Biostatistics (2020), **0**, 0, *pp.* **??**–29
doi:10.1093/biostatistics/output

Supplementary Material for "Marginal Modeling of Cluster-Period Means and Intraclass Correlations in Stepped Wedge Designs with Binary Outcomes"

FAN LI[1,2,*], HENGSHI YU[3], PAUL J. RATHOUZ[4], ELIZABETH L. TURNER[5], JOHN S. PREISSER[6]

[1] *Department of Biostatistics, Yale School of Public Health, New Haven, CT, U.S.A.*
[2] *Center for Methods in Implementation and Prevention Science, Yale School of Public Health, New Haven, CT, U.S.A.*
[3] *Department of Biostatistics, University of Michigan, Ann Arbor, MI, U.S.A.*
[4] *Department of Population Health, The University of Texas at Austin, Austin, TX, U.S.A.*
[5] *Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, U.S.A.*
[6] *Department of Biostatistics, University of North Carolina, Chapel Hill, NC, U.S.A.*

*fan.f.li@yale.edu

APPENDIX

A. Aggregated Linear Mixed Models Do Not Estimate Valid Individual-level ICCs

We provide intuitions as to why the Hussey and Hughes (2007) linear mixed model based on aggregated cluster-period means does not estimate valid ICCs defined from individual-level data with variable cluster sizes. For simplicity, we assume a continuous outcome Y_{ijk}, measured for individual k during period j in cluster i. A linear mixed model defined on individual-level data

*To whom correspondence should be addressed.

is (Hussey and Hughes, 2007):

$$Y_{ijk} = \beta_j + \delta X_{ij} + b_i + \epsilon_{ijk}, \quad b_i \sim N(0, \sigma_b^2), \quad \epsilon_{ijk} \sim N(0, \sigma_\epsilon^2) \tag{A.1}$$

where the random intercept and error are assumed independent, $b_i \perp \epsilon_{ijk}$. This model implies a simple exchangeable correlation structure (Li *and others*, 2020), with a common (individual-level) ICC, $\alpha = \sigma_b^2/(\sigma_b^2 + \sigma_\epsilon^2)$. When collapsing over cluster-periods, the induced cluster-period mean model becomes

$$\overline{Y}_{ij} = \beta_j + \delta X_{ij} + b_i + \bar{\epsilon}_{ij}, \quad b_i \sim N(0, \sigma_b^2), \quad \bar{\epsilon}_{ij} \sim N(0, \sigma_\epsilon^2/n_{ij}). \tag{A.2}$$

The implied cluster-period mean model is clearly heteroskedastic if the n_{ij}'s vary across both i and j. However, the usual implementation assumes a homoscedastic model for the cluster-period means (Hussey and Hughes, 2007), that is

$$\overline{Y}_{ij} = \beta_j + \delta X_{ij} + b_i + \bar{\epsilon}^*_{ij}, \quad b_i \sim N(0, \sigma_b^2), \quad \bar{\epsilon}^*_{ij} \sim N(0, \tau^2). \tag{A.3}$$

This is equivalent to assuming (A.2) but put a common variance on the residual for cluster-period means. Such a homoscedastic cluster-period mean model estimates a common ICC for $\overline{Y}_{ij}$, as $\alpha^* = \sigma_b^2/(\sigma_b^2 + \tau^2)$. If the individual-level model (A.1) is true, model (A.3) is misspecified but τ^2 is still well defined asymptotically. By the results of White (1982), τ^2 is least favorable value that minimizes the Kullback-Leibler distance between the probability distribution based on the misspecified model and the true model (A.1). Importantly, with variable cluster-period sizes, τ^2 will be a complex function of n_{ij} and σ_ϵ^2, and therefore $\alpha^* \neq \alpha$. One can also provide an approximate relationship between τ^2, n_{ij} and σ_ϵ^2 by the Law of total variance $\tau^2 \approx \sigma_\epsilon^2 E(n_{ij}^{-1}) \neq \sigma_\epsilon^2$, and show $\alpha^* > \alpha$. These argumemts provide an intuitive explanation for why model (A.3) does not provide a valid estimate of the individual-level ICC. The argument for binary outcomes follow the exact same logic.

B. Extensions to Continuous and Count Outcomes

Marginal model (1.1) in the main text is a generalized linear formulation that also accommodates continuous and count outcomes, providing the variance functions are appropriately chosen. However, the estimating equations for the correlation parameters may require additional considerations, as a common dispersion parameter $\phi > 0$ is necessary to characterize these types of outcomes. While our motivating example concerns binary outcomes, for completeness we outline an extension to continuous and count outcomes based on cluster-period means.

We define the variance of the individual-level outcome as $\phi\nu_{ij}$, where ν_{ij} is a function of the marginal mean. For example, a Gaussian variance function $\nu_{ij} = 1$ may be used for continuous outcomes, while the Poisson variance $\nu_{ij} = \mu_{ij}$ or the negative binomial variance $\nu_{ij} = \mu_{ij} + \mu_{ij}^2/\omega$ for some $\omega > 0$ may be chosen for count outcomes. When the correlation structure for individual-level outcomes is nested exchangeable, the diagonal element of $\boldsymbol{V}_{1i}$ is

$$\sigma_{ijj} = \text{var}(\overline{Y}_{ij+}) = \frac{\phi\nu_{ij}}{n_{ij}} \left\{1 + (n_{ij} - 1)\alpha_0\right\}, \tag{B.1}$$

and the off-diagonal elements is $\sigma_{ijl} = \text{cov}(\overline{Y}_{ij+}, \overline{Y}_{il+}) = \sqrt{\nu_{ij}\nu_{il}}\phi\alpha_1$. By defining $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \phi)^T$, we can still solve estimating equations (2.5) in the main text to obtain $\hat{\boldsymbol{\alpha}}$. In this case, the updates for the two ICC parameters become

$$\hat{\alpha}_0 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{n_{ij}-1}{n_{ij}}\right)\left(\hat{\phi}^{-1}\tilde{s}_{ijj}\hat{\nu}_{ij} - \frac{\hat{\nu}_{ij}^2}{n_{ij}}\right)}{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{n_{ij}-1}{n_{ij}}\right)^2\hat{\nu}_{ij}^2}, \quad \hat{\alpha}_1 = \frac{\sum_{i=1}^{I}\sum_{j\neq l}\hat{\phi}^{-1}\tilde{s}_{ijl}\sqrt{\hat{\nu}_{ij}\hat{\nu}_{il}}}{\sum_{i=1}^{I}\sum_{j\neq l}\hat{\nu}_{ij}\hat{\nu}_{il}},$$

and an additional update for the dispersion parameter is required by

$$\hat{\phi} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_{ij}}\tilde{s}_{ijj}\hat{\nu}_{ij}\left\{1 + (n_{ij} - 1)\hat{\alpha}_0\right\}/n_{ij} + \sum_{i=1}^{I}\sum_{j\neq l}\tilde{s}_{ijl}\sqrt{\hat{\nu}_{ij}\hat{\nu}_{il}}\hat{\alpha}_1}{\sum_{i=1}^{I}\sum_{j=1}^{n_{ij}}\hat{\nu}_{ij}^2\left\{1 + (n_{ij} - 1)\hat{\alpha}_0\right\}^2/n_{ij}^2 + \sum_{i=1}^{I}\sum_{j\neq l}\hat{\nu}_{ij}\hat{\nu}_{il}\hat{\alpha}_1^2}$$

When the correlation structure for the individual-level outcomes exhibits exponential decay, the diagonal element of $\boldsymbol{V}_{1i}$ remain (B.1), but the off-diagonal element becomes $\sigma_{ijl} = \text{cov}(\overline{Y}_{ij+}, \overline{Y}_{il+}) = \sqrt{\nu_{ij}\nu_{il}}\phi\alpha_o\rho^{|j-l|}$. Following Section 2.2 of the main text, the updates for the two correlation parameters based on the $\boldsymbol{\alpha}$-estimating equations are given by (2.9) and (2.10), except that $\tilde{s}_{ijj}$ and $\tilde{s}_{ijk}$ are replaced by $\hat{\phi}^{-1}\tilde{s}_{ijj}$ and $\hat{\phi}^{-1}\tilde{s}_{ijk}$. Further, an additional step is required to update the dispersion parameter and is given by

$$\hat{\phi} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_{ij}} \tilde{s}_{ijj}\hat{\nu}_{ij}\left\{1+(n_{ij}-1)\hat{\alpha}_0\right\}/n_{ij} + \sum_{i=1}^{I}\sum_{j\neq l}\tilde{s}_{ijl}\sqrt{\hat{\nu}_{ij}\hat{\nu}_{il}}\hat{\alpha}_0\hat{\rho}^{|j-l|}}{\sum_{i=1}^{I}\sum_{j=1}^{n_{ij}} \hat{\nu}_{ij}^2\left\{1+(n_{ij}-1)\hat{\alpha}_0\right\}^2/n_{ij}^2 + \sum_{i=1}^{I}\sum_{j\neq l}\hat{\nu}_{ij}\hat{\nu}_{il}\hat{\alpha}_0^2\hat{\rho}^{2|j-l|}}.$$

The remaining steps follow the case with binary outcomes. Finally, the estimation and inference with binary outcomes can essentially be viewed as a special case where the common dispersion is fixed at $\phi = 1$.

C. Block Toeplitz Correlation Structure

For the purpose of making Theorem 3.1 broadly applicable to a general class of correlation structures, we introduce the notion of a block Toeplitz correlation structure. We characterize such a correlation structure based on individual-level outcomes; in other words, all correlation parameters defined below are specific to pairs of individual observations. The block Toeplitz structure assumes the jth within-period correlation as r_{jj}, and allows the between-period correlation to arbitrarily vary as function of distance in time (measured in discrete period). Specifically, the correlation between two outcomes measured in the jth and lth periods is r_{jl} and $r_{jl} = r_{lj}$. In matrix notations, the block Toeplitz correlation structure can be written as

$$\begin{bmatrix} (1-r_{11})\boldsymbol{I}_{n_{i1}} + r_{11}\boldsymbol{J}_{n_{i1}} & r_{12}\boldsymbol{J}_{n_{i1}\times n_{i2}} & r_{13}\boldsymbol{J}_{n_{i1}\times n_{i3}} & \cdots & r_{1J}\boldsymbol{J}_{n_{i1}\times n_{iJ}} \\ r_{12}\boldsymbol{J}_{n_{i2}\times n_{i1}} & (1-r_{22})\boldsymbol{I}_{n_{i2}} + r_{22}\boldsymbol{J}_{n_{i2}} & r_{23}\boldsymbol{J}_{n_{i2}\times n_{i3}} & \cdots & r_{2J}\boldsymbol{J}_{n_{i2}\times n_{iJ}} \\ r_{13}\boldsymbol{J}_{n_{i3}\times n_{i1}} & r_{23}\boldsymbol{J}_{n_{i3}\times n_{i2}} & (1-r_{33})\boldsymbol{I}_{n_{i3}} + r_{33}\boldsymbol{J}_{n_{i3}} & \cdots & r_{3J}\boldsymbol{J}_{n_{i3}\times n_{iJ}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{1J}\boldsymbol{J}_{n_{iJ}\times n_{i1}} & r_{2J}\boldsymbol{J}_{n_{iJ}\times n_{i2}} & r_{3J}\boldsymbol{J}_{n_{iJ}\times n_{i3}} & \cdots & (1-r_{JJ})\boldsymbol{I}_{n_{iJ}} + r_{JJ}\boldsymbol{J}_{n_{iJ}} \end{bmatrix},$$

where $\boldsymbol{I}_s$ is the $s\times s$ identity matrix, $\boldsymbol{J}_s$ is the $s \times s$ matrix of ones, and $\boldsymbol{J}_{s\times t}$ is the $s\times t$ matrix of ones. Current correlation matrix structures assumed for cross-sectional SW-CRTs are all special cases of this block Toeplitz structure. For example, the independence correlation matrix (identity matrix) is obtained with $r_{jl} = 0$ for all $j, l = 1, \ldots, J$. The simple exchangeable correlation matrix is obtained with $r_{jl} = \alpha_0$ for all $j, l = 1, \ldots, J$. The next two special cases both assume the within-period correlation is not a function of time so that $r_{jj} = \alpha_0$ for all $j = 1, \ldots, J$. Given this, the nested exchangeable correlation matrix is obtained with $r_{jl} = \alpha_1$ for all $|j-l| \geqslant 1$. The exponential decay correlation matrix is obtained with $r_{jl} = \alpha_0 \rho^{|j-l|}$ for all $|j-l| \geqslant 1$. We make a note here that although the block Toeplitz correlation structure is more general, it is rarely adopted in the design and analysis of SW-CRTs because it could involve up to $J(J-1)/2$ distinct correlation parameters and becomes challenging to interpret or estimate. For better interpretation, we will focus on the application of Theorem 3.1 on nested exchangeable and exponential decay correlation models.

D. Proof of Theorem 3.1

Recall that the marginal mean model is $g(\mu_{ijk}) = g(\mu_{ij}) = \beta_j + X_{ij}\delta$, and that the cluster-period-level estimating equations for the marginal mean are

$$\sum_{i=1}^{I} \boldsymbol{D}_{1i}^T \boldsymbol{V}_{1i}^{-1}(\overline{\boldsymbol{Y}}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}, \tag{D.1}$$

where $\overline{\boldsymbol{Y}}_i = (\overline{Y}_{i1}, \ldots, \overline{Y}_{iJ})^T$, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iJ})^T$, $\boldsymbol{D}_{1i} = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\theta}^T$ and $\boldsymbol{V}_{1i} = \text{cov}(\overline{\boldsymbol{Y}}_i)$ is the working variance matrix induced by the individual-level variances and pairwise correlations. Further, the individual-level estimating equations for the marginal mean is

$$\sum_{i=1}^{I} \boldsymbol{E}_{1i}^T \boldsymbol{M}_{1i}^{-1}(\boldsymbol{Y}_i - \boldsymbol{\vartheta}_i) = \boldsymbol{0}, \tag{D.2}$$

where $\boldsymbol{Y}_i = (Y_{i11}, Y_{i12}, \ldots, Y_{i21}, \ldots)^T$, $\boldsymbol{\vartheta}_i = (\mu_{i1}\boldsymbol{1}_{n_{i1}}^T, \ldots, \mu_{iJ}\boldsymbol{1}_{n_{iJ}}^T)^T$, $\boldsymbol{1}_s$ is the $s \times 1$ vector of ones, $\boldsymbol{M}_{1i} = \text{cov}(\boldsymbol{Y}_i)$, and $\boldsymbol{E}_{1i} = \partial \boldsymbol{\vartheta}_i / \partial \boldsymbol{\theta}^T$.

Define $\boldsymbol{G}_i = \oplus_{j=1}^{J} \boldsymbol{1}_{n_{ij}}$, where "$\oplus$" is the block diagonal operator with nonzero matrices along the diagonal block and zero values everywhere else, and let $\boldsymbol{L}_i = \text{diag}(n_{i1}, \ldots, n_{iJ})$, then one can verify $\boldsymbol{G}_i^T \boldsymbol{G}_i = \boldsymbol{L}_i$. Further, we have $\boldsymbol{E}_{1i} = \boldsymbol{G}_i \boldsymbol{D}_{1i}$ and $(\overline{\boldsymbol{Y}}_i - \boldsymbol{\mu}_i) = \boldsymbol{L}_i^{-1} \boldsymbol{G}_i^T (\boldsymbol{Y}_i - \boldsymbol{\vartheta}_i)$. By definition of $\boldsymbol{V}_{1i}$, we also have $\boldsymbol{V}_{1i} = \boldsymbol{L}_i^{-1} \boldsymbol{G}_i^T \boldsymbol{M}_{1i} \boldsymbol{G}_i \boldsymbol{L}_i^{-1}$. These relationships allow us to rewrite the summand in (D.1) and (D.2) as

$$\boldsymbol{D}_{1i}^T\boldsymbol{V}_{1i}^{-1}(\overline{\boldsymbol{Y}}_i-\boldsymbol{\mu}_i)=\boldsymbol{D}_{1i}^T\boldsymbol{L}_i(\boldsymbol{G}_i^T\boldsymbol{M}_{1i}\boldsymbol{G}_i)^{-1}\boldsymbol{G}_i^T(\boldsymbol{Y}_i-\boldsymbol{\vartheta}_i)$$

$$\boldsymbol{E}_{1i}^T\boldsymbol{M}_{1i}^{-1}(\boldsymbol{Y}_i-\boldsymbol{\vartheta}_i)=\boldsymbol{D}_{1i}^T\boldsymbol{G}_i^T\boldsymbol{M}_{1i}^{-1}(\boldsymbol{Y}_i-\boldsymbol{\vartheta}_i)$$

Therefore it is sufficient to show

$$\boldsymbol{L}_i(\boldsymbol{G}_i^T\boldsymbol{M}_{1i}\boldsymbol{G}_i)^{-1}\boldsymbol{G}_i^T=\boldsymbol{G}_i^T\boldsymbol{M}_{1i}^{-1}\quad\forall\, i=1,\ldots,I \tag{D.3}$$

to establish Theorem 3.1.

D.1 *The Base Step with $J=2$*

We first assume that cluster i has only $J=2$ periods. This is of course an unrealistic assumption in stepped wedge designs (since $J\geqslant 3$), but serves as a base step for us to proceed by mathematical induction. Partition the covariance matrix $\boldsymbol{M}_{1i}$ and its inverse by cluster-periods so that

$$\boldsymbol{M}_{1i}=\begin{bmatrix}\boldsymbol{N}_{i[J]} & \boldsymbol{N}_{i[J]J}\\ \boldsymbol{N}_{iJ[J]} & \boldsymbol{N}_{iJ}\end{bmatrix}\qquad \boldsymbol{M}_{1i}^{-1}=\begin{bmatrix}\boldsymbol{W}_{i[J]} & \boldsymbol{W}_{i[J]J}\\ \boldsymbol{W}_{iJ[J]} & \boldsymbol{W}_{iJ}\end{bmatrix}$$

In the partition for $\boldsymbol{M}_{1i}$, $\boldsymbol{N}_{iJ}$ represents the variance matrix of all observations corresponding to the last period $J=2$; $\boldsymbol{N}_{i[J]}$ represents the variance matrix of the remaining elements in $\boldsymbol{Y}_i$, which includes all observations in the first period because $J=2$. The off-diagonal blocks are covariance matrices for observations collected in the two different periods. The same rule applies to the partition for $\boldsymbol{M}_{1i}^{-1}$. By block matrix inversion (Graybill, 1983), the following explicit relationships hold for the partitioned matrices:

$$\boldsymbol{W}_{i[J]}=(\boldsymbol{N}_{i[J]}-\boldsymbol{N}_{i[J]J}\boldsymbol{N}_{iJ}^{-1}\boldsymbol{N}_{iJ[J]})^{-1}$$

$$\boldsymbol{W}_{iJ}=(\boldsymbol{N}_{iJ}-\boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J})^{-1}$$

$$\boldsymbol{W}_{i[J]J}=\boldsymbol{W}_{iJ[J]}^T=-\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J}(\boldsymbol{N}_{iJ}-\boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J})^{-1}$$

$$\boldsymbol{W}_{iJ[J]}=\boldsymbol{W}_{i[J]J}^T=-\boldsymbol{N}_{iJ}^{-1}\boldsymbol{N}_{iJ[J]}(\boldsymbol{N}_{i[J]}-\boldsymbol{N}_{i[J]J}\boldsymbol{N}_{iJ}^{-1}\boldsymbol{N}_{iJ[J]})^{-1}$$

Similarly, we partition matrix $\boldsymbol{G}_i$ by cluster-periods and so

$$\boldsymbol{G}_i = \begin{bmatrix} \boldsymbol{G}_{i[J]} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}_{iJ} \end{bmatrix},$$

where $\boldsymbol{G}_{i[J]} = \boldsymbol{1}_{n_{i1}}$ and $\boldsymbol{G}_{iJ} = \boldsymbol{1}_{n_{i2}}$. Direct matrix inversion and multiplication then gives

$$\boldsymbol{L}_i(\boldsymbol{G}_i^T \boldsymbol{M}_{1i} \boldsymbol{G}_i)^{-1} = \frac{1}{\pi} \begin{bmatrix} n_{i1}(\boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ} \boldsymbol{G}_{iJ}) & -n_{i1}(\boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]J} \boldsymbol{G}_{iJ}) \\ -n_{i2}(\boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ[J]} \boldsymbol{G}_{i[J]}) & n_{i2}(\boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]} \boldsymbol{G}_{i[J]}) \end{bmatrix},$$

where $\pi = (\boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]} \boldsymbol{G}_{i[J]})(\boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ} \boldsymbol{G}_{iJ}) - (\boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]J} \boldsymbol{G}_{iJ})(\boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ[J]} \boldsymbol{G}_{i[J]})$,

$$(\boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]} \boldsymbol{G}_{i[J]}) = n_{i1}\nu_{i1}\{1 + (n_{i1} - 1)r_{11}\}$$

$$(\boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ} \boldsymbol{G}_{iJ}) = n_{i2}\nu_{i2}\{1 + (n_{i2} - 1)r_{22}\}$$

$$(\boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]J} \boldsymbol{G}_{iJ}) = (\boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ[J]} \boldsymbol{G}_{i[J]}) = n_{i1}n_{i2}\sqrt{\nu_{i1}\nu_{i2}}r_{12},$$

and r_{11}, r_{22}, r_{12} are defined in Appendix C. Further observe that

$$\boldsymbol{G}_i^T \boldsymbol{M}_{1i} = \begin{bmatrix} \boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]} & \boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]J} \\ \boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ[J]} & \boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ} \end{bmatrix},$$

where

$$\boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]} = \nu_{i1}\{1 + (n_{i1} - 1)r_{11}\}\boldsymbol{G}_{i[J]}^T, \qquad \boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]J} = n_{i1}\sqrt{\nu_{i1}\nu_{i2}}r_{12}\boldsymbol{G}_{iJ}^T$$

$$\boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ[J]} = n_{i2}\sqrt{\nu_{i1}\nu_{i2}}r_{12}\boldsymbol{G}_{i[J]}^T, \qquad \boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ} = \nu_{i2}\{1 + (n_{i2} - 1)r_{22}\}\boldsymbol{G}_{iJ}^T,$$

By direct matrix multiplication, we can verify that $\boldsymbol{L}_i(\boldsymbol{G}_i^T \boldsymbol{M}_{1i} \boldsymbol{G}_i)^{-1} \boldsymbol{G}_i^T \boldsymbol{M}_{1i} = \boldsymbol{G}_i^T$, by which equation (D.3) holds as $\boldsymbol{M}_{1i}$ is a positive definite variance matrix.

D.2 *The Induction Step for $J \geqslant 3$*

For general number of periods $J \geqslant 3$, we show that equation (D.3) holds by induction. Specifically, for $J \geqslant 3$, we partition the covariance matrix and its inverse so that

$$\boldsymbol{M}_{1i} = \begin{bmatrix} \boldsymbol{N}_{i[J]} & \boldsymbol{N}_{i[J]J} \\ \boldsymbol{N}_{iJ[J]} & \boldsymbol{N}_{iJ} \end{bmatrix} \qquad \boldsymbol{M}_{1i}^{-1} = \begin{bmatrix} \boldsymbol{W}_{i[J]} & \boldsymbol{W}_{i[J]J} \\ \boldsymbol{W}_{iJ[J]} & \boldsymbol{W}_{iJ} \end{bmatrix}$$

In the partition for $\boldsymbol{M}_{1i}$, $\boldsymbol{N}_{iJ}$ represents the variance matrix of all observations corresponding to the last period (period J); $\boldsymbol{N}_{i[J]}$ represents the variance matrix of the remaining elements in $\boldsymbol{Y}_i$, which includes all observations in the first $J-1$ periods. The off-diagonal blocks correspond to covariance matrices for observations collected in different sets of periods. The same rule applies to the partition for $\boldsymbol{M}_{1i}^{-1}$. We similarly partition matrix $\boldsymbol{G}_i$ and $\boldsymbol{L}_i$ such that

$$\boldsymbol{G}_i = \begin{bmatrix} \boldsymbol{G}_{i[J]} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}_{iJ} \end{bmatrix} \quad \boldsymbol{L}_i = \begin{bmatrix} \boldsymbol{L}_{i[J]} & \boldsymbol{0} \\ \boldsymbol{0} & n_{iJ} \end{bmatrix}$$

where $\boldsymbol{G}_{i[J]} = \oplus_{j=1}^{J-1} \boldsymbol{1}_{n_{ij}}$, $\boldsymbol{G}_{iJ} = \boldsymbol{1}_{n_{iJ}}$, $\boldsymbol{L}_{i[J]} = \text{diag}(n_{i1}, \ldots, n_{i,J-1})$.

For the induction hypothesis, we assume that equation (D.3) holds if we omit the last period (and so the total number of periods becomes $J-1$). In other words, we assume

$$\boldsymbol{L}_{i[J]}(\boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]} \boldsymbol{G}_{i[J]})^{-1} \boldsymbol{G}_{i[J]}^T = \boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]}^{-1} \tag{D.4}$$

and we aim to show $\boldsymbol{L}_i(\boldsymbol{G}_i^T \boldsymbol{M}_{1i} \boldsymbol{G}_i)^{-1} \boldsymbol{G}_i^T = \boldsymbol{G}_i^T \boldsymbol{M}_{1i}^{-1}$. Write

$$\begin{aligned}
\boldsymbol{L}_i(\boldsymbol{G}_i^T \boldsymbol{M}_{1i} \boldsymbol{G}_i)^{-1} \boldsymbol{G}_i^T &= \boldsymbol{L}_i \begin{bmatrix} \boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]} \boldsymbol{G}_{i[J]} & \boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]J} \boldsymbol{G}_{iJ} \\ \boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ[J]} \boldsymbol{G}_{i[J]} & \boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ} \boldsymbol{G}_{iJ} \end{bmatrix}^{-1} \boldsymbol{G}_i^T \\
=& \boldsymbol{L}_i \begin{bmatrix} \boldsymbol{\Lambda}_{i[J]} & \boldsymbol{\Lambda}_{i[J]J} \\ \boldsymbol{\Lambda}_{iJ[J]} & \boldsymbol{\Lambda}_{iJ} \end{bmatrix} \boldsymbol{G}_i^T = \begin{bmatrix} \boldsymbol{L}_{i[J]} \boldsymbol{\Lambda}_{i[J]} \boldsymbol{G}_{i[J]}^T & \boldsymbol{L}_{i[J]} \boldsymbol{\Lambda}_{i[J]J} \boldsymbol{G}_{iJ}^T \\ n_{iJ} \boldsymbol{\Lambda}_{iJ[J]} \boldsymbol{G}_{i[J]}^T & n_{iJ} \boldsymbol{\Lambda}_{iJ} \boldsymbol{G}_{iJ}^T \end{bmatrix},
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\Lambda}_{i[J]} &= \left\{ \boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]} \boldsymbol{G}_{i[J]} - \boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]J} \boldsymbol{G}_{iJ} \left(\boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ} \boldsymbol{G}_{iJ} \right)^{-1} \boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ[J]} \boldsymbol{G}_{i[J]} \right\}^{-1} \\
\boldsymbol{\Lambda}_{iJ} &= \left\{ \boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ} \boldsymbol{G}_{iJ} - \boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ[J]} \boldsymbol{G}_{i[J]} \left(\boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]} \boldsymbol{G}_{i[J]} \right)^{-1} \boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]J} \boldsymbol{G}_{iJ} \right\}^{-1} \\
\boldsymbol{\Lambda}_{i[J]J} &= \boldsymbol{\Lambda}_{iJ[J]}^T = -(\boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]} \boldsymbol{G}_{i[J]})^{-1} \boldsymbol{G}_{i[J]}^T \boldsymbol{N}_{i[J]J} \boldsymbol{G}_{iJ} \times \boldsymbol{\Lambda}_{iJ} \\
\boldsymbol{\Lambda}_{iJ[J]} &= \boldsymbol{\Lambda}_{i[J]J}^T = -\left(\boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ} \boldsymbol{G}_{iJ} \right)^{-1} \boldsymbol{G}_{iJ}^T \boldsymbol{N}_{iJ[J]} \boldsymbol{G}_{i[J]} \times \boldsymbol{\Lambda}_{i[J]}
\end{aligned}$$

We next look at each block of the right-hand side matrix separately.

(i) Upper left block $\boldsymbol{L}_{i[J]} \boldsymbol{\Lambda}_{i[J]} \boldsymbol{G}_{i[J]}^T$

Recall that the Woodbury matrix identity (Henderson and Searle, 1981) is

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U} \left(\boldsymbol{B}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U} \right)^{-1} \boldsymbol{V}\boldsymbol{A}^{-1}, \tag{D.5}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are invertible matrices. This Woodbury matrix identity is used to expand $\boldsymbol{\Lambda}_{i[J]}$, and using the induction hypothesis (D.4), we have

$$
\begin{aligned}
\boldsymbol{L}_{i[J]}\boldsymbol{\Lambda}_{i[J]}\boldsymbol{G}_{i[J]}^T =& \boldsymbol{L}_{i[J]}(\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}\boldsymbol{G}_{i[J]})^{-1}\boldsymbol{G}_{i[J]}^T + \boldsymbol{L}_{i[J]}(\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}\boldsymbol{G}_{i[J]})^{-1}\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]J}\boldsymbol{G}_{iJ}\times \\
& \left\{\boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ}\boldsymbol{G}_{iJ} - \boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ[J]}\boldsymbol{G}_{i[J]}(\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}\boldsymbol{G}_{i[J]})^{-1}\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]J}\boldsymbol{G}_{iJ}\right\}^{-1}\times \\
& \boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ[J]}\boldsymbol{G}_{i[J]}(\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}\boldsymbol{G}_{i[J]})^{-1}\boldsymbol{G}_{i[J]}^T \\
=& \boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}^{-1} + \boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J}\boldsymbol{G}_{iJ}\times \\
& \left\{\boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ}\boldsymbol{G}_{iJ} - \boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ[J]}\boldsymbol{G}_{i[J]}(\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}\boldsymbol{G}_{i[J]})^{-1}\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]J}\boldsymbol{G}_{iJ}\right\}^{-1}\times \\
& \boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ[J]}\boldsymbol{G}_{i[J]}\boldsymbol{L}_{i[J]}^{-1}\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}^{-1}
\end{aligned}
$$

Recall that in the block Toeplitz structure, the sub-matrix

$$
\boldsymbol{N}_{iJ[J]} = \boldsymbol{N}_{i[J]J}^T = \left[\begin{array}{cccc} \sqrt{\nu_{i1}\nu_{iJ}}r_{1J}\boldsymbol{J}_{n_{iJ}\times n_{i1}} & \sqrt{\nu_{i2}\nu_{iJ}}r_{2J}\boldsymbol{J}_{n_{iJ}\times n_{i2}} & \dots & \sqrt{\nu_{i,J-1}\nu_{iJ}}r_{J-1,J}\boldsymbol{J}_{n_{iJ}\times n_{i,J-1}} \end{array}\right]
$$

Because of this special structure, the following set of matrix identities hold:

$$
\boldsymbol{N}_{i[J]J} = \boldsymbol{G}_{i[J]}(\boldsymbol{L}_{i[J]}^{-1}\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]J}\boldsymbol{G}_{iJ}n_{iJ}^{-1})\boldsymbol{G}_{iJ}^T \tag{D.6}
$$

$$
\boldsymbol{N}_{iJ[J]} = \boldsymbol{G}_{iJ}(n_{iJ}^{-1}\boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ[J]}\boldsymbol{G}_{i[J]}\boldsymbol{L}_{i[J]}^{-1})\boldsymbol{G}_{i[J]}^T \tag{D.7}
$$

$$
\boldsymbol{N}_{i[J]J} = \boldsymbol{N}_{i[J]J}\boldsymbol{G}_{iJ}n_{iJ}^{-1}\boldsymbol{G}_{iJ}^T \tag{D.8}
$$

$$
\boldsymbol{N}_{iJ[J]} = \boldsymbol{G}_{iJ}n_{iJ}^{-1}\boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ[J]}. \tag{D.9}
$$

Identity (D.6) and (D.7) lead to the following two useful identities:

$$
\boldsymbol{N}_{iJ[J]}\boldsymbol{G}_{i[J]}\boldsymbol{L}_{i[J]}^{-1}\boldsymbol{G}_{i[J]}^T = \boldsymbol{G}_{iJ}(n_{iJ}^{-1}\boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ[J]}\boldsymbol{G}_{i[J]}\boldsymbol{L}_{i[J]}^{-1})\boldsymbol{G}_{i[J]}^T = \boldsymbol{N}_{iJ[J]} \tag{D.10}
$$

$$
\begin{aligned}
\boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J} &= \boldsymbol{G}_{iJ}(n_{iJ}^{-1}\boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ[J]}\boldsymbol{G}_{i[J]}\boldsymbol{L}_{i[J]}^{-1})\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{G}_{i[J]}(\boldsymbol{L}_{i[J]}^{-1}\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]J}\boldsymbol{G}_{iJ}n_{iJ}^{-1})\boldsymbol{G}_{iJ}^T \\
&= \boldsymbol{G}_{iJ}\underbrace{n_{iJ}^{-1}(\boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ[J]}\boldsymbol{G}_{i[J]})(\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}\boldsymbol{G}_{i[J]})^{-1}(\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]J}\boldsymbol{G}_{iJ})n_{iJ}^{-1}}_{\text{scalar, defined as } \varpi_J}\boldsymbol{G}_{iJ}^T \\
&= \varpi_J\boldsymbol{J}_{n_{iJ}} \qquad\qquad (\text{D.11})
\end{aligned}
$$

Recall that $\boldsymbol{N}_{iJ} = \nu_{iJ}\{(1-r_{JJ})\boldsymbol{I}_{n_{iJ}} + r_{JJ}\boldsymbol{J}_{n_{iJ}}\}$ has an exchangeable structure, we also have

$$
\begin{aligned}
&(\boldsymbol{N}_{iJ} - \boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J})^{-1} \\
=&\{\nu_{iJ}(1-r_{JJ})\boldsymbol{I}_{n_{iJ}} + (\nu_{iJ}r_{JJ} - \varpi_J)\boldsymbol{J}_{n_{iJ}}\}^{-1} \\
=&\frac{1}{\nu_{iJ}(1-r_{JJ})}\boldsymbol{I}_{n_{iJ}} - \frac{\nu_{iJ}r_{JJ} - \varpi_J}{\nu_{iJ}(1-r_{JJ})\{\nu_{iJ}(1-r_{JJ}) + n_{iJ}(\nu_{iJ}r_{JJ} - \varpi_J)\}}\boldsymbol{J}_{n_{iJ}} \qquad \text{(D.12)}
\end{aligned}
$$

then

$$
\boldsymbol{G}_{iJ}^T(\boldsymbol{N}_{iJ} - \boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J})^{-1}\boldsymbol{G}_{iJ} = \frac{n_{iJ}}{\nu_{iJ}\{1+(n_{iJ}-1)r_{JJ}\} - n_{iJ}\varpi_J}
$$

The above intermediate results allow us to rewrite

$$
\begin{aligned}
&\left\{\boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ}\boldsymbol{G}_{iJ} - \boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ[J]}\boldsymbol{G}_{i[J]}(\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}\boldsymbol{G}_{i[J]})^{-1}\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]J}\boldsymbol{G}_{iJ}\right\}^{-1} \\
=&\left\{n_{iJ}\nu_{iJ}\{1+(n_{iJ}-1)r_{JJ}\} - n_{iJ}^2\varpi_J\right\}^{-1} \\
=&n_{iJ}^{-2}\boldsymbol{G}_{iJ}^T(\boldsymbol{N}_{iJ} - \boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J})^{-1}\boldsymbol{G}_{iJ} \qquad \text{(D.13)}
\end{aligned}
$$

where the second equality comes from identity (D.11). Therefore, using identity (D.10), we obtain

$$
\begin{aligned}
\boldsymbol{L}_{i[J]}\boldsymbol{\Lambda}_{i[J]}\boldsymbol{G}_{i[J]}^T =&\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}^{-1} + \boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J}\boldsymbol{G}_{iJ}n_{iJ}^{-1}\boldsymbol{G}_{iJ}^T\times \\
&(\boldsymbol{N}_{iJ} - \boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J})^{-1}\boldsymbol{G}_{iJ}n_{iJ}^{-1}\boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1} \\
=&\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}^{-1} + \boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J}(\boldsymbol{N}_{iJ} - \boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J})^{-1}\boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1} \\
=&\boldsymbol{G}_{i[J]}^T\left(\boldsymbol{N}_{i[J]} - \boldsymbol{N}_{i[J]J}\boldsymbol{N}_{iJ}^{-1}\boldsymbol{N}_{iJ[J]}\right)^{-1} \\
=&\boldsymbol{G}_{i[J]}^T\boldsymbol{W}_{i[J]}
\end{aligned}
$$

(ii) Lower Right Block $n_{iJ}\boldsymbol{\Lambda}_{iJ}\boldsymbol{G}_{iJ}^T$

Based on identity (D.13) and (D.12), we can write

$$
\begin{aligned}
n_{iJ}\boldsymbol{\Lambda}_{iJ}\boldsymbol{G}_{iJ}^T =& n_{iJ}\left\{\boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ}\boldsymbol{G}_{iJ} - \boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ[J]}\boldsymbol{G}_{i[J]}(\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}\boldsymbol{G}_{i[J]})^{-1}\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]J}\boldsymbol{G}_{iJ}\right\}^{-1}\boldsymbol{G}_{iJ}^T \\
=& n_{iJ}^{-1}\boldsymbol{G}_{iJ}^T(\boldsymbol{N}_{iJ} - \boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J})^{-1}\boldsymbol{G}_{iJ}\boldsymbol{G}_{iJ}^T \\
=& n_{iJ}^{-1}\boldsymbol{G}_{iJ}^T\left\{\frac{1}{\nu_{iJ}(1-r_{JJ})}\boldsymbol{I}_{n_{iJ}} - \frac{\nu_{iJ}r_{JJ}-\varpi_J}{\nu_{iJ}(1-r_{JJ})\{\nu_{iJ}(1-r_{JJ})+n_{iJ}(\nu_{iJ}r_{JJ}-\varpi_J)\}}\boldsymbol{J}_{n_{iJ}}\right\}\boldsymbol{J}_{n_{iJ}} \\
=& n_{iJ}^{-1}\boldsymbol{G}_{iJ}^T\boldsymbol{J}_{n_{iJ}}\left\{\frac{1}{\nu_{iJ}(1-r_{JJ})} - \frac{n_{iJ}(\nu_{iJ}r_{JJ}-\varpi_J)}{\nu_{iJ}(1-r_{JJ})\{\nu_{iJ}(1-r_{JJ})+n_{iJ}(\nu_{iJ}r_{JJ}-\varpi_J)\}}\right\} \\
=& \boldsymbol{G}_{iJ}^T\left\{\frac{1}{\nu_{iJ}(1-r_{JJ})} - \frac{n_{iJ}(\nu_{iJ}r_{JJ}-\varpi_J)}{\nu_{iJ}(1-r_{JJ})\{\nu_{iJ}(1-r_{JJ})+n_{iJ}(\nu_{iJ}r_{JJ}-\varpi_J)\}}\right\} \\
=& \left\{\frac{1}{\nu_{iJ}(1-r_{JJ})}\boldsymbol{G}_{iJ}^T\boldsymbol{I}_{n_{iJ}} - \frac{(\nu_{iJ}r_{JJ}-\varpi_J)}{\nu_{iJ}(1-r_{JJ})\{\nu_{iJ}(1-r_{JJ})+n_{iJ}(\nu_{iJ}r_{JJ}-\varpi_J)\}}\boldsymbol{G}_{iJ}^T\boldsymbol{J}_{n_{iJ}}\right\} \\
=& \boldsymbol{G}_{iJ}^T(\boldsymbol{N}_{iJ} - \boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J})^{-1} = \boldsymbol{G}_{iJ}^T\boldsymbol{W}_{iJ}
\end{aligned}
$$

(iii) Off-Diagonal Blocks

From the above systems of equations, we have implicitly verified that

$$
\begin{aligned}
&\left\{\boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ}\boldsymbol{G}_{iJ} - \boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ[J]}\boldsymbol{G}_{i[J]}(\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}\boldsymbol{G}_{i[J]})^{-1}\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]J}\boldsymbol{G}_{iJ}\right\}^{-1}\boldsymbol{G}_{iJ}^T \\
=& n_{iJ}^{-1}\boldsymbol{G}_{iJ}^T(\boldsymbol{N}_{iJ} - \boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J})^{-1}. \qquad \text{(D.14)}
\end{aligned}
$$

Hence we write

$$
\begin{aligned}
\boldsymbol{L}_{i[J]}\boldsymbol{\Lambda}_{i[J]J}\boldsymbol{G}_{iJ}^T =& -\boldsymbol{L}_{i[J]}(\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}\boldsymbol{G}_{i[J]})^{-1}\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]J}\boldsymbol{G}_{iJ}\times \\
&\left\{\boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ}\boldsymbol{G}_{iJ} - \boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ[J]}\boldsymbol{G}_{i[J]}(\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}\boldsymbol{G}_{i[J]})^{-1}\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]J}\boldsymbol{G}_{iJ}\right\}^{-1}\boldsymbol{G}_{iJ}^T \\
=& -\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}^{-1}\times\boldsymbol{N}_{i[J]J}\boldsymbol{G}_{iJ}n_{iJ}^{-1}\boldsymbol{G}_{iJ}^T\times(\boldsymbol{N}_{iJ} - \boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J})^{-1} \\
=& -\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J}(\boldsymbol{N}_{iJ} - \boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J})^{-1} \\
=& \boldsymbol{G}_{i[J]}^T\boldsymbol{W}_{i[J]J}
\end{aligned}
$$

where the second equality comes from the induction hypothesis (D.4) and identity (D.14), and the third equality results from identity (D.8).

Finally, we write

$$
\begin{aligned}
n_{iJ}\boldsymbol{\Lambda}_{iJ[J]}\boldsymbol{G}_{i[J]}^T = & - n_{iJ}\left\{\boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ}\boldsymbol{G}_{iJ} - \boldsymbol{G}_{iJ}^T\boldsymbol{N}_{iJ[J]}\boldsymbol{G}_{i[J]}(\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}\boldsymbol{G}_{i[J]})^{-1}\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]J}\boldsymbol{G}_{iJ}\right\}^{-1}\boldsymbol{G}_{iJ}^T\times \\
& \boldsymbol{N}_{iJ[J]}\boldsymbol{G}_{i[J]}(\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}\boldsymbol{G}_{i[J]})^{-1}\boldsymbol{G}_{i[J]}^T \\
= & - \boldsymbol{G}_{iJ}^T(\boldsymbol{N}_{iJ} - \boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J})^{-1}\boldsymbol{N}_{iJ[J]}\boldsymbol{G}_{i[J]}\boldsymbol{L}_{i[J]}^{-1}\boldsymbol{G}_{i[J]}^T\boldsymbol{N}_{i[J]}^{-1} \\
= & - \boldsymbol{G}_{iJ}^T(\boldsymbol{N}_{iJ} - \boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1}\boldsymbol{N}_{i[J]J})^{-1}\boldsymbol{N}_{iJ[J]}\boldsymbol{N}_{i[J]}^{-1} \\
= & \boldsymbol{G}_{iJ}^T\boldsymbol{W}_{iJ[J]}
\end{aligned}
$$

where the second equality is based on (D.14) and the induction hypothesis (D.4), the third equality is from (D.10). The results in steps (i), (ii) and (iii) indicate that $\boldsymbol{L}_i(\boldsymbol{G}_i^T\boldsymbol{M}_{1i}\boldsymbol{G}_i)^{-1}\boldsymbol{G}_i^T = \boldsymbol{G}_i^T\boldsymbol{M}_{1i}^{-1}$, which concludes the induction step. We have thus established the general case of (D.3) for all $J \geqslant 3$.

D.3 *A Corollary*

The above results can be used to show that $\boldsymbol{D}_{1i}^T\boldsymbol{V}_{1i}^{-1}\boldsymbol{D}_{1i} = \boldsymbol{E}_{1i}^T\boldsymbol{M}_{1i}^{-1}\boldsymbol{E}_{1i}$ for all $i = 1,\dots,I$. This is stated in the last sentence of Theorem 3.1. To see why it holds, we can write the left hand side as $\boldsymbol{D}_{1i}^T\boldsymbol{L}_i(\boldsymbol{G}_i^T\boldsymbol{M}_{1i}\boldsymbol{G}_i)^{-1}\boldsymbol{L}_i\boldsymbol{D}_{1i}$, and the right hand side as $\boldsymbol{D}_{1i}^T\boldsymbol{G}_i^T\boldsymbol{M}_{1i}^{-1}\boldsymbol{G}_i\boldsymbol{D}_{1i}$. Multiply both sides of identity (D.3) by $\boldsymbol{G}_i$, we have $\boldsymbol{L}_i(\boldsymbol{G}_i^T\boldsymbol{M}_{1i}\boldsymbol{G}_i)^{-1}\boldsymbol{L}_i = \boldsymbol{G}_i^T\boldsymbol{M}_{1i}^{-1}\boldsymbol{G}_i$ and hence the equality holds.

E. Additional Subgroup Analysis of the Washington State EPT Study

We additionally compare the correlation structures among the adolescent population. Assuming the nested exchangeable correlation model, the between-period correlation is estimated to be almost equal to the within-period correlation, suggesting no apparent correlation decay. In fact, the decay parameter is estimated as $\hat{\rho} = 1.061$ under the exponential decay model. Because this value exceeds the natural bound of ρ, we do not further consider this model in STable 5. Consistent with these observations, the simple exchangeable correlation model provides the smallest CIC_{cp}, and appears appropriate for the adolescent subgroup. On the other hand, the EPT intervention has a minimum effect on the Chlamydia positivity among the adult subgroup (STable 6). Based on the exponential decay model, the odds ratio due to the intervention is 0.99, with the 95% confidence interval straddling one. The estimated CIC_{cp} favors the exponential decay correlation model for the adult subgroup analysis.

F. Sufficient sample sizes for making valid inference on ICC parameters

While unbiased estimation of the ICC parameters could be achieved with MAEE even when the number of clusters is as small as $I = 12$, statistical inference for the ICC parameters is generally challenging in small samples. In our second simulation experiments, we find that, with MAEE and the BC2 variance,

- under a nested exchangeable correlation structure ($\alpha_0 = 0.03$, $\alpha_1 = 0.015$), close to nominal coverage of α_0 requires at least $I = 48$ clusters, and close to nominal coverage of α_1 may require at least $I > 100$ clusters (Figure 2 of the main text);
- under a nested exchangeable correlation structure ($\alpha_0 = 0.1$, $\alpha_1 = 0.05$), close to nominal coverage of α_0 requires at least $I = 36$ clusters, and close to nominal coverage of α_1 may require at least $I > 100$ clusters (SFigure 3 in the Supplementary Material);
- under an exponential decay correlation structure ($\alpha_0 = 0.03$, $\rho = 0.8$), close to nominal coverage of α_0 can require as small as $I = 12$ clusters (because the true ICC is small and close to the boundary 0, the coverage may slightly fluctuate around the nominal level as I becomes larger), and close to nominal coverage of α_1 may require at least $I > 100$ clusters (SFigure 4 in the Supplementary Material);
- under an exponential decay correlation structure ($\alpha_0 = 0.1$, $\rho = 0.5$), close to nominal

coverage of α_0 requires at least $I = 36$ clusters, and close to nominal coverage of α_1 can require at least $I > 100$ clusters (SFigure 5 in the Supplementary Material);.

In general, these findings on sufficient sample sizes for inference on ICCs can be compared with previous simulations by Preisser *and others* (2008) based on individual-level GEE/MAEE. Interestingly, both our simulations and those in Preisser *and others* (2008) suggest that 30 to 40 clusters may be sufficient to provide close to nominal coverage for α_0. However, we find a larger number of clusters may be needed to achieve close to nominal coverage for the between-period correlation (α_1 or ρ), while frequently a smaller number of clusters may be needed to achieve close to nominal coverage for α_1 in Preisser *and others* (2008) (the prior work only focused on the nested exchangeable structure). However, we should note that the different findings in this work and our earlier work can be due to

- *different randomization designs*: our simulation focuses on stepped wedge designs with multiple periods, while Preisser *and others* (2008) studied pre-test post-test parallel designs;
- *equal versus unequal cluster sizes*: we have simulated unequal cluster-period sizes, which would make the sandwich variance estimator more variable and the associated inference more challenging (Kauermann and Carroll, 2001), while Preisser *and others* (2008) simulated equal cluster sizes, which represents a more ideal scenario.

Fortunately, compared to UEE, the use of cluster-period MAEE can substantially mitigate, if not eliminate, the under-coverage of ICC parameters in small samples.

Finally, a reviewer has suggested the simulation results in Maas and Hox (2005) as a comparator on the sufficient sample sizes for inference with these second-order ICC parameters. We notice that Maas and Hox (2005) focused on linear mixed models with continuous outcomes and equal cluster sizes, which may not be as comparable as the simulation study of Preisser *and others* (2008). However, the messages from Maas and Hox (2005) were actually qualitatively similar to our findings. For example, Maas and Hox (2005) indicated that the variance of variance components parameters (second-order parameters just like the ICCs) became more or less accurate when the number of clusters is at least 100, which aligns with our findings for α_1 and ρ. On the other hand, we have already achieved close to nominal coverage for α_0 with fewer than 100 clusters via MAEE, which can be considered as an improvement.

G. Cluster-period Analysis using Generalized Linear Mixed Models

A keen reviewer has raised the issue of performing cluster-period analysis using generalized linear mixed models (GLMMs), and we further elucidate that point here. With a binary outcome, one often assumes the individual-level GLMM takes the following form in SW-CRTs

$$g(\widetilde{\mu}_{ijk}) = \widetilde{\beta}_j + X_{ij}\widetilde{\delta} + \gamma_{ij}, \tag{G.1}$$

where g is the link function, $\widetilde{\mu}_{ijk}$ is the conditional mean (conditional on the latent random effects γ_{ij}), $\widetilde{\beta}_j$ is the conditional period effect, $\widetilde{\delta}$ is the conditional treatment effect, and the latent random effects $(\gamma_{i1}, \ldots, \gamma_{iJ})^T \sim N(0, \sigma_\gamma^2 \widetilde{\boldsymbol{M}})$ for some symmetric Toeplitz correlation matrix $\widetilde{\boldsymbol{M}}$ (see Li *and others* (2020) for more details). With a Gaussian response and identity link function, Li *and others* (2020) reviewed special cases of this model (i.e. resulting in nested exchangeable and exponential decay correlation structures similar to those studied in the main text) in cross-sectional designs. With a binary outcome, g in (G.1) can be more general such as the logit link. Frequently, likelihood-based approach is used to estimate conditional mean and variance components of model (G.1), with the added distributional assumption that $Y_{ijk} \sim \text{Bernoulli}(\widetilde{\mu}_{ijk})$. This distribution assumption in fact sheds light on the appropriate cluster-period analysis, since the cluster-period total $\sum_{k=1}^{n_{ij}} Y_{ijk} \sim \text{Binomial}(n_{ij}, \widetilde{\mu}_{ijk})$, where $\widetilde{\mu}_{ij} = \widetilde{\mu}_{ijk}$ for all k (notice that the right hand side of (G.1) does not depend on k). This suggests that one can obtain the cluster-period total and cluster-period size, and perform a likelihood-based analysis using the Binomial likelihood (instead of the Bernoulli likelihood) to estimate conditional mean parameters and variance components. Due to the likelihood principle, the resulting inference based on Bernoulli likelihood and the Binomial likelihood should be identical.

For binary outcomes, the treatment effect parameters in conditional model (G.1) and population-averaged model (1) in the main text may have different interpretations, depending on the link function g. When g is identity or the log link, one can show that $\widetilde{\delta} = \delta$ due to collapsibility. However, this is general not true for other link functions. For example, when g is the logit link, $\widetilde{\delta}$ tends to be larger than δ (Zeger *and others*, 1988; Preisser *and others*, 2019). As we mentioned in Section 1 of the main text and our earlier work (Preisser *and others*, 2003; Li *and others*, 2018), the treatment effect δ is averaged over all clusters. Even though the treatment indicator changes within a cluster, the marginal model does not conditional on and therefore becomes marginal to the unobserved random effects. In contrast, the meaning of $\widetilde{\delta}$ is conditional on the unobserved random effects γ_{ij} and therefore is cluster-specific, or, strictly speaking, applies to the population with the same value of the unobserved random effects (Preisser *and others*, 2003).

H. Supplementary Tables

STable 1. Examples of simple exchangeable, nested exchangeable and exponential decay correlation structures. The illustration is based on a stepped wedge trial with $J = 3$ periods and $(n_{i1}, n_{i2}, n_{i3}) = (2, 3, 3)$ observations for the three cluster-periods in cluster i. Define $\boldsymbol{Y}_i = (Y_{i11}, Y_{i12}, Y_{i21}, Y_{i22}, Y_{i23}, Y_{i31}, Y_{i32}, Y_{i33})'$.

	$\text{corr}(\boldsymbol{Y}_i)$
Simple exchangeable	$\left[\begin{array}{cc\|ccc\|ccc} 1 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 \\ \alpha_0 & 1 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 \\ \hline \alpha_0 & \alpha_0 & 1 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 \\ \alpha_0 & \alpha_0 & \alpha_0 & 1 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 \\ \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & 1 & \alpha_0 & \alpha_0 & \alpha_0 \\ \hline \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & 1 & \alpha_0 & \alpha_0 \\ \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & 1 & \alpha_0 \\ \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & 1 \end{array}\right]$
Nested exchangeable	$\left[\begin{array}{cc\|ccc\|ccc} 1 & \alpha_0 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 \\ \alpha_0 & 1 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 \\ \hline \alpha_1 & \alpha_1 & 1 & \alpha_0 & \alpha_0 & \alpha_1 & \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 & \alpha_0 & 1 & \alpha_0 & \alpha_1 & \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 & \alpha_0 & \alpha_0 & 1 & \alpha_1 & \alpha_1 & \alpha_1 \\ \hline \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 & 1 & \alpha_0 & \alpha_0 \\ \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_0 & 1 & \alpha_0 \\ \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_0 & \alpha_0 & 1 \end{array}\right]$
Exponential decay	$\left[\begin{array}{cc\|ccc\|ccc} 1 & \alpha_0 & \alpha_0\rho & \alpha_0\rho & \alpha_0\rho & \alpha_0\rho^2 & \alpha_0\rho^2 & \alpha_0\rho^2 \\ \alpha_0 & 1 & \alpha_0\rho & \alpha_0\rho & \alpha_0\rho & \alpha_0\rho^2 & \alpha_0\rho^2 & \alpha_0\rho^2 \\ \hline \alpha_0\rho & \alpha_0\rho & 1 & \alpha_0 & \alpha_0 & \alpha_0\rho & \alpha_0\rho & \alpha_0\rho \\ \alpha_0\rho & \alpha_0\rho & \alpha_0 & 1 & \alpha_0 & \alpha_0\rho & \alpha_0\rho & \alpha_0\rho \\ \alpha_0\rho & \alpha_0\rho & \alpha_0 & \alpha_0 & 1 & \alpha_0\rho & \alpha_0\rho & \alpha_0\rho \\ \hline \alpha_0\rho^2 & \alpha_0\rho^2 & \alpha_0\rho & \alpha_0\rho & \alpha_0\rho & 1 & \alpha_0 & \alpha_0 \\ \alpha_0\rho^2 & \alpha_0\rho^2 & \alpha_0\rho & \alpha_0\rho & \alpha_0\rho & \alpha_0 & 1 & \alpha_0 \\ \alpha_0\rho^2 & \alpha_0\rho^2 & \alpha_0\rho & \alpha_0\rho & \alpha_0\rho & \alpha_0 & \alpha_0 & 1 \end{array}\right]$

STable 2. Coverage of 95% confidence intervals for δ based on the t_{I-2} quantiles as a function of number of clusters I under two different correlation structures: nested exchangeable (NE) and exponential decay (ED). The cluster-period sizes are randomly drawn from DiscreteUniform$(50, 150)$. Results are based on 3000 simulations and empirical coverage between 94.2% and 95.8% is considered as close to nominal, and highlighted in bold font.

		UEE					MAEE				
		MB	BC0	BC1	BC2	BC3	MB	BC0	BC1	BC2	BC3
Correlation structure		$I = 12$									
NE(α_0, α_1)	$(0.03, 0.015)$	**95.2**	92.3	**94.5**	96.3	**94.2**	96.3	92.4	**94.5**	96.3	**94.3**
	$(0.1, 0.05)$	**95.6**	91.7	94.0	**95.8**	93.9	96.8	91.6	94.1	**95.8**	93.8
ED(α_0, ρ)	$(0.03, 0.8)$	**95.4**	92.6	**94.9**	96.2	**94.7**	96.6	92.7	**95.0**	96.3	**94.8**
	$(0.1, 0.5)$	**94.8**	92.4	**94.7**	96.1	**94.6**	95.9	92.4	**94.7**	96.1	**94.6**
Correlation structure		$I = 24$									
NE(α_0, α_1)	$(0.03, 0.015)$	**94.5**	92.9	94.1	**94.8**	93.9	**94.9**	92.9	94.1	**94.8**	93.9
	$(0.1, 0.05)$	**94.8**	93.2	**94.4**	**95.1**	**94.3**	**95.3**	93.2	**94.4**	**95.1**	**94.3**
ED(α_0, ρ)	$(0.03, 0.8)$	**94.9**	93.4	**94.5**	**95.2**	**94.2**	**95.4**	93.5	**94.4**	**95.1**	**94.3**
	$(0.1, 0.5)$	**94.7**	93.1	94.1	**95.1**	94.1	**95.4**	93.1	94.1	**95.0**	94.1
Correlation structure		$I = 36$									
NE(α_0, α_1)	$(0.03, 0.015)$	**95.0**	**94.2**	**94.6**	**95.5**	**94.5**	**95.3**	94.1	**94.7**	**95.5**	**94.6**
	$(0.1, 0.05)$	**94.8**	94.0	**94.5**	**95.1**	**94.5**	**95.3**	94.0	**94.5**	**95.1**	**94.5**
ED(α_0, ρ)	$(0.03, 0.8)$	**95.6**	**94.3**	**95.0**	**95.5**	**94.9**	96.0	**94.2**	**94.9**	**95.6**	**94.9**
	$(0.1, 0.5)$	**94.7**	94.0	**94.6**	**95.1**	**94.5**	**95.1**	93.9	**94.6**	**95.2**	**94.5**

STable 3. Coverage of 95% confidence intervals for correlation parameters based on the t_{I-2} quantiles as a function of number of clusters I under the nested exchangeable (NE) correlation structure. The cluster-period sizes are randomly drawn from DiscreteUniform(50, 150). Results are based on 3000 simulations and empirical coverage between 94.2% and 95.8% is considered as close to nominal, and highlighted in bold font.

		UEE				MAEE			
		BC0	BC1	BC2	BC3	BC0	BC1	BC2	BC3
(α_0, α_1)	I	Coverage of α_0							
$(0.03, 0.015)$	12	80.6	81.8	83.0	81.8	89.3	90.1	91.3	90.1
	24	87.7	88.3	88.7	88.3	91.9	92.5	93.0	92.5
	36	88.5	88.9	89.2	88.9	91.8	92.2	92.4	92.2
$(0.1, 0.05)$	12	83.1	84.8	85.5	84.8	89.6	90.5	91.7	90.5
	24	87.1	87.9	88.4	87.9	92.1	92.4	92.7	92.4
	36	90.0	90.4	90.8	90.4	93.2	93.5	93.5	93.5
(α_0, α_1)	I	Coverage of α_1							
$(0.03, 0.015)$	12	80.7	81.7	82.9	81.8	84.5	85.7	86.6	85.7
	24	86.9	87.3	87.8	87.3	89.0	89.5	90.0	89.5
	36	88.8	89.2	89.5	89.2	90.4	90.6	90.8	90.6
$(0.1, 0.05)$	12	82.5	83.5	84.4	83.5	85.8	86.7	87.7	86.7
	24	87.3	87.9	88.2	87.9	89.3	89.8	90.3	89.8
	36	89.1	89.4	89.8	89.4	91.0	91.2	91.4	91.2

STable 4. Coverage of 95% confidence intervals for correlation parameters based on the t_{I-2} quantiles as a function of number of clusters I under the exponential decay (ED) correlation structure. The cluster-period sizes are randomly drawn from DiscreteUniform$(50, 150)$. Results are based on 3000 simulations and empirical coverage between 94.2% and 95.8% is considered as close to nominal, and highlighted in bold font.

		UEE				MAEE			
		BC0	BC1	BC2	BC3	BC0	BC1	BC2	BC3
(α_0, ρ)	I	Coverage of α_0							
$(0.03, 0.8)$	12	80.5	81.5	82.6	81.5	87.4	88.4	89.1	88.4
	24	86.6	87.1	87.4	87.1	91.0	91.6	92.0	91.6
	36	88.7	89.1	89.4	89.1	92.0	92.4	92.6	92.4
$(0.1, 0.5)$	12	84.3	85.5	86.6	85.5	91.3	92.2	93.2	92.2
	24	88.0	88.7	89.2	88.7	92.3	92.8	93.3	92.7
	36	91.1	91.6	91.9	91.6	**94.3**	**94.6**	**94.7**	**94.6**
(α_0, ρ)	I	Coverage of ρ							
$(0.03, 0.8)$	12	85.6	86.7	87.4	86.6	87.3	88.4	89.3	88.4
	24	90.0	90.5	91.0	90.4	90.8	91.3	92.0	91.3
	36	91.8	92.0	92.3	92.0	92.2	92.6	92.8	92.6
$(0.1, 0.5)$	12	83.4	85.0	86.5	84.9	84.0	85.6	86.6	85.6
	24	87.8	88.8	89.3	88.8	88.5	88.9	89.5	88.9
	36	89.7	90.1	90.5	90.1	89.9	90.4	90.8	90.4

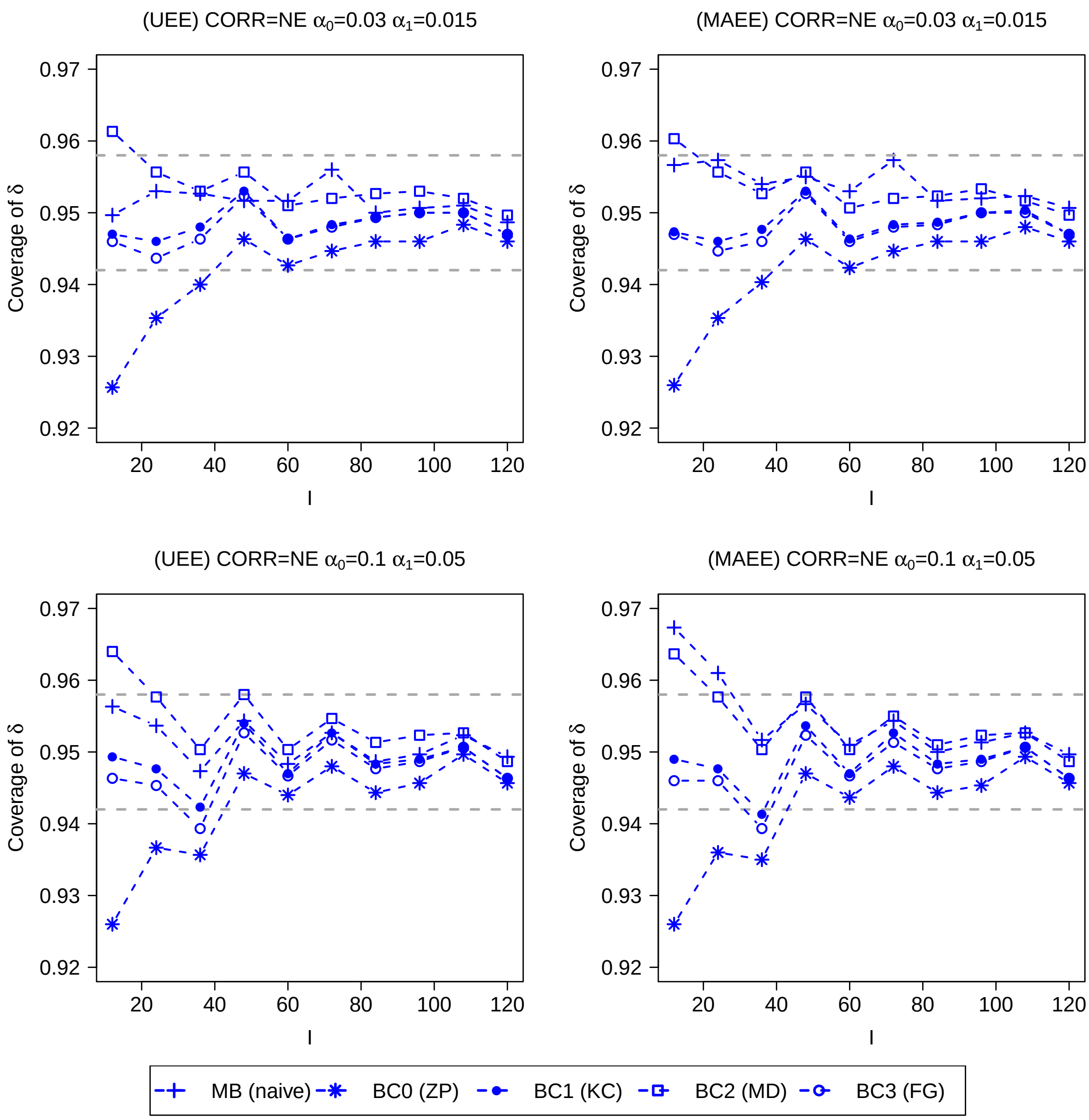
STable 5. Parameter estimates of marginal mean and correlation parameters from the subgroup analysis of Washington State EPT Trial using MAEE. The subgroup includes adolescents aged between 14 and 19. Standard error of the marginal mean parameters are based on BC1 and standard error of the intraclass correlation parameters are based on BC2. All standard error estimates are reported in the parenthesis. Results under the exponential decay structure is not reported because the decay parameter estimate is out of its natural bound.

	Simple exchangeable	Nested exchangeable	Exponential decay
Marginal mean			
β_1 (period 1)	-2.290 (0.098)	-2.290 (0.097)	–
β_2 (period 2)	-2.309 (0.123)	-2.313 (0.128)	–
β_3 (period 3)	-2.264 (0.095)	-2.272 (0.092)	–
β_4 (period 4)	-2.354 (0.160)	-2.360 (0.152)	–
β_5 (period 5)	-2.247 (0.180)	-2.255 (0.188)	–
δ (treatment)	-0.254 (0.106)	-0.249 (0.106)	–
Intraclass correlation			
α_0	.0047 (.0022)	.0041 (.0058)	–
α_1	–	.0049 (.0019)	–
ρ	–	–	–
Correlation selection criteria			
CIC_{cp}	11.69	11.85	–

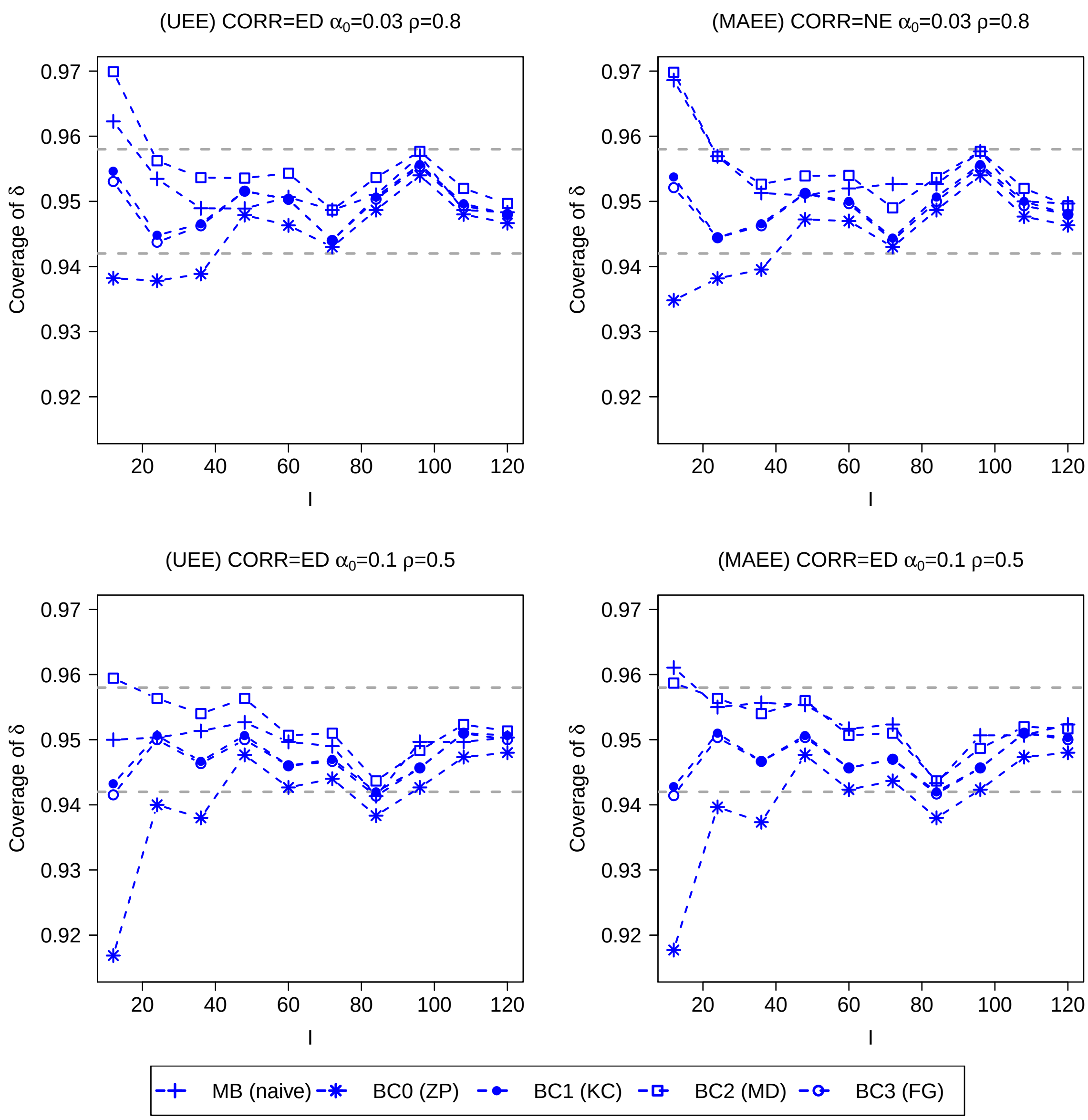
STable 6. Parameter estimates of marginal mean and correlation parameters from the subgroup analysis of Washington State EPT Trial using MAEE. The subgroup includes women aged between 19 and 25. Standard error of the marginal mean parameters are based on BC1 and standard error of the intraclass correlation parameters are based on BC2. All standard error estimates are reported in the parenthesis.

	Simple exchangeable	Nested exchangeable	Exponential decay
Marginal mean			
β_1 (period 1)	-2.573 (0.108)	-2.577 (0.112)	-2.563 (0.108)
β_2 (period 2)	-2.581 (0.078)	-2.556 (0.075)	-2.579 (0.078)
β_3 (period 3)	-2.747 (0.110)	-2.713 (0.113)	-2.747 (0.107)
β_4 (period 4)	-2.807 (0.111)	-2.812 (0.092)	-2.828 (0.101)
β_5 (period 5)	-2.750 (0.139)	-2.782 (0.130)	-2.783 (0.128)
δ (treatment)	-0.033 (0.095)	-0.024 (0.087)	-0.009 (0.089)
Intraclass correlation			
α_0	.0038 (.0029)	.0063 (.0063)	.0043 (.0049)
α_1	–	.0024 (.0019)	–
ρ	–	–	.9082 (.4111)
Correlation selection criteria			
CIC_{cp}	9.94	10.20	9.52

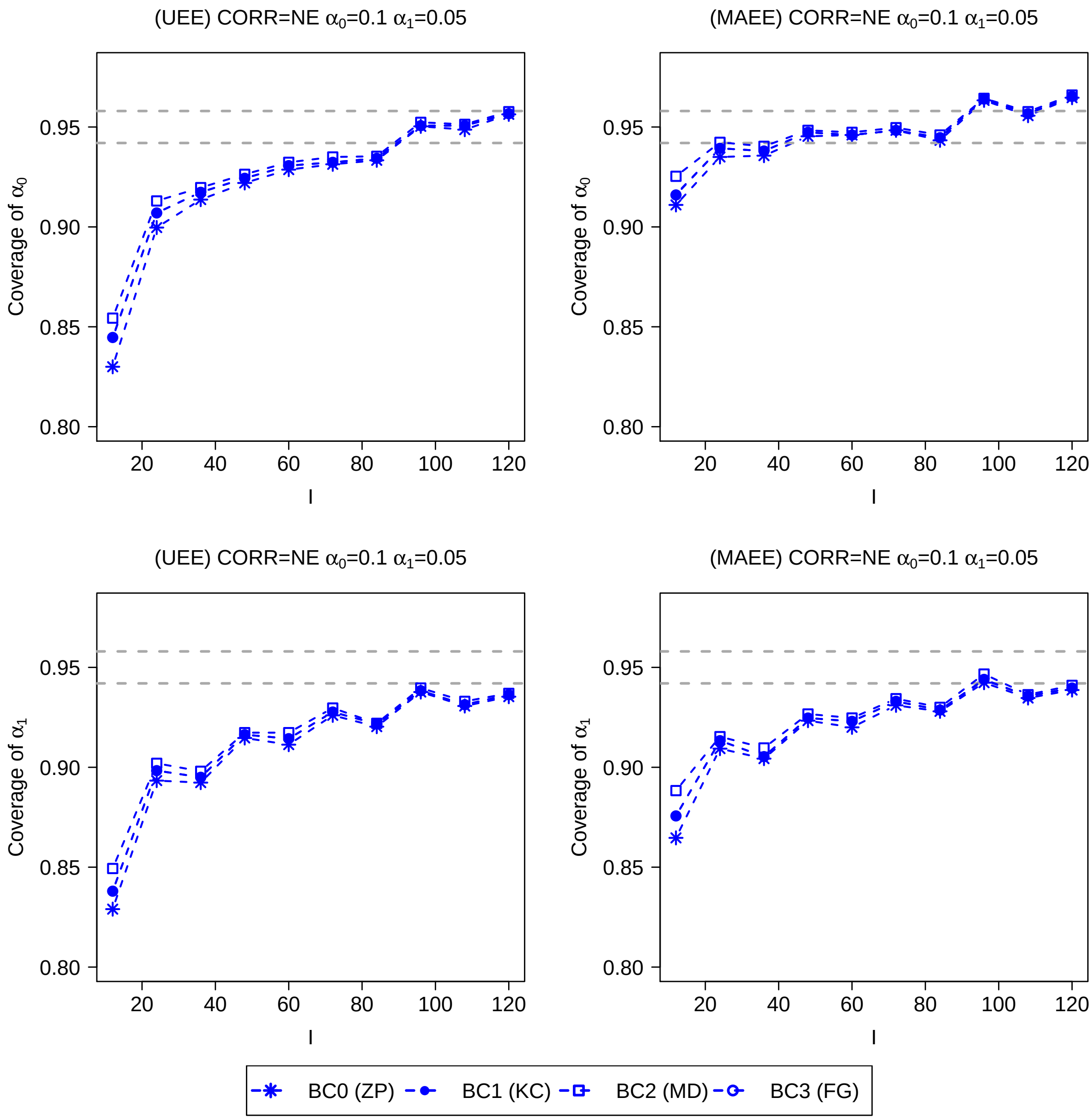
I. Supplementary Figures



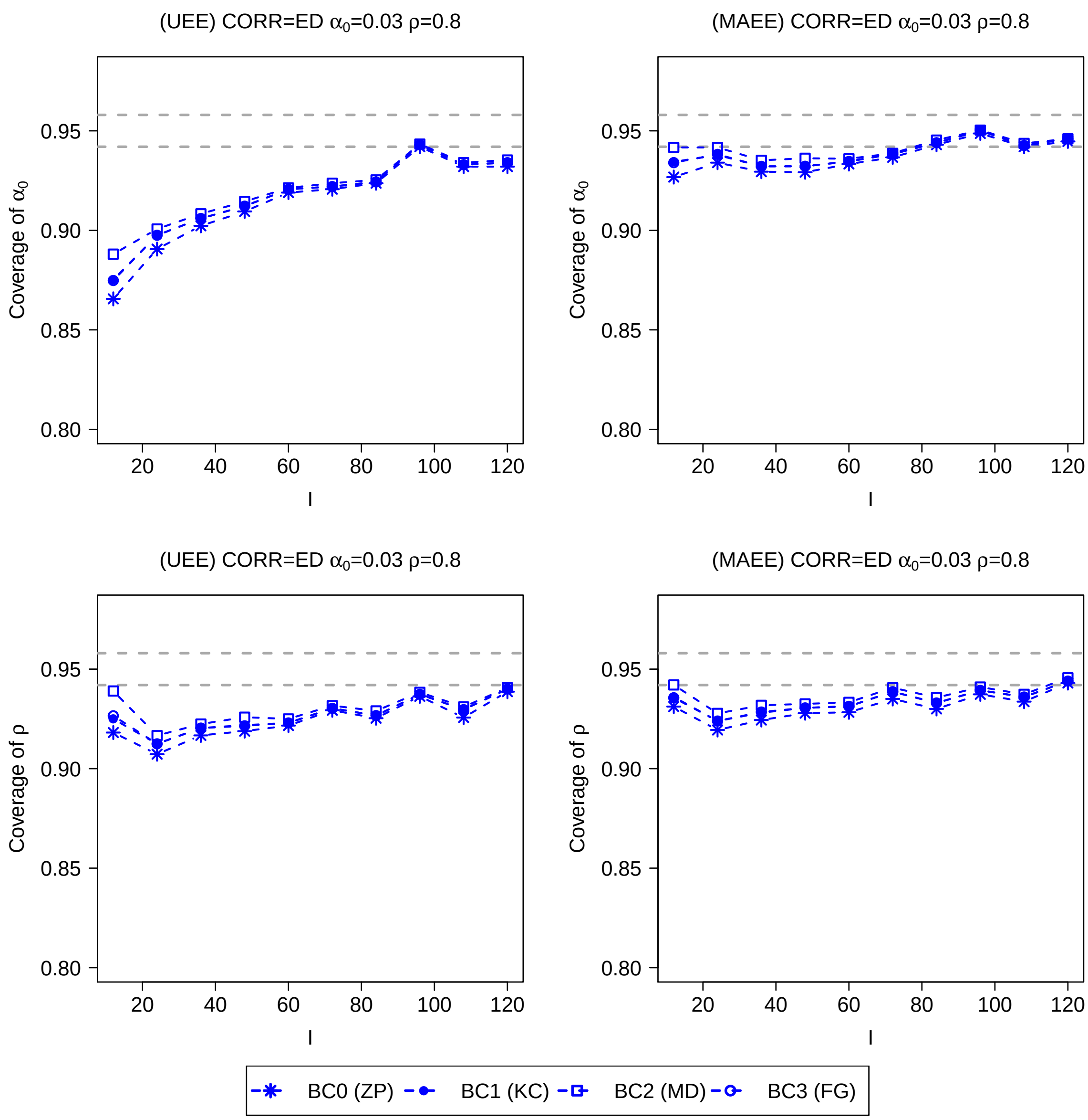
SFigure 1. Coverage of 95% confidence intervals for the intervention effect parameter δ based on the t_{I-2} quantiles as a function of number of clusters I under the nested exchangeable (NE) correlation structure. The cluster-period sizes are randomly drawn from $\mathrm{DiscreteUniform}(25, 50)$. The acceptable bounds according to simulation error based on 3000 replicates are shown with the dashed horizontal lines.



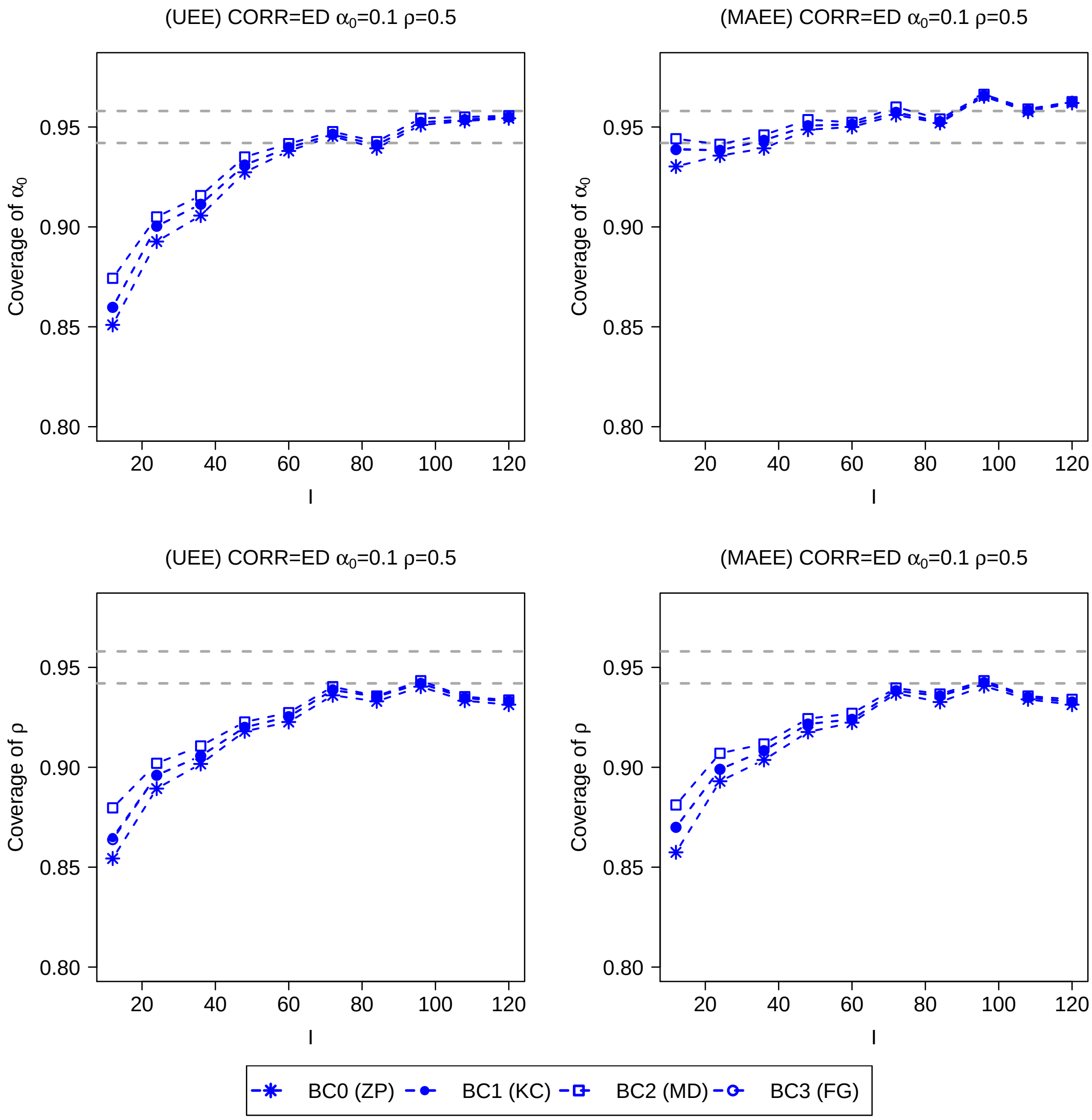
SFigure 2. Coverage of 95% confidence intervals for the intervention effect parameter δ based on the t_{I-2} quantiles as a function of number of clusters I under the exponential decay (ED) correlation structure. The cluster-period sizes are randomly drawn from DiscreteUniform$(25, 50)$. The acceptable bounds according to simulation error based on 3000 replicates are shown with the dashed horizontal lines.



SFigure 3. Coverage of 95% confidence intervals for correlation parameters based on the t_{I-2} quantiles as a function of number of clusters I under the nested exchangeable (NE) correlation structure when $\alpha_0 = 0.1$ and $\alpha_1 = 0.05$. The cluster-period sizes are randomly drawn from DiscreteUniform$(25, 50)$. The acceptable bounds according to simulation error based on 3000 replicates are shown with the dashed horizontal lines.



SFigure 4. Coverage of 95% confidence intervals for correlation parameters based on the t_{I-2} quantiles as a function of number of clusters I under the exponential decay (ED) correlation structure when $\alpha_0 = 0.03$ and $\rho = 0.8$. The cluster-period sizes are randomly drawn from DiscreteUniform$(25, 50)$. The acceptable bounds according to simulation error based on 3000 replicates are shown with the dashed horizontal lines.



SFigure 5. Coverage of 95% confidence intervals for correlation parameters based on the t_{I-2} quantiles as a function of number of clusters I under the exponential decay (ED) correlation structure when $\alpha_0 = 0.1$ and $\rho = 0.5$. The cluster-period sizes are randomly drawn from DiscreteUniform$(25, 50)$. The acceptable bounds according to simulation error based on 3000 replicates are shown with the dashed horizontal lines.

References

Graybill, Franklin A. (1983). *Matrices with Applications in Statistics*. Wadsworth, Belmont, CA: Brooks/Cole.

Henderson, Harold V. and Searle, Shayle R. (1981). On deriving the inverse of a sum of matrices. *Siam Review* **23**, 53–60.

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

Kauermann, G. and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* **96**, 1387–1396.

Li, F., Hughes, J. P., Hemming, K., Taljaard, M., Melnick, E. R. and Heagerty, P. J. (2020). Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: An overview. *Statistical Methods in Medical Research*.

Li, F., Turner, E. L. and Preisser, J. S. (2018). Sample size determination for GEE analyses of stepped wedge cluster randomized trials. *Biometrics* **74**, 1450–1458.

Maas, Cora JM and Hox, Joop J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology* **1**, 86–92.

Preisser, John S, Inan, Gul, Powers, James M and Chu, Haitao. (2019). A population-averaged approach to diagnostic test meta-analysis. *Biometrical Journal* **61**, 126–137.

Preisser, J. S., Lu, B. and Qaqish, B. F. (2008). Finite sample adjustments in estimating equations and covariance estimators for intracluster correlations. *Statistics in Medicine* **27**, 5764–5785.

Preisser, John S., Young, Mary L., Zaccaro, Daniel J. and Wolfson, Mark. (2003). An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Statistics in Medicine* **22**, 1235–1254.

White, Halbert. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society* **50**, 1–25.

Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**, 1049–1060.