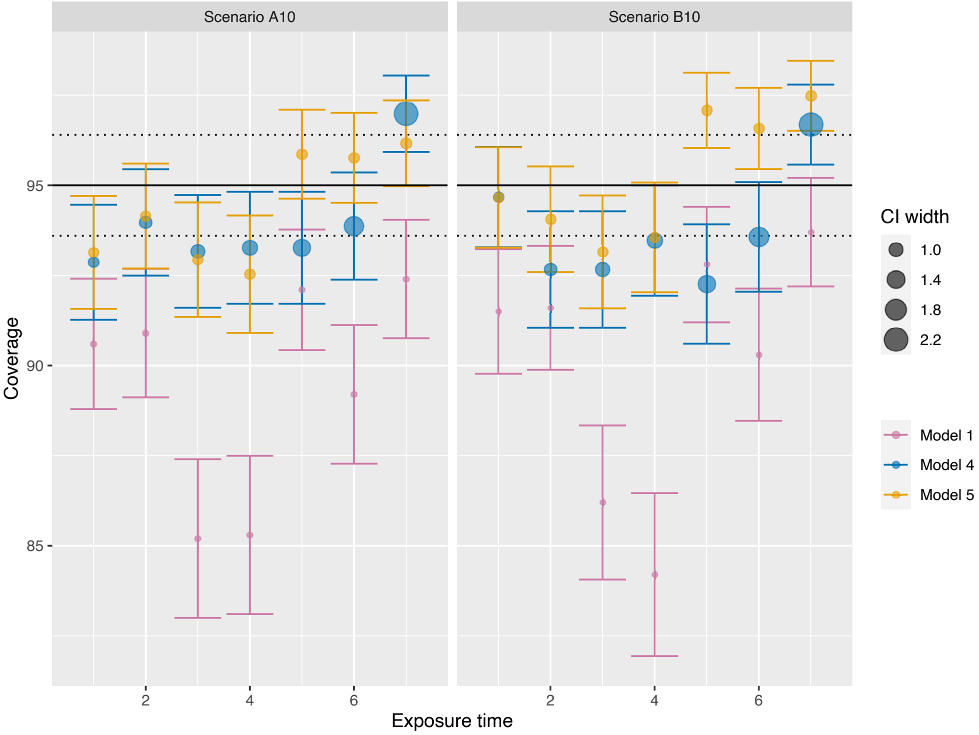
Scenario	**Treatment effect as a function of E_{kt}**
1	$g(e) = \log(1.2)$
2	$g(e) = \log(1.2)/4e$
3	$g(e) = -\log(1.2)/4e + 2\log(1.2)$
4	$g(e) = \log(1)I(e=1) + \frac{7}{6}\log(1.2)I(e>1)$
5	$g(e) = \{6\log(1.2) - 5\log(1.1)\}I(e=1) + \log(1.1)I(e>1)$
6	$g(e) = 0.28\sin\left\{\frac{0.8\pi(e-1)}{6}\right\} + 2\log(1.2)$
7	$g(e) = -0.28\sin\left\{\frac{0.8\pi(e-1)}{6}\right\} + 2\log(1.2)$
8	$g(e) = 0.14\sin\left\{\frac{2\pi(e-1)}{6}\right\} + \log(1.2)$
9	$g(e) = -0.14\sin\left\{\frac{2\pi(e-1)}{6}\right\} + \log(1.2)$
10	Model 5 with $\phi = \log(1.2)$ and $\delta_{E_{kt}} \sim \mathcal{N}(0, 0.1^2)$
11	Model 5 with $\phi = \log(1.2)$ and $\delta_{E_{kt}} = 0.022Z + 0.025Z^2 + 0.01Z^3,\ Z \sim \mathcal{N}(0,1)$
12	Model 5 with $\phi = \log(1.2)$ and $\delta \sim \mathcal{N}(0, 0.1^2\mathbf{M})$, $\mathbf{M}$ AR-1 with correlation 0.2

Figure S3: Graphical illustration of (A) baseline prevalence in simulation scenarios A and B, defined as $expit(\text{logit}(0.7) + f(t))$ as a function of time t; and (B) exposure-time treatment effect function, $g(e)$, on the link function scale for each simulation scenario 1-12. e is the exposure time unit ranging from 1 to 7.



Model 4 increase with increasing exposure time, ranging from 0.8 to 2.2 in both scenarios A10 and B10. The CI widths for estimates based on Model 5, however, are stable over time (all 0.8) due to increased efficiency by modeling the set of treatment parameters via random effects.

Figure S4: Summary of 95% confidence interval (CI) coverage of exposure-time specific treatment effect, using within-cluster bootstrapped standard error, from scenario 10 (A and B) in simulation study. Circle size is proportional to CI width. Model 1 is pink, 4 is blue, and 5 is orange. Black line indicates 95% coverage, lower and upper dotted lines indicate 93.6% and 96.4% coverage, respectively.



Web Appendix F. Additional information on trial planning

F.1 Derivation of variance expression for the average treatment effect estimator in Model 4

In this section we briefly describe how to compute the weighted least squares variance for trial planning with an average treatment effect target based on Model 4. Let

$$Y_{kti} = \mu + \beta_t + \theta_{E_{kt}} + \alpha_k + \epsilon_{kti},$$

where μ is the intercept, β_t $(t = 1, \ldots, T)$ is the fixed effect for time ($\beta_1 = 0$ for identifiability), $\theta_{E_{kt}}$ is the treatment effect for each exposure time ($\theta_0 = 0$ for identifiability), $\alpha_k \sim \mathcal{N}(0, \sigma_\alpha^2)$ is a random cluster intercept, and $\epsilon_{kti} \sim \mathcal{N}(0, \sigma_\epsilon^2)$. To facilitate the derivation, we can re-parameterize the model by dropping the overall mean, such that the model becomes

$$Y_{kti} = \beta_t + \theta_{E_{kt}} + \alpha_k + \epsilon_{kti}, \tag{S6}$$

where there is no restriction on β_t but the restriction that $\theta_0 = 0$ remains. Assuming Model S6, the variance of the estimated average treatment effect, $E^{-1} \sum_{e=1}^{E} \widehat{\theta}_e$ is the lower right entry of $\mathbf{C}'\mathbf{Var}(\widehat{\boldsymbol{\theta}})\mathbf{C}$, where $\mathbf{Var}(\widehat{\boldsymbol{\theta}})$ is the lower $E \times E$ block of $(\mathbf{F}'\mathbf{V}^{-1}\mathbf{F})^{-1}$, $\mathbf{C} = E^{-1}\mathbf{1}_E$ and

$$\mathbf{F} = \begin{pmatrix} \mathbf{I}_T & \mathbf{Z}_1 \\ \mathbf{I}_T & \mathbf{Z}_2 \\ \vdots & \vdots \\ \mathbf{I}_T & \mathbf{Z}_K \end{pmatrix} \otimes \mathbf{1}_n$$

where $\mathbf{Z}_k$ is a $T \times E$ matrix of the form

$$\mathbf{Z}_k = \begin{pmatrix} \mathbf{0}_{(s-1)\times E} \\ \mathbf{I}_{(T-s+1)\times E} \end{pmatrix},$$

where s is the period in which treatment k starts to receive treatment, $\mathbf{1}_n$ is a n-vector of ones, $\mathbf{I}_n$ is an $n \times n$ identity matrix, and the covariance matrix for all observations can be partitioned as

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0}' & \mathbf{V}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \dots & \mathbf{V}_K \end{pmatrix},$$

where $\mathbf{V}_k = \sigma_\epsilon^2 \mathbf{I}_n + \sigma_\alpha^2 \mathbf{J}_n$, $\mathbf{J}_n = \mathbf{1}_n \mathbf{1}_n'$. Using the property of Kronecker products that $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$, we have

$$\begin{aligned} \mathbf{F}' &= \begin{pmatrix} \mathbf{I}_T & \mathbf{Z}_1 \\ \mathbf{I}_T & \mathbf{Z}_2 \\ \vdots & \vdots \\ \mathbf{I}_T & \mathbf{Z}_K \end{pmatrix}' \otimes \mathbf{1}_n' \\ &= \begin{pmatrix} \mathbf{I}_T & \mathbf{I}_T & \dots & \mathbf{I}_T \\ \mathbf{Z}_1' & \mathbf{Z}_2' & \dots & \mathbf{Z}_K' \end{pmatrix} \otimes \mathbf{1}_n'. \end{aligned}$$

Next, we have

$$\mathbf{F}'\mathbf{V}^{-1} = \left(\begin{pmatrix} \mathbf{I}_T & \mathbf{I}_T & \cdots & \mathbf{I}_T \\ \mathbf{Z}_1' & \mathbf{Z}_2' & \cdots & \mathbf{Z}_K' \end{pmatrix} \otimes \mathbf{1}_n' \right) \begin{pmatrix} \mathbf{V}_1^{-1} & 0 & \dots & 0 \\ 0 & \mathbf{V}_2^{-1} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \mathbf{V}_K^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} \left(\mathbf{I}_T \otimes \mathbf{1}_n'\right)\mathbf{V}_1^{-1} & \left(\mathbf{I}_T \otimes \mathbf{1}_n'\right)\mathbf{V}_2^{-1} & \cdots & \left(\mathbf{I}_T \otimes \mathbf{1}_n'\right)\mathbf{V}_K^{-1} \\ \left(\mathbf{Z}_1' \otimes \mathbf{1}_n'\right)\mathbf{V}_1^{-1} & \left(\mathbf{Z}_2' \otimes \mathbf{1}_n'\right)\mathbf{V}_2^{-1} & \cdots & \left(\mathbf{Z}_K' \otimes \mathbf{1}_n'\right)\mathbf{V}_K^{-1} \end{pmatrix}.$$

Then, $\mathbf{F}'\mathbf{V}^{-1}\mathbf{F}$ equals

$$= \begin{pmatrix} \left(\mathbf{I}_T \otimes \mathbf{1}_n'\right)\mathbf{V}_1^{-1} & \left(\mathbf{I}_T \otimes \mathbf{1}_n'\right)\mathbf{V}_2^{-1} & \cdots & \left(\mathbf{I}_T \otimes \mathbf{1}_n'\right)\mathbf{V}_K^{-1} \\ \left(\mathbf{Z}_1' \otimes \mathbf{1}_n'\right)\mathbf{V}_1^{-1} & \left(\mathbf{Z}_2' \otimes \mathbf{1}_n'\right)\mathbf{V}_2^{-1} & \cdots & \left(\mathbf{Z}_K' \otimes \mathbf{1}_n'\right)\mathbf{V}_K^{-1} \end{pmatrix} \left(\begin{pmatrix} \mathbf{I}_T & \mathbf{Z}_1 \\ \mathbf{I}_T & \mathbf{Z}_2 \\ \vdots & \vdots \\ \mathbf{I}_T & \mathbf{Z}_K \end{pmatrix} \otimes \mathbf{1}_n \right)$$

$$= \begin{pmatrix} \sum_{k=1}^K \left(\mathbf{I}_T \otimes \mathbf{1}_n'\right)\mathbf{V}_k^{-1}\left(\mathbf{I}_T \otimes \mathbf{1}_n\right) & \sum_{k=1}^K \left(\mathbf{I}_T \otimes \mathbf{1}_n'\right)\mathbf{V}_k^{-1}\left(\mathbf{Z}_k \otimes \mathbf{1}_n\right) \\ \sum_{k=1}^K \left(\mathbf{Z}_k' \otimes \mathbf{1}_n'\right)\mathbf{V}_k^{-1}\left(\mathbf{I}_T \otimes \mathbf{1}_n\right) & \sum_{k=1}^K \left(\mathbf{Z}_k' \otimes \mathbf{1}_n'\right)\mathbf{V}_k^{-1}\left(\mathbf{Z}_k \otimes \mathbf{1}_n\right) \end{pmatrix}.$$

Since we are assuming that the covariance matrix is the same for each cluster (due to the conventional assumption of equal cluster size), then $\mathbf{V}_1 = \cdots = \mathbf{V}_K = \mathbf{V}_*$, and so

$$\mathbf{F}'\mathbf{V}^{-1}\mathbf{F} = \begin{pmatrix} K\left(\mathbf{I}_T \otimes \mathbf{1}_n'\right)\mathbf{V}_*^{-1}\left(\mathbf{I}_T \otimes \mathbf{1}_n\right) & \left(\mathbf{I}_T \otimes \mathbf{1}_n'\right)\mathbf{V}_*^{-1}\left(\sum_{k=1}^K \mathbf{Z}_k \otimes \mathbf{1}_n\right) \\ \left(\sum_{k=1}^K \mathbf{Z}_k' \otimes \mathbf{1}_n'\right)\mathbf{V}_*^{-1}\left(\mathbf{I}_T \otimes \mathbf{1}_n\right) & \sum_{k=1}^K \left(\mathbf{Z}_k' \otimes \mathbf{1}_n'\right)\mathbf{V}_*^{-1}\left(\mathbf{Z}_k \otimes \mathbf{1}_n\right) \end{pmatrix}$$

Let

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} = \begin{pmatrix} K\left(\mathbf{I}_T \otimes \mathbf{1}_n'\right)\mathbf{V}_*^{-1}\left(\mathbf{I}_T \otimes \mathbf{1}_n\right) & \left(\mathbf{I}_T \otimes \mathbf{1}_n'\right)\mathbf{V}_*^{-1}\left(\sum_{k=1}^K \mathbf{Z}_k \otimes \mathbf{1}_n\right) \\ \left(\sum_{k=1}^K \mathbf{Z}_k' \otimes \mathbf{1}_n'\right)\mathbf{V}_*^{-1}\left(\mathbf{I}_T \otimes \mathbf{1}_n\right) & \sum_{k=1}^K \left(\mathbf{Z}_k' \otimes \mathbf{1}_n'\right)\mathbf{V}_*^{-1}\left(\mathbf{Z}_k \otimes \mathbf{1}_n\right) \end{pmatrix}$$

Then, using the definition of the inverse of a partitioned matrix, the lower $E \times E$ matrix of $(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-1}$ is given by $\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1}$ and so

$$\mathbf{Var}(\widehat{\theta}) = \left(\sum_{k=1}^{K} \mathbf{D}_k'\mathbf{V}_*^{-1}\mathbf{D}_k - \frac{1}{K}\mathbf{B}'\mathbf{V}_*^{-1}\mathbf{G}\left(\mathbf{G}'\mathbf{V}_*^{-1}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{V}_*^{-1}\mathbf{B}\right)^{-1}$$

where $\mathbf{B} = \left(\sum_{k=1}^{K} \mathbf{Z}_k \otimes \mathbf{1}_n\right)$, $\mathbf{G} = (\mathbf{I}_T \otimes \mathbf{1}_n)$ and $\mathbf{D}_k = \mathbf{Z}_k \otimes \mathbf{1}_n$, and

$$\text{Var}\left(\frac{1}{K}\sum_{e=1}^{E}\widehat{\theta}_e\right) = \mathbf{C}'\left(\sum_{k=1}^{K} \mathbf{D}_k'\mathbf{V}_*^{-1}\mathbf{D}_k - \frac{1}{K}\mathbf{B}'\mathbf{V}_*^{-1}\mathbf{G}\left(\mathbf{G}'\mathbf{V}_*^{-1}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{V}_*^{-1}\mathbf{B}\right)^{-1}\mathbf{C}$$

We then further simplify the expression. Importantly, we define the (generalized version of) design constant as $\mathbf{U}_1 = \sum_{k=1}^{K} \mathbf{Z}_k'\mathbf{1}_T$ as the E vector, $\mathbf{W}_1 = \left(\sum_{k=1}^{K} \mathbf{Z}_k'\right)\left(\sum_{k=1}^{K} \mathbf{Z}_k\right)$ is an $E \times E$ matrix, $\mathbf{W}_2 = \sum_{k=1}^{K} \mathbf{Z}_k'\mathbf{J}_T\mathbf{Z}_k$ is also an $E \times E$ matrix, and finally $\mathbf{U}_2 = \sum_{k=1}^{K} \mathbf{Z}_k'\mathbf{Z}_k$ is an $E \times E$ matrix. First we notice that the common variance matrix of the outcome vector in each cluster is simple exchangeable, given by $\mathbf{V}^* = \sigma_y^2\left[(1-\rho)\mathbf{I}_{Tn} + \rho\mathbf{J}_{Tn}\right] = \sigma_y^2\mathbf{R}$,

where $\sigma_y^2 = \sigma_\alpha^2 + \sigma_e^2$ is the total outcome variance, and $\rho = \sigma_\alpha^2/\sigma_y^2$ is the common ICC. The correlation matrix $\mathbf{R}$ has two eigenvalues, $\lambda_1 = 1-\rho, \lambda_2 = 1 + (Tn-1)\rho$. Thus, the inverse is given by

$$\mathbf{R}^{-1} = \frac{1}{1-\rho}\mathbf{I}_{Tn} - \frac{\rho}{(1-\rho)[1+(Tn-1)\rho]}\mathbf{J}_{Tn} = \frac{1}{\lambda_1}\mathbf{I}_{Tn} - \frac{\lambda_2 - \lambda_1}{Tn\lambda_1\lambda_2}\mathbf{J}_{Tn} = a\mathbf{I}_{Tn} + b\mathbf{J}_{Tn}, \quad \text{(S7)}$$

with $a = 1/\lambda_1$ and $b = -(\lambda_2 - \lambda_1)/(Tn\lambda_1\lambda_2)$. Then we observe

$$\begin{aligned}
\mathbf{G}'\mathbf{R}^{-1}\mathbf{G} &= \left(\mathbf{I}_T \otimes \mathbf{1}_n'\right)\left(a\mathbf{I}_{Tn} + b\mathbf{J}_{Tn}\right)\left(\mathbf{I}_T \otimes \mathbf{1}_n\right) = \left(\mathbf{I}_T \otimes \mathbf{1}_n'\right)\left(a\mathbf{I}_T \otimes \mathbf{I}_n + b\mathbf{J}_T \otimes \mathbf{J}_n\right)\left(\mathbf{I}_T \otimes \mathbf{1}_n\right) \\
&= an\mathbf{I}_T + bn^2\mathbf{J}_T, \\
\left(\mathbf{G}'\mathbf{R}^{-1}\mathbf{G}\right)^{-1} &= \frac{1}{an}\mathbf{I}_T - \frac{bn^2}{an\left(an + Tbn^2\right)}\mathbf{J}_T = \frac{\lambda_1}{n}\mathbf{I}_T + \frac{\lambda_2 - \lambda_1}{Tn}\mathbf{J}_T = \frac{1}{n}\left(\lambda_1\mathbf{I}_T + \frac{\lambda_2 - \lambda_1}{T}\mathbf{J}_T\right) \\
\mathbf{B}'\mathbf{R}^{-1}\mathbf{G} &= \left(\sum_{k=1}^{K}\mathbf{Z}_k' \otimes \mathbf{1}_n\right)\left(a\mathbf{I}_T \otimes \mathbf{I}_n + b\mathbf{J}_T \otimes \mathbf{J}_n\right)\left(\mathbf{I}_T \otimes \mathbf{1}_n\right) \\
&= an\sum_{k=1}^{K}\mathbf{Z}_k' + bn^2\sum_{k=1}^{K}\mathbf{Z}_k'\mathbf{J}_T = \frac{n}{\lambda_1}\sum_{k=1}^{K}\mathbf{Z}_k' - \frac{\left(\lambda_2 - \lambda_1\right)n}{\lambda_1\lambda_2 T}\mathbf{U}_1\mathbf{1}_T' \\
\mathbf{G}'\mathbf{R}^{-1}\mathbf{B} &= an\sum_{k=1}^{K}\mathbf{Z}_k + bn^2\sum_{k=1}^{K}\mathbf{J}_T\mathbf{Z}_k = \frac{n}{\lambda_1}\sum_{k=1}^{K}\mathbf{Z}_k - \frac{\left(\lambda_2 - \lambda_1\right)n}{\lambda_1\lambda_2 T}\left(\mathbf{U}_1\mathbf{1}_T'\right)'
\end{aligned}$$

We then simplify $\left(\mathbf{B}'\mathbf{R}^{-1}\mathbf{G}\right)\left(\mathbf{G}'\mathbf{R}^{-1}\mathbf{G}\right)^{-1}\left(\mathbf{G}'\mathbf{R}^{-1}\mathbf{B}\right)$ as

$$\begin{aligned}
=&n\left(\frac{1}{\lambda_1}\sum_{k=1}^{K}\mathbf{Z}_k' - \frac{\lambda_2 - \lambda_1}{\lambda_1\lambda_2 T}\mathbf{U}_1'\mathbf{1}_T'\right)\left(\lambda_1\mathbf{I}_T + \frac{\lambda_2 - \lambda_1}{T}\mathbf{J}_T\right)\left(\frac{1}{\lambda_1}\sum_{k=1}^{K}\mathbf{Z}_k - \frac{\lambda_2 - \lambda_1}{\lambda_1\lambda_2 T}\left(\mathbf{U}_1\mathbf{1}_T'\right)'\right)\alpha^{(n)} \\
=&\frac{1}{\lambda_1}\mathbf{W}_1 - \frac{\lambda_2 - \lambda_1}{\lambda_1\lambda_2 T}\mathbf{U}_1^{\otimes 2} + \frac{\lambda_2 - \lambda_1}{\lambda_1^2 T}\mathbf{U}_1^{\otimes 2} - \frac{\lambda_2 - \lambda_1}{\lambda_1\lambda_2 T}\mathbf{U}_1^{\otimes 2} - \frac{\left(\lambda_2 - \lambda_1\right)^2}{\lambda_1^2\lambda_2 T}\mathbf{U}_1^{\otimes 2} \\
&+ \frac{\left(\lambda_2 - \lambda_1\right)^2}{\lambda_1\lambda_2^2 T}\mathbf{U}_1^{\otimes 2} - \frac{\left(\lambda_2 - \lambda_1\right)^2}{\lambda_1^2\lambda_2 T}\mathbf{U}_1^{\otimes 2} + \frac{\left(\lambda_2 - \lambda_1\right)^3}{\lambda_1^2\lambda_2^2 T}\mathbf{U}_1^{\otimes 2} \\
=&\frac{1}{\lambda_1}\mathbf{W}_1 - \frac{2\left(\lambda_2 - \lambda_1\right)}{\lambda_1\lambda_2 T}\mathbf{U}_1^{\otimes 2} + \frac{\lambda_2 - \lambda_1}{\lambda_1^2 T}\mathbf{U}_1^{\otimes 2} - \frac{2\left(\lambda_2 - \lambda_1\right)^2}{\lambda_1^2\lambda_2 T}\mathbf{U}_1^{\otimes 2} \\
&+ \frac{\left(\lambda_2 - \lambda_1\right)^2}{\lambda_1\lambda_2^2 T}\mathbf{U}_1^{\otimes 2} + \frac{\left(\lambda_2 - \lambda_1\right)^3}{\lambda_1^2\lambda_2^2 T}\mathbf{U}_1^{\otimes 2} \\
=&\frac{1}{\lambda_1}\mathbf{W}_1 + \frac{\left(\lambda_2 - \lambda_1\right)}{\lambda_1 T}\left[-\frac{2}{\lambda_2} + \frac{1}{\lambda_1} - \frac{2\left(\lambda_2 - \lambda_1\right)}{\lambda_1\lambda_2} + \frac{\left(\lambda_2 - \lambda_1\right)}{\lambda_2^2} + \frac{\left(\lambda_2 - \lambda_1\right)^2}{\lambda_1\lambda_2^2}\right]\mathbf{U}_1^{\otimes 2} \\
=&\frac{1}{\lambda_1}\mathbf{W}_1 - \frac{\left(\lambda_2 - \lambda_1\right)}{\lambda_1\lambda_2 T}\mathbf{U}_1^{\otimes 2}
\end{aligned}$$

and further notice that

$$\mathbf{D}_k'\mathbf{R}^{-1}\mathbf{D}_k = \left(\mathbf{Z}_k' \otimes \mathbf{1}_n'\right)\left(a\mathbf{I}_T \otimes \mathbf{I}_n + b\mathbf{J}_T \otimes \mathbf{J}_n\right)\left(\mathbf{Z}_k \otimes \mathbf{1}_n\right) = an\mathbf{Z}_k'\mathbf{Z}_k + bn^2\mathbf{Z}_k'\mathbf{J}_T$$

Then the inverse variance is proportional to

$$
\begin{aligned}
&\sum_{k=1}^{K} an\mathbf{Z}_k'\mathbf{Z}_k + \sum_{k=1}^{K} bn^2\mathbf{Z}_k'\mathbf{J}_T\mathbf{Z}_k - \frac{1}{K}\left(\mathbf{B}'\mathbf{R}^{-1}\mathbf{G}\right)\left(\mathbf{G}'\mathbf{R}^{-1}\mathbf{G}\right)^{-1}\left(\mathbf{G}'\mathbf{R}^{-1}\mathbf{B}\right)\\
=&\frac{n}{\lambda_1}\mathbf{U}_2 - \frac{\lambda_2-\lambda_1}{T\lambda_1\lambda_2}n\mathbf{W}_2 - \frac{n}{K}\frac{1}{\lambda_1}\mathbf{W}_1 + \frac{n}{K}\frac{(\lambda_2-\lambda_1)}{\lambda_1\lambda_2 T}\mathbf{U}_1^{\otimes 2}\\
=&\frac{n}{K\lambda_1}\left(K\mathbf{U}_2-\mathbf{W}_1\right) - \frac{n\left(\lambda_2-\lambda_1\right)}{KT\lambda_1\lambda_2}\left(K\mathbf{W}_2-\mathbf{U}_1^{\otimes 2}\right)
\end{aligned}
$$

The variance matrix for the vector of treatment effect estimators become

$$
\begin{aligned}
\mathbf{Var}(\widehat{\theta}) &= \sigma_y^2\left[\frac{n}{K\lambda_1}\left(K\mathbf{U}_2-\mathbf{W}_1\right) - \frac{n\left(\lambda_2-\lambda_1\right)}{KT\lambda_1\lambda_2}\left(K\mathbf{W}_2-\mathbf{U}_1^{\otimes 2}\right)\right]^{-1}\\
&= \sigma_y^2\frac{KT\lambda_1\lambda_2}{n}\left[\lambda_2\left(KT\mathbf{U}_2-T\mathbf{W}_1\right)-\left(\lambda_2-\lambda_1\right)\left(K\mathbf{W}_2-\mathbf{U}_1^{\otimes 2}\right)\right]^{-1}\\
&= \sigma_y^2\frac{KT\lambda_1\lambda_2}{n}\left[\lambda_2\left(\mathbf{U}_1^{\otimes 2}+KT\mathbf{U}_2-T\mathbf{W}_1-K\mathbf{W}_2\right)+\lambda_1\left(K\mathbf{W}_2-\mathbf{U}_1^{\otimes 2}\right)\right]^{-1},
\end{aligned}
$$

where $\mathbf{U}_1^{\otimes 2}=\mathbf{U}_1\mathbf{U}_1'$. Notice that this form resembles the variance expression in Hussey and Hughes (2007) and reviewed in Li *et al.* (2021; equation (7)), but now with the generalized version of design constants. Finally, the variance of the average treatment effect estimator from Model 4 is

$$
\begin{aligned}
&\text{Var}\left(\frac{1}{E}\sum_e \widehat{\theta}_e\right) = \frac{1}{E^2}\text{Var}\left(\mathbf{1}_E'\widehat{\theta}\right)\\
=&\frac{KT\lambda_1\lambda_2\sigma_y^2}{nE^2}\mathbf{1}_E'\left[\lambda_2\left(\mathbf{U}_1^{\otimes 2}+KT\mathbf{U}_2-T\mathbf{W}_1-K\mathbf{W}_2\right)+\lambda_1\left(K\mathbf{W}_2-\mathbf{U}_1^{\otimes 2}\right)\right]^{-1}\mathbf{1}_E.
\end{aligned}
\tag{S8}
$$

To explicate the design constants needed in the final variance expression (S8), we given a specific example below. We assume a standard stepped wedge design with 1 baseline period, and one cluster crosses over to intervention during each period. In this setup, we can define the exposure-time treatment indicator matrix for each cluster as $\mathbf{Z}_k$. This is a matrix of dimension $T \times E$ in a standard design where the earliest treatment timing is period 2. Then, if cluster k receives treatment at time 2, the form of this design matrix is written as

$$\mathbf{Z}_k = \begin{pmatrix} 0 & 0 & \ldots & 0 & 0 \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 1 \end{pmatrix}$$

If cluster k receives treatment at time 3, the form of this design matrix is written as

$$\mathbf{Z}_k = \begin{pmatrix} 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots. & 0 & 0 \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix}$$

And we define the period when cluster k initiates treatment is $s_k \in \{2, \ldots, T\}$. And without loss of generality, we assume there is only one baseline period, and we order k such that cluster 1 has the smallest s_1 (start treatment earliest), and cluster K has the largest s_K (start treatment the latest); in other words, s_k is increasing in k. So there are $s_k - 1$ rows of zeros and $T - s_k + 1$ rows of basis vector. We can write

$$\mathbf{Z}_k = \begin{pmatrix} \mathbf{0}' \\ \mathbf{0}' \\ \vdots \\ \mathbf{e}_1' \\ \mathbf{e}_2' \\ \vdots \\ \mathbf{e}_{T-s_k+1}' \end{pmatrix}$$

Where $\mathbf{e}_l$ is the $E \times 1$ orthonormal basis with the l-th element equal to 1 and zero everywhere else. Then notice the following intermediate steps

$$\mathbf{Z}_k'\mathbf{Z}_k = \sum_{l=1}^{T-s_k+1} \mathbf{e}_l\mathbf{e}_l'$$

$$\mathbf{Z}_k'\mathbf{1}_T = \sum_{l=1}^{T-s_k+1} \mathbf{e}_l$$

$$\mathbf{Z}_k'\mathbf{J}_T\mathbf{Z}_k = \mathbf{Z}_k'\mathbf{1}_T\left(\mathbf{Z}_k'\mathbf{1}_T\right)' = \left(\sum_{l=1}^{T-s_k+1} \mathbf{e}_l\right)\left(\sum_{l=1}^{T-s_k+1} \mathbf{e}_l\right)'$$

Then

$$\mathbf{U}_1 = \sum_{k=1}^{K}\sum_{l=1}^{T-s_k+1} \mathbf{e}_l$$

$$\mathbf{U}_2 = \sum_{k=1}^{K}\sum_{l=1}^{T-s_k+1} \mathbf{e}_l\mathbf{e}_l'$$

$$\mathbf{W}_2 = \sum_{k=1}^{K}\left(\sum_{l=1}^{T-s_k+1} \mathbf{e}_l\right)\left(\sum_{l=1}^{T-s_k+1} \mathbf{e}_l\right)'$$

$$\mathbf{W}_1 = \sum_{k=1}^{K}\left(\sum_{l=1}^{T-s_k+1} \mathbf{e}_l\right)\left(\sum_{l=1}^{T-s_k+1} \mathbf{e}_l\right)' = \mathbf{W}_2$$

Thus, in this standard design, we have $\mathbf{W}_1 = \mathbf{W}_2$ and the variance can be further written as

$$\mathrm{Var}\left(\frac{1}{E}\sum_{e}\widehat{\theta}_e\right) = \frac{KT\lambda_1\lambda_2\sigma_y^2}{nE^2}\mathbf{1}_E'\left[\lambda_2\left(\mathbf{U}_1^{\otimes 2} + KT\mathbf{U}_2 - T\mathbf{W}_1 - K\mathbf{W}_1\right) + \lambda_1\left(K\mathbf{W}_1 - \mathbf{U}_1^{\otimes 2}\right)\right]^{-1}\mathbf{1}_E$$

With a binary outcome, the variance expression of the average treatment effect estimator in a generalized linear mixed model is more complicated. One possible direction is to follow the general linearization approach outlined in Section 3.2 in Davis-Plourde *et al.* (2021). The resulting expression can again be approximated by a weighted least squares variance but is analytically less tractable because the binomial variance is an explicit function of the mean.

F.2 An illustration of trial planning

F.2.1 Planning a trial with continuous outcomes

In this section, we illustrate the planning of a stepped wedge CRT with continuous outcomes using our derived analytical variance expression (S8) for the average treatment effect estimate based on Model 4 (the general time-on-treatment effect formulation). As noted in the main paper, this formula does not require information about the exposure-time specific treatment effects, and only concerns the design resources, randomization schedule, and the ICC parameter ρ.

By way of example, suppose we are designing a trial with $T = 8$ steps, $E = 7$ exposure times, and $K = 14$ clusters as in the XpertMTB/RIF Tuberculosis study described in Section 3 of the main article. Suppose further that the primary outcome Y_{kti} is continuous and follows Model 4. Then, the analytic variance formula (S8) can be used to facilitate power calculation. We provide an R function *ATE.fixed.power* in the Supporting Information for its implementation. This function first computes the analytical variance from (S8) using the input parameters described in Table S4. It then uses a Wald test to compute the power of such a trial to detect a pre-specified average treatment effect.

Table S4: Arguments for *ATE.fixed.power*

Argument	Description
`ate`	Δ, average treatment effect
`t_max`	T, number of periods
`k_max`	K, number of clusters
`n_per`	n, constant cluster-period size
`crossover_t`	Vector of length K indicating the crossover time of each cluster
`rho`	ρ, outcome ICC for Model 4, defined by $\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2}$
`sigma_epsilon_sq`	σ_ϵ^2, individual-level random error
`alpha`	Type I error (defaults to 0.05)

Assuming that $\rho = 0.01$, $\sigma_\epsilon^2 = 1$, and $n = 34$, the code below returns the study power for detecting an average treatment effect of 0.206 with significance level of 0.05:

```
ATE.fixed.power(ate = 0.206, t_max = 8, k_max = 14, n_per = 34,
    crossover_t = rep(2:8,each=2), rho = 0.1, sigma_epsilon_sq = 1, alpha = 0.05)
```

In practice, because the outcome ICC is often unknown, we recommend examining settings with a variety of parameter values. To illustrate this, we explore the detectable effect sizes of the hypothetical stepped wedge CRT above for varying outcome ICC. Table S5 shows the minimum detectable average treatment effect size in order for the trial to have 80% power when analyzed via Model 4 at a significance level of 0.05.

Table S5: Detectable effect size with 80% power

Outcome ICC (ρ)	Δ
0.00	0.143
0.01	0.206
0.05	0.246
0.10	0.256
0.20	0.261

F.2.2 Planning a trial with binary outcomes

In this section, we illustrate the use of simulation-based methods, implemented in an R function *ATE.sim.power.binary* provided in the Supporting Information, to aid in the design of stepped wedge CRTs with binary outcomes. This function approximates the power to detect a pre-specified average treatment effect for analysis with either Model 4 (the general time-on-treatment effect formulation) or Model 5 (our proposed model formulation which uses random effects to model treatment effect heterogeneity). In practice, one would pre-specify which model they plan to use beforehand and use the power estimate from *ATE.sim.power.binary* which corresponds to their respective model. The function *ATE.sim.power.binary* first generates data from Model 4 using the formula logit $\{P(Y_{kti} = 1)\} = \mu + \beta_t + \theta_{E_{kt}} + \alpha_k$, where where μ is the intercept, β_t $(t = 1, \ldots, T)$ is the fixed effect for background calendar time with $\beta_1 = 0$ for identifiability, $\theta_{E_{kt}}$ is the treatment effect as a function of the exposure time, and $\alpha_k \sim \mathcal{N}(0, \sigma^2_\alpha)$ is a random intercept that allows for correlation within clusters. We then use the simulated data to fit both Models 4 and 5, and record their respective estimated average treatment effects. After many iterations, we take the empirical variance of the average treatment effect estimates for each model. Using each empirical variance, we then use a Wald test to approximate the power to detect a pre-specified average treatment effect for both Models 4 and 5. The required input parameters are provided in Table S6. Unlike the analytical formula (S8) for continuous outcomes, this function requires apriori full model specification as well as specification of the design parameters and randomization schedule to simulate trial data.

Table S6: Arguments for *ATE.sim.power.binary*

Argument	Description
`t_max`	T, number of periods
`k_max`	K, number of clusters
`n_per`	n, constant cluster-period size or a vector of cluster-period sizes
`crossover_t`	Vector of length K indicating the crossover time of each cluster
`intercept`	μ, intercept
`betas`	$(\beta_2, \ldots, \beta_T)$, background calendar time trend
`expt_eff`	$(\theta_1, \ldots, \theta_E)$, exposure-time specific with mean Δ
`sigma_alpha_sq`	σ^2_α, cluster-level heterogeneity
`nsims`	Number of simulation iterations
`alpha`	Type I error (defaults to 0.05)

We use *ATE.sim.power.binary* to calculate power of a hypothetical stepped wedge CRT as in the XpertMTB/RIF Tuberculosis study with $T = 8$ steps, $E = 7$ exposure times, and $K = 14$ clusters. The number of individuals per cluster-period “n_per” (median [range]: 34 [6-96]) are set to be the same as in the trial. The average treatment effect “ate”, background calendar time trend “betas”, and “sigma_alpha_sq” are estimated from the trial data. The exposure-time specific treatment effect parameters “expt_eff” are set to be the same ones used in the synthetic data with induced heterogeneity as described in Section 3 of the main article. The code below returns the power of such a trial to detect a pre-specified average treatment effect estimated via either Model 4 or Model 5 based on 1000 simulated datasets.

```
ATE.sim.power.binary(t_max = 8, k_max = 14, n_per, crossover_t = rep(2:8,each=2),
    intercept = 0,  betas = c(-0.0134, 0.1667, 0.0365, -0.0660, -0.1915, -0.1881,
    0.0059), expt_eff = c(-0.375, -0.252,  0.412, -0.141, -0.119, 0.471,  0.004),
    sigma_alpha_sq = 0.34, nsims = 1000, alpha = 0.05)
```

As shown in Table S7, when the pre-specified average log-odds ratio is log(1.19), the trial only has 7.52% power for testing the null hypothesis of no average treatment effect with a significance level of 0.05 when analyzed using Model 4. The power slightly increases to 7.77% when data are analyzed using Model 5. When the pre-specified average log-odds ratio is log(3), the power to detect a significant treatment effect is about 78% based on Model 4 and 82% based on Model 5, both with significance level 0.05.

Table S7: Simulation-based power estimates based on Models 4 and 5.

Average Treatment Effect (Δ)	Model 4	Model 5
log(1.19)	7.52	7.77
log(3)	78.15	82.31

Web Appendix G. Schematic illustration of different modelling assumptions

Figure S5 shows a schematic comparison of the intervention effects in the three modelling frameworks. Here, Model 5 focuses on horizontal treatment heterogeneity across exposure time, the Hemming *et al.* (2018) model allows vertical treatment effect heterogeneity across clusters, and Model 6 incorporates both horizontal and vertical treatment effect heterogeneity.

Figure S5: Schematic illustrations of differing intervention effects in a stepped wedge design based on three models with $K = 4$ clusters and $T = 5$ periods. Each cell with a zero entry indicates a control cluster-period and each cell with a non-zero entry indicates an intervention cluster-period.

(a) *Model 5:* $\theta(t,k) = \phi + \delta_{E_{tk}}$

$k=1$	0	$\phi+\delta_1$	$\phi+\delta_2$	$\phi+\delta_3$	$\phi+\delta_4$
$k=2$	0	0	$\phi+\delta_1$	$\phi+\delta_2$	$\phi+\delta_3$
$k=3$	0	0	0	$\phi+\delta_1$	$\phi+\delta_2$
$k=4$	0	0	0	0	$\phi+\delta_1$

(b) *Hemming et al.:* $\theta(t,k) = \psi + \nu_k$

$k=1$	0	$\psi+\nu_1$	$\psi+\nu_1$	$\psi+\nu_1$	$\psi+\nu_1$
$k=2$	0	0	$\psi+\nu_2$	$\psi+\nu_2$	$\psi+\nu_2$
$k=3$	0	0	0	$\psi+\nu_3$	$\psi+\nu_3$
$k=4$	0	0	0	0	$\psi+\nu_4$

(c) *Model 6:* $\theta(t,k) = \pi + \delta_{E_{tk}} + \nu_k$

$k=1$	0	$\pi+\delta_1+\nu_1$	$\pi+\delta_2+\nu_1$	$\pi+\delta_3+\nu_1$	$\pi+\delta_4+\nu_1$
$k=2$	0	0	$\pi+\delta_1+\nu_2$	$\pi+\delta_2+\nu_2$	$\pi+\delta_3+\nu_2$
$k=3$	0	0	0	$\pi+\delta_1+\nu_3$	$\pi+\delta_2+\nu_3$
$k=4$	0	0	0	0	$\pi+\delta_1+\nu_4$

References

Davis-Plourde, K., Taljaard, M. and Li, F. (2021) Sample size considerations for stepped wedge designs with subclusters. *Biometrics*. Online Early View.

Hemming, K., Taljaard, M., and Forbes, A. (2018) Modeling clustering and treatment effect heterogeneity in parallel and stepped-wedge cluster randomized trials. *Statistics in medicine*, 37(6), 883–898.

Hussey, M. A. and Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, 28(2), 182-191.

Li, F., Hughes, J.P., Hemming, K., Taljaard, M., Melnick, E.R. and Heagerty, P.J. (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.

Sitlani, C. M., Heagerty, P. J., Blood, E. A., Tosteson, T. D. (2012). Longitudinal structural mixed models for the analysis of surgical trials with noncompliance. *Statistics in Medicine*, 31(16), 1738-1760.

Trajman, A., Durovni, B., Saraceni, V., Menezes, A., Cordeiro-Santos, M., Cobelens, F., *et al.* (2015) Impact on patients' treatment outcomes of XpertMTB/RIF implementation for the diagnosis of tuberculosis: follow-up of a stepped-wedge randomized clinical trial. *PloS One*, 10(4), e0123252.

VanderWeele, T. J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology*, 20(6), 880-883.