HHS Public Access

Author manuscript

Stat Methods Med Res. Author manuscript; available in PMC 2021 March 31.

Published in final edited form as:

Stat Methods Med Res. 2021 February ; 30(2): 612–639. doi:10.1177/0962280220932962.

Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: An overview

Fan Li[1,2], **James P Hughes**[3], **Karla Hemming**[4], **Monica Taljaard**[5], **Edward R. Melnick**[6], **Patrick J Heagerty**[3]

[1]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA

[2]Center for Methods in Implementation and Preventive Science, Yale University, New Haven, CT, USA

[3]Department of Biostatistics, School of Public Health, University of Washington, Seattle, WA, USA

[4]Institute of Applied Health Research, University of Birmingham, Birmingham, UK

[5]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada

[6]Department of Emergency Medicine, Yale School of Medicine, New Haven, CT, USA

Abstract

The stepped wedge cluster randomized design has received increasing attention in pragmatic clinical trials and implementation science research. The key feature of the design is the unidirectional crossover of clusters from the control to intervention conditions on a staggered schedule, which induces confounding of the intervention effect by time. The stepped wedge design first appeared in the Gambia hepatitis study in the 1980s. However, the statistical model used for the design and analysis was not formally introduced until 2007 in an article by Hussey and Hughes. Since then, a variety of mixed-effects model extensions have been proposed for the design and analysis of these trials. In this article, we explore these extensions under a unified perspective. We provide a general model representation and regard various model extensions as alternative ways to characterize the secular trend, intervention effect, as well as sources of heterogeneity. We review the key model ingredients and clarify their implications for the design and analysis. The article serves as an entry point to the evolving statistical literatures on stepped wedge designs.

Keywords

Cluster randomized trials; group-randomized trials; heterogeneity; intraclass correlation coefficient; mixed-effects regression; pragmatic clinical trials; sample size calculation

Corresponding author: Fan Li, Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA. fan.f.li@yale.edu.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

1 Introduction

Cluster-randomized trials (CRTs), also known as group-randomized trials, are frequently designed to evaluate the effect of an intervention administered at the cluster level, such as clinics, hospitals or geographical units.[1–4] Common reasons for randomizing at the cluster level include minimization of treatment contamination, administrative convenience, among others. The design and analysis of CRTs have been an active area of research over the past four decades and comprehensive reviews of recent methodological developments can be found in Turner et al.[5,6] and Murray et al.[7] In parallel designs, usually half of the clusters are randomized to each arm. While parallel randomization ensures valid comparisons of post-treatment outcomes at the same point in time, concurrent implementation of the intervention may demand extensive administrative planning and logistical infrastructure.[8]

The stepped wedge CRT is an alternative design that allows for phased implementation of an intervention. In a stepped wedge CRT, clusters are randomized to intervention sequences that differ by the time points when the intervention starts to roll out.[9,10] An attractive feature of the stepped wedge CRT is that all clusters eventually receive the intervention, which can help facilitate recruitment when cluster stakeholders perceive the intervention to be beneficial.[11–13] Stepped wedge designs have also received increasing attention in recent pragmatic clinical trials (PCTs) embedded in health care delivery systems; see, for example, the Lumbar Imaging with Reporting of Epidemiology study (LIRE)[14] and the Trauma Survivors Outcomes and Support study (TSOS)[15] as two recent Demonstration Projects supported by the U.S. National Institutes of Health (NIH) Health Care Systems Research Collaboratory.[16] Because of their unique features, stepped wedge CRTs usually require more complex statistical considerations compared to parallel CRTs. Mixed-effects regression is one of several approaches proposed for CRTs, and has been the most commonly used approach in analyzing stepped wedge CRTs. The objective of this article is to provide an overview of mixed-effects models developed for stepped wedge CRTs. In an effort to clarify their assumptions and implications, this article provides an entry point to the evolving statistical literatures on stepped wedge CRTs.

Several systematic reviews emphasized the conducting and reporting related to both the design and analysis of stepped wedge CRTs. For example, Martin et al.[17] and Grayling et al.[18] assessed the quality of reporting and design features of stepped wedge CRTs and found that many studies did not adhere to the guidelines recommended in the earlier CONSORT extension to CRTs.[19] In particular, statistical methods for the sample size determination varied across studies, and insufficient details on modeling assumptions were provided. Variations in statistical models had been first observed in an earlier systematic review by Brown and Lilford,[20] even before the standard method was published in Hussey and Hughes.[9] Davey et al.[21] and Barker et al.[22] surveyed the statistical methodology used for stepped wedge CRTs in practice and also noticed substantial variations in model specification, to which the sample size calculation and model-based inference could be sensitive. The various model specifications in practice motivated us to integrate the current toolkit of analytical models for stepped wedge designs.

Hemming et al.[23,24] recently introduced the CONSORT extension for the stepped wedge CRTs and encouraged clear reporting of analytical models specified for sample size calculation (item 7a) as well as for the primary and secondary analyses (items 12a and 12 b). In what follows, we consider the specification of the secular trend, the intervention, and sources of heterogeneity as three essential components of a model, and describe different formulations of each component. Our overview complements the CONSORT extension in clarifying the similarities and differences among models and in facilitating their proper application. The scope of this article differs from previous systematic reviews due to the fact that it is focusing on the statistical formulations and assumptions of the models used to describe the individual-level outcome trajectories. We took a top-down approach by providing a general model representation that separates the three essential components (i.e. secular trend, intervention effect, and sources of heterogeneity). We then cast a number of model variants as special cases of the general representation to explain their assumptions and implications for the design and analysis.

The rest of this article is organized as follows. Section 2 introduces the notation and the general model representation. Section 3 provides an overview of existing mixed-effects models and clarifies their assumptions and properties. Section 4 reviews the estimation and inference strategies in stepped wedge trials, and Section 5 concludes with a discussion.

2 A general model representation

2.1 Notation

Throughout the paper, we consider a stepped wedge CRT with I participating clusters followed over J ($J \geq 3$) time periods. We assume that individuals are included in each cluster and the outcome assessment is scheduled during each period at the individual level; in other words, we only consider *complete* designs,[25] and refer readers to Kasza and Forbes[26] and Kasza et al.[27] for methodological developments on *incomplete* designs. Based on the terminology of Murray and Hannan[28] and Feldman and Mckinlay,[29] we will distinguish between *cross-sectional* and *closed-cohort* stepped wedge designs. In a cross-sectional design, different individuals are observed in each cluster over time, whereas in a closed-cohort design, individuals are identified at the start of the trial and scheduled for repeated outcome assessment. In addition, Copas et al.[30] discussed a third option, the *open-cohort* design, which allows for attrition of members from and addition of new members to the original cohort in each cluster. We will describe the notation for each one of these three designs, and consistently use these notation when discussing model development.

For the cross-sectional design, we assume N_{ij} individuals are included during period j ($j = 1, \ldots, J$) in cluster i ($i = 1, \ldots, I$); the cluster-period sizes may vary. For the closed-cohort design, we define N_i as the cohort size in cluster i as repeated measurements are taken from the same individuals. The open-cohort design can be considered as a mix of a cross-sectional design and a closed-cohort design, and we still assume N_{ij} individuals are included during period j in cluster i. However, in this case, there exists an *overlapping* number ($0 \leq n_i(j, l) \leq \min\{N_{ij}, N_{il}\}$) of individuals for period j and period l in cluster i, depending on the degree of cohort openness. The notation of the open-cohort design generalizes that of the previous two designs, because the cross-sectional design is obtained as a special case with $n_i(j, l) = 0$ for

all j and l (maximum degree of openness) and the closed-cohort design is obtained with $n_i(j, l) = N_{ij} = N_{il}$ for all j and l (minimum degree of openness). Such notation becomes useful in Section 3.6. For all three types of designs, each cluster typically starts out in the control condition; clusters or sets of clusters are then randomized to intervention sequences and all clusters will be exposed to the intervention condition before the end of the trial. Figure 1 provides a schematic illustration of a design with I=8 clusters and J=5 periods. Notably, each one of the four distinct intervention sequences is fully determined by the time period during which the intervention is first implemented. We define the total number of distinct intervention sequences by $S(S \leq J-1)$, and there are in total S=4 pre-planned sequences in Figure 1.

2.2 Outcome model

The analysis of stepped wedge CRTs usually involves the characterization of a cluster-level, time-specific outcome trajectory. Here, we focus on the class of conditional models that require specification of fixed effects for the group-average structure and random effects for the heterogeneity. We will return to a brief discussion of marginal models in Section 5. The conditional models and marginal models have their own advantages and disadvantages, and our experience suggests there are more off-the-shelf software routines to fit conditional models with a complex random-effects structure. The review of Barker et al.[22] also suggested that 61 out of 102 stepped wedge CRTs specified a linear or generalized linear mixed model for the primary analysis.

We define $Y_{ijk}(s)$ as the potential outcome of individual k during period j in cluster i, had cluster i received, possibly to contrary to fact, an intervention sequence s.[31] We borrow the potential outcome framework of Rubin[32] to clearly indicate the dependence of elements on intervention sequences. We index each distinct sequence by s, which is defined as the time interval when the intervention will be first introduced. Formally, $s \in \mathcal{S} \subseteq \{2, \ldots, J\}$, and the total number of sequences is the cardinality, $S = \text{card}(\mathcal{S})$. Define $\mu_{ijk}(s)$ as the expectation of $Y_{ijk}(s)$. We use a generalized linear mixed model to characterize the mean potential outcome as

$$g\left[\mu_{ijk}(s)\right] = \boldsymbol{F}_i(j, s)'\boldsymbol{\theta} + \boldsymbol{R}_{ik}(j, s)'\boldsymbol{\alpha}_i \quad (1)$$

where g is a link function. Similar model representation has been previously introduced by Sitlani et al. in the context of longitudinal observational studies.[33]

On the link function scale, $\boldsymbol{F}_i(j, s)'\boldsymbol{\theta}$ represents the group-average component and vector $\boldsymbol{\theta}$ includes the parameter of interest (i.e. the intervention effect), while $\boldsymbol{R}_{ik}(j, s)'\boldsymbol{\alpha}_i$ represents the cluster-specific, time-specific, and/or individual-specific departure from the group average. By design, the assignment of intervention to clusters is monotone and confounded with time. Hence, it is common practice to separate $\boldsymbol{F}_i(j, s)$ into a baseline component $\boldsymbol{F}^0(j)$ characterizing the background secular trend in the absence of intervention, and a time-dependent intervention component $F_i^1(j, s) = \mathbb{I}_{[j \geq s]}$. Then, the group-average component can be expressed as

$$\boldsymbol{F}_i(j,s)'\boldsymbol{\theta} = \boldsymbol{F}^0(j)'\boldsymbol{\beta} + F_i^1(j,s)\Delta(j,s) \tag{2}$$

where $\boldsymbol{\beta}$ is the parameter encoding the secular trend without intervention, and $\Delta(j, s)$ is the change in the mean outcome at period j due to sequence s. To summarize, the ingredients for a potential mean outcome model are

$$g\left[\mu_{ijk}(s)\right] = \underbrace{\boldsymbol{F}^0(j)'\boldsymbol{\beta}}_{\text{secular trend}} + \underbrace{F_i^1(j,s)\Delta(j,s)}_{\text{intervention effect}} + \underbrace{\boldsymbol{R}_{ik}(j,s)'\boldsymbol{\alpha}_i}_{\text{heterogeneity}} \tag{3}$$

With such a formulation, the potential outcome $Y_{ijk}(s)$ is then assumed to follow a parametric distribution with mean $\mu_{ijk}(s)$ and variance as a function of $\mu_{ijk}(s)$. For example, if the potential outcome is continuous and assumed normally distributed, we use an identity link for g and obtain the linear mixed model

$$Y_{ijk}(s) = \boldsymbol{F}^0(j)'\boldsymbol{\beta} + F_i^1(j,s)\Delta(j,s) + \boldsymbol{R}_{ik}(j,s)'\boldsymbol{\alpha}_i + \epsilon_{ijk} \tag{4}$$

where ϵ_{ijk}'s are independent and identically distributed as $N\left(0, \sigma_\epsilon^2\right)$. Assume that there are no hidden variations of the intervention (i.e. the intervention is well defined), and we can link the observed outcome to the potential outcome by equating $Y_{ijk} = Y_{ijk}(s)$, if cluster i receives sequence s. This allow us to use the observed data to estimate all model parameters. As will be seen in Section 3, another typical assumption of models (3) and (4) is that the heterogeneity parameter α_i is assumed independent across clusters and follows a common parametric distribution. This assumption implies that the potential outcomes are independent across clusters, and would not be affected by the intervention sequences received by other clusters.[31] On the other hand, the heterogeneity parameter α_i can induce correlation between potential outcomes of different individuals in the same cluster. Finally, because the majority of the literature on stepped wedge designs has focused on a continuous outcome, we will start with the identity link function and review existing models as special cases of the general representation (4).

3 Modeling considerations and implications

3.1 The Hussey and Hughes model

The standard analytical model for stepped wedge designs was proposed in the seminal paper by Hussey and Hughes.[9] Assuming an identity link function g, the observed outcome Y_{ijk} is modeled as

$$Y_{ijk} = \mu + \beta_j + \delta X_{ij} + \alpha_i + \epsilon_{ijk} \tag{5}$$

where μ is the grand mean, β_j is the jth period effect (with $\beta_1 = 0$ for identifiability), X_{ij} is a time-varying intervention indicator for cluster i during period j (X_{ij}=1, if exposed to intervention; and X_{ij}=0, otherwise), δ is the intervention effect, α_i is the random cluster effect assumed to follow $N\left(0, \tau_\alpha^2\right)$, and $\epsilon_{ijk} \sim N\left(0, \sigma_\epsilon^2\right)$ is the residual error independent of α_i.

The Hussey and Hughes model is a special case of model (4). We observe that the secular trend is assumed as

$$\boldsymbol{F}^{0}(j)'\boldsymbol{\beta} = \mu + \beta_2 \mathbb{I}[j = 2] + \cdots + \beta_J \mathbb{I}[j = J]$$

where $\mathbb{I}[\,\cdot\,]$ is an indicator function. Because the average secular trend is assumed to be a distinct value during each period, this representation requires J parameters and is considered saturated. Further, the intervention effect, $\Delta(j, s) = \delta$, does not depend on the time interval during which the intervention was initiated. Finally, inherited from the models for analyzing parallel CRTs, the heterogeneity term, $\boldsymbol{R}_{ik}(j, s)'\boldsymbol{a}_i = a_i$, captures the cluster-specific departure from the average but is assumed to be homogeneous across time periods, intervention sequences, and individuals.

The assumptions of the Hussey and Hughes model may be considered restrictive. For example, the intervention effect $\Delta(j, s)$ could be cumulative and explicitly depend on the time when the intervention was initiated, which is not captured by a constant intervention effect. In addition, the single cluster random effect postulates a *simple exchangeable* within-cluster correlation structure. In other words, the correlation between any pair of observations k, m in any two periods j, l and across all sequences s is assumed to be a nonnegative constant

$$\operatorname{corr}\left[Y_{ijk}(s), Y_{ilm}(s)\right] = \rho = \tau_\alpha^2 / \left(\tau_\alpha^2 + \sigma_\epsilon^2\right) \tag{6}$$

where corr$[x, y]$ is a symmetric correlation operator. The value of ρ is referred to as the intraclass correlation coefficient (ICC).[1] Such a simple correlation structure also does not account for repeated measurements from the same individual, and so applies only to the cross-sectional setting.

The Hussey and Hughes model has been frequently used to estimate the required sample size for cross-sectional stepped wedge CRTs. Assuming equal cluster-period sizes $N_{ij} = N$ and known variance components, Hussey and Hughes derived the variance of the intervention effect estimator.[9] Let $\lambda_1 = 1 - \rho$ and $\lambda_2 = 1 + (JN - 1)\rho$ we can re-write the variance of the intervention effect estimator as

$$\operatorname{var}\left(\hat{\delta}\right) = \frac{\left(\sigma_{\text{tot}}^2 / N\right) IJ\lambda_1\lambda_2}{\left(U^2 + IJU - JW - IV\right)\lambda_2 - \left(U^2 - IV\right)\lambda_1} \tag{7}$$

where $\sigma_{\text{tot}}^2 = \tau_\alpha^2 + \sigma_\epsilon^2$ is the total variance, $U = \sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}$, $W = \sum_{j=1}^{J}\left(\sum_{i=1}^{I} X_{ij}\right)^2$ and $V = \sum_{i=1}^{I}\left(\sum_{j=1}^{J} X_{ij}\right)^2$ are design constants that depend on the assignment of intervention sequences to clusters. Using the results in Li et al.,[34] we can show that λ_1 and λ_2 are two distinct eigenvalues of the simple exchangeable correlation matrix. In fact, we will see in due course that expression (7) is a general form that applies to several other model variants, with slight changes in values for the total variance and the eigenvalues. The Hussey and Hughes variance formula is the basis for a number of subsequent methodological

investigations. For instance, Woertman et al.[35] used the variance to derive a design effect, or variance inflation factor, relative to the individually randomized trial, under a *balanced* allocation of clusters to intervention sequences. The variance formula or design effect also motivated the study of optimal stepped wedge designs in the cross-sectional setting; see, for example, Lawrie et al.[36] and Thompson et al.[37] Girling and Hemming[38] considered optimal designs within a larger design space that includes hybrid designs (i.e. designs having both parallel and stepped wedge components), and found that the most efficient design was a hybrid design. Grayling et al.[39] proposed a group sequential design for stepped wedge CRTs. Rhoda et al.[40] and Hemming and Girling[41] studied the relative efficiency between stepped wedge and parallel designs, and found that the relative efficiency depends on the number of periods J, cluster-period sizes N and the intraclass correlation coefficient ρ. The impact of variable cluster sizes, based on the Hussey and Hughes model, was studied in Kristunas et al. and Martin et al.[42,43] Even though the reduction in efficiency due to unequal cluster sizes can be dramatic in a given randomization scheme,[43] the *average* reduction in efficiency is generally smaller in a stepped wedge CRT compared to that in a parallel CRT. Harrison et al.[44] further developed an optimization algorithm for power calculation that accounts for unequal cluster sizes.

Taljaard et al.[45] and Bond[46] pointed out a possible limitation of the Hussey and Hughes model from a variance perspective. Specifically, one can show that the variance of the intervention effect estimator, $\mathrm{var}(\hat{\delta})$, converges to zero if the cluster-period sizes $N \to \infty$. This implies that the required number of clusters for an anticipated power converges to unity as N increases indefinitely, which may not be realistic. Nevertheless, both Barker et al.[22] and Martin et al.[17] found in their systematic reviews that the Hussey and Hughes model was the most widely used approach for designing and analyzing stepped wedge CRTs.

In what follows, we will review extensions of the Hussey and Hughes model, with an emphasis on alternative considerations on modeling $\boldsymbol{F}^0(j)'\beta$, $\Delta(j, s)$ and $\boldsymbol{R}_{ik}(j, s)'\boldsymbol{\alpha}_i$. The considerations for modeling the group-average component (i.e. secular trend and intervention effect) are typically the same between cross-sectional, closed-cohort and open-cohort designs; therefore, we will not consider them separately. However, the considerations for modeling heterogeneity can differ between designs, and will be separately discussed in Sections 3.4, 3.5 and 3.6.

3.2 Considerations for modeling the secular trend

Because the intervention is confounded with time, modeling the background secular trend is necessary to remove the bias in estimating the effect attributed solely to the intervention.[9,47] Recall that $\boldsymbol{F}^0(j)'\boldsymbol{\beta}$ models the group-average secular trend in the absence of intervention across J time periods, and one may generally write

$$\boldsymbol{F}^0(j)'\boldsymbol{\beta} = \beta_1 B_1(j) + \cdots + \beta_p B_p(j) \tag{8}$$

where $\boldsymbol{F}^0(j) = (B_1(j), \ldots, B_p(j))'$ is a p-dimensional basis function, and p is generally no larger than J for identifiability. Different choices of the basis function results in different

formulations of the average secular trend. For instance, the Hussey and Hughes model assumes a saturated J-dimensional basis function with

$$\boldsymbol{F}^0(j) = \left(1, \mathbb{I}[j = 2] \cdots, \mathbb{I}[j = J]\right)'$$

while Hemming et al.[47] explored a linear trend specification such that $p = 2$ and $\boldsymbol{F}^0(j) = (1, j)'$. In principle, as long as $p \leq J$ is required for identifiability, one could expand on the linear trend specification by including higher-order polynomial terms or their orthogonal counterparts.[48] In a recent simulation study, Nickless et al.[49] examined the quadratic specification with $\boldsymbol{F}^0(j) = (1, j, j^2)'$ and found that such models performed generally well in terms of bias when the approximation to the true secular trend was adequate, even if the data were generated from complex nonlinear time effects.

From a bias perspective, it is natural to consider a nonparametric representation of $\boldsymbol{F}^0(j)'\boldsymbol{\beta}$, which would favor the saturated specification as in the Hussey and Hughes model. For example, when the true secular trend is nonlinear, the saturated specification could adequately control for the time effect, while the linear trend specification may lead to a biased intervention effect estimate. While the saturated time parameterization is adequate for trials with a limited number of discrete periods ($J = 5$) such as in the Washington State EPT Study,[50] it may not be the most efficient if there are a large number of periods relative to the number of clusters, due to the reduced degree of freedom available for estimating the intervention effect. For example, Hemming et al.[47] analyzed a stepped wedge CRT of 10 midwifery teams (with each team forming a cluster) to evaluate the effectiveness of a training package to promote sweeping membranes in post-term women in the UK. The trial collected outcomes from each team during each of the 40 weeks of the study, and would have required 39 categorical time parameters if the Hussey and Hughes model had been considered. In general, including many fixed-effects parameters with a limited number of clusters may decrease the precision of the intervention effect, so that it becomes much less likely to locate a true effect signal.[47,51] In this particular case, it seems attractive to look at a parsimonious specification of $\boldsymbol{F}^0(j)'\boldsymbol{\beta}$, such as the linear trend or a polynomial specification to a fixed degree.

Grantham et al.[52] provided an interesting result on time parameterization in stepped wedge CRTs from a variance perspective. In the planning stage, sample size and power calculation critically depend on the variance of the intervention effect, $\text{var}\left(\hat{\delta}\right)$. In the linear mixed model setting with equal cluster-period sizes $N_{ij} = N$, Grantham et al.[52] showed that $\text{var}\left(\hat{\delta}\right)$ was invariant to time parameterization as long as the sum of the intervention sequences across clusters

$$\left(\sum_{i=1}^{I} X_{i1}, \ldots, \sum_{i=1}^{I} X_{iJ}\right)'$$

lay in the column space of $\boldsymbol{F}^0$, where $\boldsymbol{F}^0 = (\boldsymbol{F}^0(1), \ldots, \boldsymbol{F}^0(J))'$ is the design matrix for the secular trend. An implication from this result is that, if there is a *balanced* allocation of

clusters to each sequence ranging from $(0, \ldots, 0, 1)'$ to $(0, 1, \ldots, 1)'$, the saturated time specification in the Hussey and Hughes model and the linear trend specification yield the same expression for $\text{var}(\hat{\delta})$. Further, $\text{var}(\hat{\delta})$ does not change with polynomial specifications as long as the linear time term is included. This invariance property suggests that, with the same trial configuration, the sample size estimates become identical irrespective of the above two time parameterizations.[52] However, it is important to realize that such variance comparisons assume known variance components, and are relevant only for design purposes. In the analysis stage, under-specification of the secular trend could result in bias relative to the true intervention effect, and thus variance appears to be a secondary consideration.[49]

For sample size and power calculation, Heo et al.[53] used a linear mixed model that forwent any secular trend, namely assuming $\boldsymbol{F}^0(j)'\beta = 0$. While Zhou et al.[54] argued that ignoring the secular trend might be reasonable in trials with a very short duration, a number of authors[9,47] have cautioned against the general application of models without a secular trend due to the potential of bias. In fact, one can show analytically that, holding all other conditions equal, the variance of the intervention effect estimator, $\text{var}(\hat{\delta})$, becomes strictly smaller when the secular trend is omitted.[54] This implies that the required sample size could be underestimated when it is incorrectly assumed that there is no time effect.

3.3 Considerations for modeling the intervention effect

In the general model formulation (4), the intervention effect, $\Delta(j, s)$, depends on both period index j and sequence index s, which suggests the possibility for going beyond a constant treatment effect. Formal extensions on modeling a time-varying intervention effect appeared in Hussey and Hughes[9] and Hughes et al.[55] From Hughes et al.,[55] a saturated but stationary intervention effect representation is given by the *general time-on-treatment effect*, where

$$\Delta(j, s) = \delta_{j-s} = \delta_0 \mathbb{I}[j = s] + \delta_1 \mathbb{I}[j = s + 1] + \cdots + \delta_{J-s} \mathbb{I}[j = J] \tag{9}$$

We call this representation stationary because $\Delta(j, s)$ is not a saturated function of (j, s) but a saturated function of $j - s$ for $j \geq s$. The general time-on-treatment effect allows the group-average intervention effect to be different depending on the elapsed number of time intervals since the intervention was first introduced. For example, the model assumes that the intervention effect at time $j \geq s$ is δ_{j-s}, if the intervention is introduced at time s. In this case, the global test for H_0: $\delta_0 = \delta_1 = \cdots = \delta_{J-2} = 0$ is used to assess the overall intervention effect. Nickless et al.[49] reported that a linear mixed model with the time-on-treatment effect assumption had minimum bias and carried close-to-nominal coverage in estimating the average intervention effect under a wide range of scenarios. Further, because the constant intervention effect representation is nested within equation (9), a global test for H_0: $\delta_0 = \delta_1 = \cdots = \delta_{J-2}$ provides a mechanism to assess the plausibility of constant intervention effect assumption.

The general time-on-treatment model requires $J - 2$ parameters for the intervention effect (as compared to only 1 parameter in the Hussey and Hughes model (5)), and could be challenging to estimate in trials with a limited number of clusters. Parsimonious versions of the time-on-treatment effect model have been suggested. Assuming that the periods are

equally spaced, Hughes et al.[55] introduced the *linear time-on-treatment effect* representation, where

$$\Delta(j, s) = \delta_0 + \delta_1(j - s) \tag{10}$$

or more simply, $\Delta(j, s) = \delta(j - s + 1)$, which was assumed as a linear function of the elapsed number of periods since the intervention was first introduced. Such parameterizations are especially useful when the intervention takes more than a single time period to fully develop, or when there is a strengthening or weakening of intervention effect over time. Alternatively, representation (10) can be considered as a constant treatment effect plus a treatment-by-linear-time interaction.

In the presence of a *delayed treatment effect*, one could also incorporate prior knowledge to such delay and model

$$\Delta(j, s) = \delta\pi_0 \mathbb{I}_{[j = s]} + \delta \mathbb{I}_{[j > s]} \tag{11}$$

where $\pi_0 \in [0, 1]$ is a constant value representing how effective the intervention will be during the time interval when it is just introduced.[55] For example, if the intervention is known to be 50% effective when it is first introduced and 100% effective afterwards, we can set $\pi_0 = 1/2$. Had one known from prior knowledge that the intervention will be $100\pi_{j-s}$ percent effective when it has been introduced $j - s$ periods (with $\pi_{j-s} = 0$ if $j < s$), the *general delayed treatment effect* representation can be formalized as

$$\Delta(j, s) = \delta\pi_0 \mathbb{I}_{[j = s]} + \delta\pi_1 \mathbb{I}_{[j = s + 1]} + \cdots + \delta\pi_{J - s} \mathbb{I}_{[j = J]} \tag{12}$$

Clearly, when prior knowledge suggests an arithmetic increase in effectiveness such that $\pi_1 - \pi_2 = \pi_2 - \pi_3 = \ldots = \pi_{j-s-1} - \pi_{j-s}$, representation (12) is an equivalent parameterization to equation (10).

Finally, et al.[55] provided an example of a nonlinear model for the time-on-treatment effect, where the intervention effect is considered to increase nonlinearly over time until it reaches the maximum long-term effect. In that model, the time indicator j indexes the exponential rate of increase, and so the model is no longer nested within equation (9). To facilitate the understanding of various intervention effect assumptions, we provide schematic illustrations of four typical examples in Figure 2.

3.4 Considerations for modeling heterogeneity in cross-sectional designs

There have been extensive discussions of alternative strategies for modeling the random-effects structure in stepped wedge trials, especially for those involving cross-sectional designs. Because such discussion has been centered on extensions to the Hussey and Hughes model, they have almost exclusively adopted the constant intervention effect and the categorical time parameterization. We conjecture that assuming no treatment-by-time interaction in the analytical model has gained popularity since trial planning and sample size estimation are more convenient once a scalar target parameter is assumed. To focus on ideas and stay consistent with the current literature, we will review variants of random-effects structures by assuming a linear link, categorical secular trend (except for the random

coefficient model which uses a linear trend specification) as well as a time-invariant intervention effect. To provide a quick reference, we also list selected model variants in Table 1.

3.4.1 Nested exchangeable correlation model—The first notable extension to the Hussey and Hughes model was found in Hooper et al.[56] and Girling and Hemming.[38] This model has also been referred to as the Hooper/Girling model,[57] and is written as

$$Y_{ijk} = \mu + \beta_j + \delta X_{ij} + \alpha_i + \gamma_{ij} + \epsilon_{ijk} \tag{13}$$

Compared to the Hussey and Hughes model, there is an additional term, $\gamma_{ij} \sim N\left(0, \tau_\gamma^2\right)$, representing the random cluster-by-time interaction. This additional random effect is assumed independent of the random cluster effect α_i. As a special case of the general model representation, the nested exchangeable correlation model specifies the heterogeneity term as

$$\boldsymbol{R}_{ik}(j,s)'\boldsymbol{\alpha}_i = \alpha_i + \gamma_{ij} \tag{14}$$

and therefore allows the deviation from the group average to be both cluster-specific and period-specific. Notice that similar ideas on random cluster-by-time interaction date back to the earlier work of Murray et al.[58] for parallel CRTs with repeated measurements. Hemming et al.[47] pointed out that it might be convenient to consider γ_{ij} as a latent factor arising from the unmeasured time-varying characteristics within a cluster.

The nested exchangeable correlation model distinguishes between two different types of correlation parameters: the within-period ICC and the between-period ICC. Specifically

$$\operatorname{corr}\left[Y_{ijk}(s), Y_{ilm}(s)\right] = \begin{cases} \rho_{\mathrm{w}} = \left(\tau_\alpha^2 + \tau_\gamma^2\right)/\left(\tau_\alpha^2 + \tau_\gamma^2 + \sigma_\epsilon^2\right), & j = l \\ \rho_{\mathrm{b}} = \tau_\alpha^2/\left(\tau_\alpha^2 + \tau_\gamma^2 + \sigma_\epsilon^2\right), & j \neq l \end{cases}$$

where the within-period ICC, ρ_{w}, describes the correlation between two within-cluster observations collected during the same period, and the between-period ICC, ρ_{b}, describes the correlation between two within-cluster observations collected in different periods. Since the variance components are positive, the between-period ICC is constrained to be no larger than the within-period ICC. Such a *nested exchangeable* correlation model has also been previously studied in three-level and crossover CRTs.[59,60] An example matrix form of the nested exchangeable correlation structure is provided in Table 2.

On the other hand, Hooper et al.[56] characterized the nested exchangeable correlation structure based on ρ_{w} and the cluster autocorrelation (CAC), which was defined as

$$\mathrm{CAC} = \tau_\alpha^2/\left(\tau_\alpha^2 + \tau_\gamma^2\right) = \rho_{\mathrm{b}}/\rho_{\mathrm{w}} \tag{15}$$

Different from the individual-level correlation ρ_{b}, the CAC has been interpreted as the correlation between two population means from the same cluster at different times (also see Feldman and Mckinlay[29] for this interpretation). Here we clarify that CAC should actually

be interpreted as the limit of the correlation between two cluster-period means. Specifically, if we define the cluster-period mean as $\bar{Y}_{ij+} = N_{ij}^{-1}\sum_{k=1}^{N_{ij}} Y_{ijk}$, then the variance, covariance and correlation of cluster-period means can be calculated as

$$\mathrm{var}\left(\bar{Y}_{ij+}\right) = \sigma_{\mathrm{tot}}^2\left\{\frac{1+\left(N_{ij}-1\right)\rho_{\mathrm{w}}}{N_{ij}}\right\}, \quad \mathrm{cov}\left(\bar{Y}_{ij+}, \bar{Y}_{il+}\right) = \sigma_{\mathrm{tot}}^2\rho_{\mathrm{b}}$$

$$\begin{aligned}\mathrm{corr}\left(\bar{Y}_{ij+}, \bar{Y}_{il+}\right) &= \frac{\mathrm{cov}\left(\bar{Y}_{ij+}, \bar{Y}_{il+}\right)}{\sqrt{\mathrm{var}\left(\bar{Y}_{ij+}\right)}\sqrt{\mathrm{var}\left(\bar{Y}_{il+}\right)}} \\ &= \frac{N_{ij}\rho_{\mathrm{b}}}{1+\left(N_{ij}-1\right)\rho_{\mathrm{w}}} \rightarrow \mathrm{CAC}, \text{ as } N_{ij} \rightarrow \infty\end{aligned}$$

and CAC is the limit of correlation between $\bar{Y}_{ij+}$ and $\bar{Y}_{il+}$ when the cluster-period size N_{ij} increases indefinitely. Girling and Hemming[38] also defined the cluster mean correlation (CMC) as the proportion of the variance of a cluster mean $\bar{Y}_{i++} = \sum_{j=1}^{J}\sum_{k=1}^{N_{ij}} Y_{ijk}$ that came from random effects that were independent of time. Assuming equal cluster-period sizes $N_{ij} = N$, the CMC is the proportion of variability of $\bar{Y}_{i++}$ explained by α_i, and can actually be rewritten as

$$\begin{aligned}\mathrm{CMC} &= \frac{NJ\rho_{\mathrm{b}}}{1+(N-1)\rho_{\mathrm{w}}+N(J-1)\rho_{\mathrm{b}}} \\ &= \frac{NJ\times\mathrm{CAC}}{1/\rho_{\mathrm{w}}+(N-1)+N(J-1)\mathrm{CAC}}\end{aligned} \tag{16}$$

which is a function of CAC, within-period ICC, number of periods and the cluster-period size. In what follows, we will use the individual-level ICCs to characterize different correlation structures, but the CAC and CMC are two alternative parameterizations.

Hooper et al.[56] and Girling and Hemming[38] provided a closed-form expression for the variance of the intervention effect based on model (13). Assuming equal cluster-period sizes $N_{ij} = N$, one can use the results in Li et al.[34] to show that the form of the variance formula is identical to equation (7), except that we replace

$$\sigma_{\mathrm{tot}}^2 = \tau_\alpha^2 + \tau_\gamma^2 + \sigma_\epsilon^2$$

$$\lambda_1 = 1+(N-1)\rho_{\mathrm{w}} - N\rho_{\mathrm{b}}$$

$$\lambda_2 = 1+(N-1)\rho_{\mathrm{w}} + N(J-1)\rho_{\mathrm{b}}$$

In particular, the two parameters, λ_1 and λ_2, have been shown to be two distinct eigenvalues of the nested exchangeable correlation matrix.[34] Interestingly, the cluster mean correlation

(16) also depends only on the two eigenvalues as we can show CMC = 1 λ_1/λ_2. Further, unlike the Hussey and Hughes model, the nested exchangeable correlation model is considered to be more realistic for cross-sectional studies since the limit of the variance

$$\lim_{N\to\infty} \mathrm{var}\left(\hat{\delta}\right) = \frac{\sigma_{\mathrm{tot}}^2 I(\rho_{\mathrm{w}}-\rho_{\mathrm{b}})\{\rho_{\mathrm{w}}+(J-1)\rho_{\mathrm{b}}\}}{(IU-W)\{\rho_{\mathrm{w}}+(J-1)\rho_{\mathrm{b}}\}+\left(U^2-IV\right)\rho_{\mathrm{b}}} > 0 \tag{17}$$

is a positive quantity as long as $\rho_{\mathrm{w}} \neq \rho_{\mathrm{b}}$.

3.4.2 Exponential decay model—Kasza et al.[57] extended the nested exchangeable correlation model (13) by allowing the between-period correlation to decay exponentially over time. The model is written as

$$Y_{ijk} = \mu + \beta_j + \delta X_{ij} + \gamma_{ij} + \epsilon_{ijk} \tag{18}$$

where the heterogeneity term is

$$\mathrm{R}_{ik}(j,s)'\boldsymbol{\alpha}_i = \gamma_{ij}$$

The collection of random effects in cluster i is assumed to follow $\boldsymbol{\gamma}_i = (\gamma_{i1},\ldots,\gamma_{iJ})' \sim N\left(0, \tau_\gamma^2 \widetilde{\mathbf{M}}\right)$, and $\widetilde{\mathbf{M}}$ had a symmetric Toeplitz structure

$$\widetilde{\mathbf{M}} = \begin{pmatrix} 1 & r_{12} & r_{13} & \ldots & r_{1J} \\ r_{21} & 1 & r_{23} & \ldots & r_{2J} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{J1} & r_{J2} & r_{J3} & \ldots & 1 \end{pmatrix} \tag{19}$$

where $r_{jl} = r_{lj}$ for all l and j.

Clearly, an unrestricted Toeplitz correlation structure could include up to $J(J-1)/2$ unknown parameters, which may not be easy to interpret from a design perspective. Therefore, Kasza et al.[57] focused on the following autoregressive structure for trial planning. Specifically, the structure matrix $\widetilde{\mathbf{M}}$ could include two parameters r_0 and r and is written as

$$\widetilde{\mathbf{M}} = \mathbf{M}(r_0, r) = \begin{pmatrix} 1 & r_0 r & r_0 r^2 & \ldots & r_0 r^{J-1} \\ r_0 r & 1 & r_0 r & \ldots & r_0 r^{J-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_0 r^{J-1} & r_0 r^{J-2} & r_0 r^{J-3} & \ldots & 1 \end{pmatrix} \tag{20}$$

The Hussey and Hughes model and the nested exchangeable correlation model are returned by $\mathbf{M}(1, 1)$ and $\mathbf{M}(r_0, 1)$, while the exponential decay model of Kasza et al.[57] is returned by $\mathbf{M}(1, r)$. Although the Hussey and Hughes model is a special case of the nested exchangeable correlation model, it is important to realize that the exponential decay and

nested exchangeable correlation models do not have a clear nesting relationship. The exponential decay model implies the following correlation structure

$$\operatorname{corr}\left[Y_{ijk}(s), Y_{ilm}(s)\right] = \begin{cases} \rho_{\mathrm{w}} = \tau_{\gamma}^2/\left(\tau_{\gamma}^2 + \sigma_{\epsilon}^2\right), & j = l \\ \rho_{\mathrm{b},|j-l|} = \tau_{\gamma}^2 r^{|j-l|}/\left(\tau_{\gamma}^2 + \sigma_{\epsilon}^2\right), & j \neq l \end{cases}$$

An example matrix form the exponential decay correlation structure is provided in Table 2.

For sample size estimation, the variance of $\hat{\delta}$ may not be obtained analytically with the exponential decay model, but could be computed numerically following the general variance formula of Kasza et al.[57] Kasza et al.[57] compared $\operatorname{var}\left(\hat{\delta}\right)$ using the nested exchangeable correlation model and the exponential decay model, and concluded that $\operatorname{var}\left(\hat{\delta}\right)$ was sensitive to the random-effects assumptions. Specifically, when the exponential decay model is the true model, the variance could either be overestimated or underestimated if the nested exchangeable correlation model is incorrectly assumed, and vice versa. Therefore, Kasza et al.[57] recommended examination of the plausibility of alternative correlation structures based on preliminary data, whenever possible. From a data analytic perspective, Kasza and Forbes[61] further considered the misspecification of the random-effects structure on the estimation of the treatment effect and variance components. They found that incorrectly omitting the decay parameter r (namely assuming the Hussey and Hughes model or the nested exchangeable correlation model when the true model induces an exponential correlation decay) might lead to an inflated type I error rate and invalid inference.

The exponential decay model (18) assumes that the correlation decay is a function of the distance between time periods, which is considered appropriate if all individuals in the same period are measured at approximately the same time. For this reason, this model is also more explicitly referred to as the *discrete-time* exponential decay model. Grantham et al.[62] extended the discrete-time exponential decay model to accommodate continuous enrollment, and allowed for the correlation decay to depend on the distance between the actual measurement times of each individual. They concluded that incorrectly assuming the Hussey and Hughes model in the presence of continuous-time correlation decay would likely underestimate the required sample size in the design stage. We are not aware of any existing numerical studies that examine the implications for the statistical analysis due to continuous correlation decay. In fact, Hooper and Copas indicated that the current literature on stepped wedge designs had not differentiated between continuous enrollment and discrete individual sampling, and therefore new statistical models and methods would be required to address the challenges associated with continuous enrollment.[63]

3.4.3 Random intervention model—Several authors have suggested extensions to the Hussey and Hughes model and accounted for potential variation across clusters in the magnitude of intervention effects.[47,55,64] For example, Hemming et al.[47,64] considered a model parameterized as

$$Y_{ijk} = \mu + \beta_j + \delta X_{ij} + \alpha_{1i} X_{ij} + \alpha_{0i}\left(1 - X_{ij}\right) + \epsilon_{ijk} \tag{21}$$

where

$$\begin{pmatrix}\alpha_{1i}\\ \alpha_{0i}\end{pmatrix}\sim N\left[\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}\tau_1^2 & \sigma_{10}\\ \sigma_{10} & \tau_0^2\end{pmatrix}\right]$$

and σ_{10} is the possibly non-zero covariance between random effects α_{1i} and α_{01}. Within our general model representation (4), this model assumes the heterogeneity term to be

$$\boldsymbol{R}_{ik}(j,s)'\boldsymbol{\alpha}_i = \alpha_{1i}\mathbb{I}[j \geq s] + \alpha_{0i}\mathbb{I}[j < s]$$

and now depends on the intervention sequence assigned for cluster i. The heterogeneity term can also be considered as an interaction between the random cluster effect and treatment assignment. An implication of this interaction term is that the intervention not only affects the group average through $F_i^1(j,s)'\Delta(j,s) = \delta X_{ij}$, but also affects the marginal dispersion through the variance components. Hemming et al.[64] showed that the following within-cluster correlation structure holds

$$\operatorname{corr}\left[Y_{ijk}(s), Y_{ilm}(s)\right]\begin{cases}\rho_0 = \tau_0^2/\left(\tau_0^2+\sigma_\epsilon^2\right), & j<s, l<s\\ \rho_1 = \tau_1^2/\left(\tau_1^2+\sigma_\epsilon^2\right), & j\geq s, l\geq s\\ \rho_{10} = \sigma_{10}/\left\{\sqrt{\tau_0^2+\sigma_\epsilon^2}\sqrt{\tau_1^2+\sigma_\epsilon^2}\right\}, & j\geq s, l<s, \text{ or } j<s, l\geq s\end{cases}$$

where ρ_0 is the correlation for two observations collected under the control condition, ρ_1 is the correlation for two observations collected under the intervention condition, and ρ_{10} is the correlation for two observations collected under different conditions (one under control and the other under intervention condition). The random intervention model does not permit a closed-form derivation of the variance, $\operatorname{var}\left(\hat{\delta}\right)$, and therefore sample size estimates must proceed by numerical calculations. To date, only simulation-based approaches have been examined to estimate sample size from the random intervention model.[65]

An alternative parameterization of (21) is to directly include a random cluster-by-treatment interaction in the Hussey and Hughes model.[55] The model can be written as

$$Y_{ijk} = \mu + \beta_j + (\delta + v_i)X_{ij} + \alpha_i + \epsilon_{ijk} \tag{22}$$

where

$$\begin{pmatrix}\alpha_i\\ v_i\end{pmatrix}\sim N\left[\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}\tau_\alpha^2 & \sigma_{\alpha v}\\ \sigma_{\alpha v} & \tau_v^2\end{pmatrix}\right]$$

and σ_{av} is a possibly non-zero covariance between α_i and v_i. This model assumes the heterogeneity term

$$\boldsymbol{R}_{ik}(j,s)'\boldsymbol{\alpha}_i = \alpha_i + v_i\mathbb{I}[j \geq s]$$

Hemming et al.[64] discussed alternative parameterizations that allowed for treatment effect heterogeneity, and recommended the use of equations (21) and (22) because other parameterizations induced unnecessary and sometimes implausible assumptions on the correlation structure. Baio et al.[65] pointed out that term $(\delta + v_i)$ in model (22) could be interpreted as a cluster-varying random slope for the intervention effect.

3.4.4 Random coefficient model—Another modeling technique, proposed for analyzing parallel longitudinal CRTs, is the random coefficient model.[58] Although such a model has not yet been formally investigated in the context of stepped wedge designs, there has been recent interest in exploring their operating characteristics (Section 4 of Kasza and Forbes[61] mentioned such models), and we briefly discuss the model assumptions here. The random coefficient model usually specifies a linear secular trend but allows for cluster-specific time slopes

$$Y_{ijk} = \mu + (\beta + \xi_i)T_j + \delta X_{ij} + \alpha_i + \epsilon_{ijk} \tag{23}$$

In this model, we use $T_j = j$ to represent the linear time basis function, β as the fixed time slope and ξ_i as the random slope. Within the general model representation, the heterogeneity term is written as

$$\boldsymbol{R}_{ik}(j,s)'\boldsymbol{\alpha}_i = \alpha_i + j\xi_i$$

The random intercept and slope are assumed to be independent of the residual error, but could covary following a bivariate normal distribution

$$\begin{pmatrix}\alpha_i \\ \xi_i\end{pmatrix} \sim N\left[\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\tau_\alpha^2 & \sigma_{\alpha\xi} \\ \sigma_{\alpha\xi} & \tau_\xi^2\end{pmatrix}\right]$$

The following within-cluster correlation structure between a pair of outcomes results from the above model

$$\text{corr}\left[Y_{ijk}(s), Y_{ilm}(s)\right] = \rho_{jl} = \frac{\tau_\alpha^2 + (j+l)\sigma_{\alpha\xi} + jl\tau_\xi^2}{\sqrt{\tau_\alpha^2 + 2j\sigma_{\alpha\xi} + j^2\tau_\xi^2 + \sigma_\epsilon^2}\sqrt{\tau_\alpha^2 + 2l\sigma_{\alpha\xi} + I^2\tau_\xi^2 + \sigma_\epsilon^2}}$$

which is specific to both time period indices j and l. It is not immediate what correlation pattern is implied from the above expression, except that it is symmetric, namely, $\rho_{jl} = \rho_{lj}$. Therefore, we plot the within-period and between-period ICCs in a hypothetical trial with J=5 periods in Figure 3 under different assumptions of the covariance parameters. When ran $\sigma_{a\xi} \geq 0$ as in panels (b) and (c), our finding suggests that the within-period ICC is often an increasing function of time j. In addition, the between-period ICC also increases as the distance in time, $|j - l|$, increases, which is opposite to the correlation structure implied by

the exponential decay model. Finally, when the covariance ran $\sigma_{a\xi} < 0$, the random coefficient model could imply negative between-period ICCs. These observations point out that the within-cluster correlation structure induced from the random coefficient model may be challenging to interpret, especially because the pattern of the between-period ICC contradicts that of the exponential decay model, and the latter has been considered plausible in several settings.[57] Further simulation and methodological investigations are required to study the performance and interpretation of the random coefficient model versus other alternatives in the context of stepped wedge designs.

3.5 Considerations for modeling heterogeneity in closed-cohort designs

3.5.1 The basic model—Considerations for closed-cohort designs were discussed in Copas et al.,[30] and a simple extension to the Hussey and Hughes model was introduced in Baio et al.[65] Specifically, the basic model is written as

$$Y_{ijk} = \mu + \beta_j + \delta X_{ij} + \alpha_i + \phi_{ik} + \epsilon_{ijk} \tag{24}$$

where $\phi_{ik} \sim N\left(0, \tau^2_{\phi}\right)$ is the random effect for the repeated measures from individual k in cluster i, and it is assumed to be independent of random cluster effect α_i. The heterogeneity term

$$\boldsymbol{R}_{ik}(j, s)'\boldsymbol{\alpha}_i = \alpha_i + \phi_{ik} \tag{25}$$

is modeled as a function of cluster index i and individual index k. This model assumption induces the following *nested exchangeable* within-cluster correlation structure

$$\operatorname{corr}\left[Y_{ijk}(s), Y_{ilm}(s)\right] = \begin{cases} \rho_{\mathrm{a}} = \left(\tau^2_{\alpha} + \tau^2_{\phi}\right) / \left(\tau^2_{\alpha} + \tau^2_{\phi} + \sigma^2_{\epsilon}\right), & k = m \\ \rho_{\mathrm{d}} = \tau^2_{\alpha} / \left(\tau^2_{\alpha} + \tau^2_{\phi} + \sigma^2_{\epsilon}\right), & k \neq m \end{cases}$$

where ρ_{a} is the correlation between two repeated measurements from the same individual (termed the within-individual ICC following Li et al.[34]) and ρ_{d} is the correlation between two observations collected from different individuals, regardless of time periods. Although not directly pointed out by Baio et al.,[65] the additive random structure permits a closed-form derivation of the variance of the intervention effect for trial planning once we assume equal cohort sizes, $N_i = N$. In particular, using the results of Li et al.,[34] one can show that $\operatorname{var}\left(\hat{\delta}\right)$ shares the same form with expression (7), except that we replace

$$\sigma^2_{\mathrm{tot}} = \tau^2_{\alpha} + \tau^2_{\phi} + \sigma^2_{\epsilon},$$

$$\lambda_1 = 1 - \rho_{\mathrm{a}},$$

$$\lambda_2 = 1 + J(N-1)\rho_{\mathrm{d}} + (J-1)\rho_{\mathrm{a}}$$

where λ_1 and λ_2 are again the two eigenvalues of the within-cluster correlation matrix implied by model (24).

This basic model suggested in Baio et al.[65] has the same limitation as the Hussey and Hughes model, that is, the limit of the variance, $\lim_{N\to\infty} \text{var}(\hat{\delta})$, converges to zero as the cohort size approaches infinity. In other words, the required number of clusters converges to one for any given level of power as long as one increases the cohort sizes indefinitely, which may not be realistic. Nevertheless, models assuming the heterogeneity term (25) appeared in a few previous investigations, including all models used in the simulation study of Nickless et al.[49]

3.5.2 Block exchangeable correlation model—The nested exchangeable correlation model was extended to include a similar individual-level random intercept to account for the correlations between repeated measures. The model appeared in Hooper et al.[56] and Girling and Hemming[38] as

$$Y_{ijk} = \mu + \beta_j + \delta X_{ij} + \alpha_i + \gamma_{ij} + \phi_{ik} + \epsilon_{ijk} \tag{26}$$

where $\phi_{ik} \sim N(0, \tau_\phi^2)$ is the random effect for the repeated measures from individual k in cluster i, and is assumed to be independent of α_i and γ_{ij} (α_i and γ_{ij} are defined earlier in equation (13)). Using the notation of the general model, the block exchangeable correlation model represents the heterogeneity by

$$\boldsymbol{R}_{ik}(j,s)'\boldsymbol{\alpha}_i = \alpha_i + \gamma_{ij} + \phi_{ik}$$

which depends on cluster i, period j as well as individual k. Three ICC parameters are implied by the block exchangeable correlation model in the cohort setting, as we can write

$$\text{corr}\left[Y_{ijk}(s), Y_{ilm}(s)\right] = \begin{cases} \rho_\text{a} = \left(\tau_\alpha^2 + \tau_\phi^2\right)/\left(\tau_\alpha^2 + \tau_\gamma^2 + \tau_\phi^2 + \sigma_\epsilon^2\right), & j \neq l, k = m \\ \rho_\text{w} = \left(\tau_\alpha^2 + \tau_\gamma^2\right)/\left(\tau_\alpha^2 + \tau_\gamma^2 + \tau_\phi^2 + \sigma_\epsilon^2\right), & j = l, k \neq m \\ \rho_\text{b} = \tau_\alpha^2/\left(\tau_\alpha^2 + \tau_\gamma^2 + \tau_\phi^2 + \sigma_\epsilon^2\right), & j \neq l, k \neq m \end{cases}$$

where ρ_a is the within-individual ICC for repeated measures, ρ_w and ρ_b are the within-period and between-period ICCs which have the same interpretations as their corresponding counterparts in the cross-sectional model. Constant values are assumed for three types of ICCs, and therefore the correlation structure does not depend on the intervention sequence s. An example matrix form the block exchangeable correlation structure is provided in Table 2.

In the closed-cohort setting, model (26) induces the so-called *block exchangeable* correlation structure, due to the fact that if the correlation structure is written in a matrix form, the exchangeability holds both within and across periods.[34] The variance expression for the treatment effect was derived in Hooper et al.,[56] Girling and Hemming,[38] and Li et al.,[34,66] using different notation. In our notation, the expression of $\text{var}(\hat{\delta})$ is the same as equation (7), except that we will redefine the total variance and eigenvalues by

$$\sigma^2_{\text{tot}} = \tau^2_\alpha + \tau^2_\gamma + \tau^2_\phi + \sigma^2_\epsilon,$$

$$\lambda_1 = 1 + (N-1)(\rho_{\text{w}} - \rho_{\text{b}}) - \rho_{\text{a}},$$

$$\lambda_2 = 1 + (N-1)\rho_{\text{w}} + (J-1)(N-1)\rho_{\text{b}} + (J-1)\rho_{\text{a}}$$

Further, as the cohort size N increases to infinity, the limit of the variance, $\lim_{N\to\infty} \text{var}\left(\hat{\delta}\right)$, is given in equation (17), which is a positive constant as long as $\rho_{\text{b}} \neq \rho_{\text{w}}$. Therefore, the block exchangeable correlation model is considered more realistic than the basic model in the cohort setting, for the same reason argued in Taljaard et al.[45] Finally, we can see that models (5), (13) and (24) are all nested in the block exchangeable correlation model (26).

The block exchangeable correlation model has facilitated the investigation of several design questions in the closed-cohort settings. For example, Li et al.[34] reported the roles of the three ICC parameters for design efficiency. In particular, they found that larger values of the within-period ICC reduced the design efficiency, just as the traditional ICC did in a parallel design. However, larger values of both the between-period ICC and/or within-individual ICC increase the design efficiency. Further, optimal closed-cohort designs were reported in Li et al.,[66] who generalized the earlier findings in Lawrie et al. based on the Hussey and Hughes model.[36] Girling and Hemming[38] derived the optimal design within a larger design space that includes hybrid designs, and found the hybrid design to be the most efficient within the larger design space. Their results apply to both cross-sectional and closed-cohort designs. Grayling et al.[67] developed an algorithm to search for admissible (cohort) stepped wedge designs in the presence of multiple intervention arms.[68] Girling[69] studied the relative efficiency of unequal cluster sizes versus balanced cluster sizes in closed-cohort designs, and reported that the loss of precision due to unequal cluster sizes was usually no more than 12%, which was consistent with prior investigations in cross-sectional designs.[42] Defining CV as the coefficient of variation for cohort sizes, Girling[69] showed that inflating the required cohort size by a factor of $(1 + \text{CV}^2)$, as one would do in a parallel CRT,[70] provided a valid but conservative sample size estimate for cohort stepped wedge trials. Finally, because the nested exchangeable correlation model is a special case of the block exchangeable correlation model, these results derived under the latter model apply to the cross-sectional setting by setting $\tau^2_\phi = 0$.

3.5.3 Proportional decay model—Li[71] proposed a model for the design and analysis of cohort stepped wedge designs that allowed the exponential decay of between-period ICC and within-individual ICC over time. However, Li[71] focused on a population-averaged model that allowed the direct characterization of the within-cluster correlation structure, but did not consider a mixed-effects model counterpart. With a continuous outcome Y_{ijk}, we are able to find a conditional model that leads to the same inference as the marginal model

discussed in Li.[71] Specifically, the conditional model that allows correlation decay in the closed-cohort design setting shares the same form as the exponential decay model

$$Y_{ijk} = \mu + \beta_j + \delta X_{ij} + \gamma_{ij} + \epsilon_{ijk} \tag{27}$$

where the heterogeneity term is

$$\boldsymbol{R}_{ik}(j,s)'\boldsymbol{\alpha}_i = \gamma_{ij}$$

However, in addition to assuming $\gamma_i = (\gamma_{i1}, \ldots, \gamma_{iJ})' \sim N\left(0, \tau_\gamma^2 \boldsymbol{M}(1,r)\right)$, we further assume a similar autoregressive structure for residual errors of the kth person in cluster i as

$$\boldsymbol{\epsilon}_{ik} = (\epsilon_{i1k}, \ldots, \epsilon_{iJk})' \sim N\left(0, \sigma_\epsilon^2 \boldsymbol{M}(1,r)\right), \quad \boldsymbol{\epsilon}_{ik} \perp \boldsymbol{\epsilon}_{im}, \quad k \neq m$$

where r is the decay rate shared by $\boldsymbol{\gamma}_i$ and $\boldsymbol{\epsilon}_{ik}$, and $\boldsymbol{\gamma}_i \perp \boldsymbol{\epsilon}_{ik}$.

The above decay model implies a *proportional decay* correlation structure that dates back to the analysis of multilevel longitudinal data.[72,73] The within-cluster correlations between each pair of observations is

$$\operatorname{corr}\left[Y_{ijk}(s), Y_{ilm}(s)\right] = \begin{cases} \rho_{\mathrm{a},|j-l|} = r^{|j-l|}, & j \neq l, k = m, \\ \rho_{\mathrm{w}} = \tau_\gamma^2 / \left(\tau_\gamma^2 + \sigma_\epsilon^2\right), & j = l, k \neq m, \\ \rho_{\mathrm{b},|j-l|} = \tau_\gamma^2 r^{|j-l|} / \left(\tau_\gamma^2 + \sigma_\epsilon^2\right), & j \neq l, k \neq m \end{cases}$$

where $\rho_{\mathrm{a},|j-l|}$ is the within-individual ICC that decays exponentially over time, ρ_{w} and $\rho_{\mathrm{b},|j-l|}$ are the within-period and between-period ICCs just as their counterparts in the exponential decay model. This correlation model is termed the proportional decay model as the same decay rate r applies to both the within-individual ICC and the between-period ICC for different individuals. An example matrix form the proportional decay correlation structure is provided in Table 2. A unique feature of the proportional decay correlation structure is that the correlation matrix can be written as a Kronecker product between an exchangeable correlation and a first-order autoregressive matrix.[73] This separability property allows one to derive a closed-form variance for the intervention effect to facilitate sample size and power calculation.

Under equal cohort sizes $N_i = N$, Li[71] showed that

$$\operatorname{var}\left(\hat{\delta}\right) = \frac{\left(\sigma_{\mathrm{tot}}^2 / N\right) I\left(1 - r^2\right)\{1 + (N-1)\rho_{\mathrm{w}}\}}{(IU - W)\left(1 + r^2\right) - 2(IP - Q)r} \tag{28}$$

where $U = \sum_{i=1}^{I} \sum_{j=1}^{J} X_{ij}$ and $W = \sum_{j=1}^{J} \left(\sum_{i=1}^{I} X_{ij}\right)^2$ are defined earlier as in the Hussey and Hughes model,[9]. and $P = \sum_{i=1}^{I} \sum_{j=1}^{J-1} X_{ij} X_{i,j+1}$,

$Q = \sum_{j=1}^{J-1}\left(\sum_{i=1}^{I} X_{ij}\right)\left(\sum_{i=1}^{I} X_{i,j+1}\right)$ are cross-product terms resulting from the first-order autoregressive decay. As the cohort size increases to infinity

$$\lim_{N\to\infty} \mathrm{var}\left(\hat{\delta}\right) = \frac{\sigma_{\mathrm{tot}}^2 I\left(1-r^2\right)\rho_{\mathrm{w}}}{(IU-W)\left(1+r^2\right)-2(IP-Q)r} > 0$$

which is a positive constant as long as |r| < 1. This variance expression suggests that the proportional decay model is not subject to the same criticism as the basic model (24). The variance expression also permits us to study the role of decay, r, on design efficiency. Li[71] further presented a closed-form expression of the design effect based on equation (27), and demonstrated the parabolic relationship between $\mathrm{var}\left(\hat{\delta}\right)$ and r, when all other parameters were held constant.

3.5.4 Random intervention model—The random intervention model in the closed-cohort design has been considered in Kasza et al.,[27] although in the context of incomplete designs where outcomes may not be measured in certain cluster-periods (e.g. trials with implementation periods where outcome data are not collected). The model can be represented by

$$Y_{ijk} = \mu + \beta_j + (\delta + v_i)X_{ij} + \gamma_{ij} + \phi_{ik} + \epsilon_{ijk} \tag{29}$$

where $\phi_{ik} \sim N\left(0, \tau_\phi^2\right)$ is the random effect for the repeated measures from individual k in cluster i, $\epsilon_{ijk} \sim N\left(0, \sigma_\epsilon^2\right)$ is the residual error, v_i is the cluster-specific random intervention effect, and γ_{ij} is the cluster-period-specific random deviation from the group average, as in the exponential decay model. Clearly, the heterogeneity term is modeled as

$$\boldsymbol{R}_{ik}(j,s)'\boldsymbol{\alpha}_i = \gamma_{ij} + \phi_{ik} + v_i \mathbb{I}[j \geq s]$$

Kasza et al.[27] assumed the following correlation pattern for the remaining set of random effects

$$(\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{iJ}, v_i)' \sim N\left[\begin{pmatrix}\mathbf{0}_J \times 1\\ 0 1 \times 1\end{pmatrix}, \begin{pmatrix}\tau_\gamma^2 \widetilde{\boldsymbol{M}} & \sigma_{\gamma v}\mathbf{1}\\ \sigma_{\gamma v}\mathbf{1}' & \tau_v^2\end{pmatrix}\right]$$

and the vector $(\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{iJ}, v_i)'$ was assumed to be independent of ϕ_{ik} and ϵ_{ijk}. In the above covariance structure, τ_γ^2, τ_v^2 are variance components for γ_{ij} and v_i, $\sigma_{\gamma v}$ is the possibly non-zero covariance between them, $\mathbf{1}$ is the $J \times 1$ matrix of ones, and $\widetilde{\boldsymbol{M}}$ is the symmetric Toeplitz matrix defined in equation (19). The complicated random-effects structure in fact distinguishes between eight types of ICC parameters. Specifically, when two observations are measured under the control condition (namely, $j < s$, $l < s$), the within-period, between-period and within-individual ICCs are

$$\operatorname{corr}\left[Y_{ijk}(s), Y_{ilm}(s)\right] = \begin{cases} \rho_{\mathrm{w}} = \dfrac{\tau_\gamma^2}{\tau_\gamma^2 + \tau_\phi^2 + \sigma_\epsilon^2}, & j = l, k \neq m \\ \rho_{\mathrm{b}} = \dfrac{\tau_\gamma^2 r_{jl}}{\tau_\gamma^2 + \tau_\phi^2 + \sigma_\epsilon^2}, & j \neq l, k \neq m \\ \rho_{\mathrm{a}} = \dfrac{\tau_\gamma^2 r_{jl} + \tau_\phi^2}{\tau_\gamma^2 + \tau_\phi^2 + \sigma_\epsilon^2}, & j \neq l, k = m \end{cases}$$

When two observations are measured under the intervention condition (namely, $j \geq s$, $l \geq s$), the within-period, between-period, and within-individual ICCs become

$$\operatorname{corr}\left[Y_{ijk}(s), Y_{ilm}(s)\right] = \begin{cases} \rho_{\mathrm{w}} = \dfrac{\tau_\gamma^2 + \tau_\nu^2 + 2\sigma_{\gamma\nu}}{\tau_\gamma^2 + \tau_\phi^2 + \tau_\nu^2 + 2\sigma_{\gamma\nu} + \sigma_\epsilon^2}, & j = l, k \neq m \\ \rho_{\mathrm{b}} = \dfrac{\tau_\gamma^2 r_{jl} + \tau_\nu^2 + 2\sigma_{\gamma\nu}}{\tau_\gamma^2 + \tau_\phi^2 + \tau_\nu^2 + 2\sigma_{\gamma\nu} + \sigma_\epsilon^2}, & j \neq l, k \neq m \\ \rho_{\mathrm{a}} = \dfrac{\tau_\gamma^2 r_{jl} + \tau_\nu^2 + 2\sigma_{\gamma\nu} + \tau_\phi^2}{\tau_\gamma^2 + \tau_\phi^2 + \tau_\nu^2 + 2\sigma_{\gamma\nu} + \sigma_\epsilon^2} & j \neq l, k = m \end{cases}$$

Finally, when one observation is measured under the control condition while the other one under the intervention condition ($j \geq s$, $l < s$ or $j < s$, $l \geq s$), the correlations are

$$\operatorname{corr}\left[Y_{ijk}(s), Y_{ilm}(s)\right] = \begin{cases} \rho_{\mathrm{b}} = \dfrac{\tau_\gamma^2 r_{jl} + \sigma_{\gamma\nu} + \tau_\phi^2}{\sqrt{\tau_\gamma^2 + \tau_\phi^2 + \tau_\nu^2 + 2\sigma_{\gamma\nu} + \sigma_\epsilon^2}\sqrt{\tau_\gamma^2 + \tau_\phi^2 + \sigma_\epsilon^2}}, & k = m \\ \rho_{\mathrm{a}} = \dfrac{\tau_\gamma^2 r_{jl} + \sigma_{\gamma\nu}}{\sqrt{\tau_\gamma^2 + \tau_\phi^2 + \tau_\nu^2 + 2\sigma_{\gamma\nu} + \sigma_\epsilon^2}\sqrt{\tau_\gamma^2 + \tau_\phi^2 + \sigma_\epsilon^2}}, & k \neq m \end{cases}$$

As we explained when we reviewed the exponential decay model in Section 3.4, parsimonious parameterization of $\widetilde{\mathbf{M}}$ may lead to simpler and more interpretable models. For example, when $\widetilde{\mathbf{M}} = \mathbf{M}(1, 1)$, model (29) is a direct extension of model (22) by the addition of the random intercept ϕ_{ik}. When $\widetilde{\mathbf{M}} = \mathbf{M}(r_0, 1)$, model (29) extends the block exchangeable correlation model (26) with the addition of a random intervention component. When $\widetilde{\mathbf{M}} = \mathbf{M}(1, r)$, model (29) extends the exponential decay model (18) by the addition of random intercept ϕ_{ik} and random intervention effect v_i. Notice that in the last case, there is no guarantee that the between-period ICC for any pair of observations decays at an exponential rate, and therefore model (29) does not nest the proportional decay model (27), even though both models are developed for closed-cohort designs.

3.6 Considerations for modeling heterogeneity in open-cohort designs

The model development in the cross-sectional and closed-cohort designs have important implications for the open-cohort design, as an open-cohort design can be considered a mix of the former two. Kasza et al.[74] recently discussed several open-cohort sampling schemes

for stepped wedge designs and proposed a corresponding sample size calculation procedure based on a linear mixed model. We will review the related model variants and their connections to the results in Sections 3.4 and 3.5.

3.6.1 Blended exchangeable correlation model—In principle, the block exchangeable model developed for the closed-cohort design can still be used to represent the outcome trajectory in the open-cohort design, except for a few notational caveats. Specifically, the outcome model can still be written as

$$Y_{ijk} = \mu + \beta_j + \delta X_{ij} + \alpha_i + \gamma_{ij} + \phi_{ik} + \epsilon_{ijk} \tag{30}$$

where all the parameters are defined in Section 3.5.2. Importantly, under the attrition of members from and addition of new members to the original cohort, we shall use a distinct subscript k to represent a distinct individual in each cluster. The implied within-cluster correlation matrix is neither nested exchangeable nor block exchangeable, but becomes a blend of these two. We call such a matrix a *blended exchangeable correlation structure* and an example formulation is provided in Table 2.

Assuming that the cluster-period sizes were identical ($N_{ij} = N$) and there existed the same number of overlapping individuals between any two periods ($n_1(j, l) = n$), Kasza et al.[74] derived a closed-form variance of the intervention effect, which could be rewritten in our notation as

$$\begin{aligned}&\mathrm{var}\left(\hat{\delta}\right)\\ &= \frac{\left(\sigma^2_{\mathrm{tot}}/N\right)IJ\{\lambda_1 + \chi(\rho_{\mathrm{a}} - \rho_{\mathrm{b}})\}\{\lambda_2 - \chi(J-1)(\rho_{\mathrm{a}} - \rho_{\mathrm{b}})\}}{\left(U^2 + IJU - JW - IV\right)\{\lambda_2 - \chi(J-1)(\rho_{\mathrm{a}} - \rho_{\mathrm{b}})\} - \left(U^2 - IV\right)\{\lambda_1 + \chi(\rho_{\mathrm{a}} - \rho_{\mathrm{b}})\}}\end{aligned} \tag{31}$$

where U, V, W were design constants defined in Section 3.1

$$\sigma^2_{\mathrm{tot}} = \tau^2_{\alpha} + \tau^2_{\gamma} + \tau^2_{\phi} + \sigma^2_{\epsilon},$$

$$\lambda_1 = 1 + (N-1)(\rho_{\mathrm{w}} - \rho_{\mathrm{b}}) - \rho_{\mathrm{a}},$$

$$\lambda_2 = 1 + (N-1)\rho_{\mathrm{w}} + (J-1)(N-1)\rho_{\mathrm{b}} + (J-1)\rho_{\mathrm{a}}$$

and $\chi = 1 - n/N \in [0, 1]$ was the common rate of attrition or *churn rate*. This expression permits a convenient sample size formula for open-cohort designs, and unifies the variance expressions derived under models (5), (13), (24) and (26). For example, as the churn rate approaches one, the open-cohort design reduces to the cross-sectional design and the variance (31) reduces to the variance derived under the nested exchangeable model (13). On the other hand, as the churn rate approaches zero, the open-cohort design reduces to the closed-cohort design and variance (31) reduces to the one derived under the block exchangeable correlation model. This unified perspective represents a continuum between

cross-sectional and closed-cohort designs, and may help with the efficiency comparisons between these two designs. In fact, when the within-individual ICC ρ_a is larger than the between-period ICC ρ_b, variance (31) is a monotonically increasing function of χ over [0, 1] Within a random-effects model (30), ρ_a is constrained to be no smaller than ρ_b, and therefore the closed-cohort design is usually more efficient than the cross-sectional design, provided other parameters are all held equal. On the contrary, if the within-individual ICC ρ_a is smaller than the between-period ICC ρ_b, variance (31) becomes a monotonically decreasing function of χ over [0, 1]. In this case, the closed-cohort design becomes less efficient than the cross-sectional design, providing the remaining parameters are held equal. However, even though the latter case is mathematically valid (because the resulting correlation matrix can still be positive definite, see the eigenvalue conditions of Li et al.[34]), it may not be plausible in practice because serial correlation defined for the same individual is usually believed to be stronger than correlation between individuals.

3.6.2 Blended correlation decay model—Kasza et al.[74] introduced a linear mixed model that allowed correlation decay in open-cohort designs. The model has the same conditional mean structure as the exponential decay and the proportional decay model and is written as

$$Y_{ijk} = \mu + \beta_j + \delta X_{ij} + \gamma_{ij} + \epsilon_{ijk} \tag{32}$$

where $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{iJ})' \sim N\left(0, \tau_\gamma^2 \boldsymbol{M}(1, r)\right)$ and r is the decay rate at the cluster-period level. Here, we use a distinct subscript k to represent a distinct individual in each cluster to allow for open-cohort sampling. If individual k in cluster i contributes outcome observations in a total of $J_k \leq J$ periods, the model assumes an autoregressive structure for errors of that individual as

$$\boldsymbol{\epsilon}_{ik} = \left(\epsilon_{i1k}, \ldots, \epsilon_{iJ_k k}\right)' \sim N\left(0, \sigma_\epsilon^2 \mathbf{M}(1, \eta)\right), \quad \boldsymbol{\epsilon}_{ik} \perp \boldsymbol{\epsilon}_{im}, \quad k \neq m$$

where η is the decay rate at the individual level, and the two random effects are independent, $\boldsymbol{\gamma}_i \perp \boldsymbol{\epsilon}_{ik}$. Notice that this *blended correlation decay model* is more general than the proportional decay model because? the individual-level decay rate η is allowed to differ from the cluster-period-level decay rate r.

The blended correlation decay model implies the following correlation structure

$$\text{corr}\left[Y_{ijk}(s), Y_{ilm}(s)\right] = \begin{cases} \rho_{a,|j-l|} = \left(\tau_\gamma^2 r^{|j-l|} + \sigma_\epsilon^2 \eta^{|j-l|}\right) / \left(\tau_\gamma^2 + \sigma_\epsilon^2\right), & j \neq l, k = m, \\ \rho_w = \tau_\gamma^2 / \left(\tau_\gamma^2 + \sigma_\epsilon^2\right), & j = l, k \neq m, \\ \rho_{b,|j-l|} = \tau_\gamma^2 r^{|j-l|} / \left(\tau_\gamma^2 + \sigma_\epsilon^2\right), & j \neq l, k \neq m, \end{cases}$$

where $\rho_{a,|j-l|}$ is the within-individual ICC that decays over time depending on both η and r. Furthermore, ρ_w and $\rho_{b,|j-l|}$ are the within-period and between-period ICCs just as their counterparts in the exponential decay model and the proportional decay model. The blended correlation decay model unifies the exponential decay and proportional decay models. For

instance, when the churn rate approaches one, the open-cohort design reduces to the cross-sectional design, and the blended correlation decay structure reduces to the exponential decay structure. On the other hand, as the churn rate approaches zero, the open-cohort design reduces to the closed-cohort design, and the blended correlation decay structure becomes the proportional decay structure when the two decay rates are identical, namely $\eta = r$. This indicates that the blended correlation decay model represents a continuum between cross-sectional and closed-cohort designs, and will be helpful for comparing efficiency between cross-sectional and closed-cohort designs under a range of correlation decay parameters. We provide an illustrative matrix form of the blended correlation decay structure in Table 2, where we assume an equal decay rate at each level ($\eta = r$). This illustrative formulation in Table 2 also shows that the blended correlation decay structure is a "blend" of the exponential decay and proportional decay structures. Unlike the blended exchangeable correlation model, the blended correlation decay structure does not admit a closed-form variance expression of the intervention effect, even when the churn rate is assumed to be a constant. Kasza et al.[74] provided a general matrix-based variance formula for numerically computing sample size and power, with the two decay parameters as key input. However, empirical estimates of these decay rates are lacking, and additional research effort is necessary to examine the operating characteristics of model (32) for estimating these decay parameters in stepped wedge designs with realistic sample sizes.

3.7 Considerations for modeling binary outcomes

The literature on stepped wedge designs has largely focused on the application of linear mixed models and a continuous outcome, and includes few focused discussions of binary outcomes. For sample size estimation, Hussey and Hughes[9] used variance expression (5) derived from the linear mixed model, but approximated $\sigma_\epsilon^2 \approx \mu(1-\mu)$. In this particular case, the link function g is still identity and thus the intervention effect could be interpreted as the risk difference. Although this variance approximation may be adequate when there is minimal secular trend and a small intervention effect,[25,75] it may either underestimate or overestimate the true power in other parameter regions.[76] To accurately estimate sample size, Zhou et al.[76] proposed the following variant of the Hussey and Hughes model

$$\mu_{ij} = \mu + \beta_j + \delta X_{ij} + \alpha_i \tag{33}$$

where μ_{ij} is the proportion of responses in cluster i during period j, and the heterogeneity term $\boldsymbol{R}_{ik}(j, s)'\boldsymbol{\alpha}_i = \alpha_i$ now follows a truncated normal distribution with density

$$f\left(\alpha_i \mid \tau_\alpha^2\right) \propto \mathbb{I}\{-l_0 < \alpha_i < 1 - l_1\} \exp\left(-\frac{\alpha_i^2}{2\tau_\alpha^2}\right)$$

where the truncation points $l_0 = \min\{\mu + \beta_j, \mu + \beta_j + \delta; j = 1, \ldots, J\}$ and $l_1 = \max\{\mu + \beta_j, \mu + \beta_j + \delta; j = 1, \ldots, J\}$ are defined to ensure that the probability μ_{ij} is strictly bounded between zero and one. Based on this model, Zhou et al.[76] proposed a maximum likelihood approach to compute the sample size. It was shown that their approach provided more accurate

characterization of the required sample size than the binomial approximation in Hussey and Hughes.[9]

Since μ_{ij} is a proportion, other common choices of the link function include the log link and the logit link, with the respective interpretations of the intervention effect as a risk ratio and as an odds ratio. Although sample size methods based on these nonlinear link functions have not yet been extensively discussed (except for the simulation-based approach of Baio et al.[65]), there have been some investigations of the operating characteristics of these models as tools for data analysis. For example, in the cross-sectional setting, Thompson et al.[77] compared the performance between three logistic linear mixed models in a simulation study with varying parameter constellations. The three models they examined could be considered as the logistic version of the Hussey and Hughes model (5), nested exchangeable correlation model (13) and the random intervention model (22). They found that the following logistic counterpart of the nested exchangeable correlation model

$$\text{logit}(\mu_{ij}) = \mu + \beta_j + \delta X_{ij} + \alpha_i + \gamma_{ij}, \quad \alpha_i \sim N(0, \tau_\alpha^2), \quad \gamma_{ij} \sim N(0, \tau_\gamma^2)$$

had more robust performance in terms of bias and type I error rates across a number of data generating processes. Finally, the extension of the exponential decay model to binary outcomes and its operating characteristics have not yet been investigated.

4 Estimation and inference for the intervention effect

Estimation and inference for the parameters in mixed-effects models have been extensively discussed in a number of textbooks.[48,78–80] The basic principles, such as maximum likelihood, apply to all model formulations we have reviewed in Section 3. Although not our focus, the Bayesian approach is an alternative option, and could potentially be attractive especially in the presence of complex random-effects structures.[48] Using the general model (4) and assuming that the heterogeneity parameter α_i follows a parametric distribution $f(\alpha_i; \boldsymbol{\Theta})$, one could define the likelihood of the observed outcomes by generic notation as

$$L(\theta, \boldsymbol{\Theta}) = \prod_{i=1}^{I} \int \left[\prod_{j=1}^{J} \prod_{k=1}^{N_{ij}} f(Y_{ijk} \mid \theta, \alpha_i) \right] f(\alpha_i; \boldsymbol{\Theta}) \mathrm{d}\alpha_i \tag{34}$$

and numerically search for the values of fixed-effects parameter θ and variance components $\boldsymbol{\Theta}$ that maximize the likelihood. With continuous outcomes and the normality assumption for $f(Y_{ijk}|\theta, \alpha_i)$, it is often possible to obtain closed-form expressions for iterative updates between θ and $\boldsymbol{\Theta}$.[80] More often than not, equation (34) is modified to obtain the restricted maximum likelihood (REML), because the estimates of the variance component parameters $\boldsymbol{\Theta}$ will be unbiased. With binary outcomes and binomial assumptions for $f(Y_{ijk}|\theta, \boldsymbol{\Theta})$, approximation to (34) can be carried out via the Laplace method,[78] penalized quasi-likelihood[81] or adaptive Gauss-Hermite Quadrature,[82] among others. The variance of the MLE can be obtained from computing the approximate information matrix for $(\theta, \boldsymbol{\Theta})$. Testing the null hypothesis of no intervention effect (i.e. certain components of θ equal zero) can proceed by the Wald, likelihood ratio or score statistic based on the large-sample

normality theory. These procedures are available in standard software packages, such as SAS and R.

Cluster randomized trials usually involve a limited number of clusters, and therefore the desired frequentist properties may not be guaranteed for the hypothesis testing procedures derived from large-sample theory. Recent systematic reviews confirmed that most stepped wedge CRTs recruited fewer than 30 clusters,[17,18] and so there could be an emerging interest in developing small-sample adaptation of existing testing procedures for better performance. In the recent CONSORT extension to stepped wedge CRTs, Hemming et al.[23] encouraged the incorporation of small-sample corrections in the analysis of stepped wedge designs, whenever appropriate (item 12a). Although there has not yet been much investigation of small-sample corrections for mixed-effects model-based tests applied to stepped wedge trials, there were previous reports of small-sample corrections in parallel CRTs that may inspire ideas. For example, Li and Redden[83] considered the Wald t-statistic (or the equivalent F-statistic) from the logistic linear mixed model in the analysis of parallel CRTs with 10 to 30 clusters. They compared five degree-of-freedom approximations in terms of type I error rates and power, across scenarios with varying ICCs and cluster sizes. They concluded that the between-within degree of freedom[84] carried the nominal type I error rates and had higher power than its competitors. The between-within approach divides the residual degree of freedom into the between-cluster and within-cluster portions. If a fixed-effect covariate changes within any cluster, the within-cluster degree of freedom is assigned to that effect; otherwise, the between-cluster degree of freedom is assigned to the effect. Such findings may or may not be directly generalizable to stepped wedge trials, because unlike the parallel CRT, the intervention status actually changes over time within a cluster. In fact, we can compute the between-within degree of freedom for testing the intervention effect in the Hussey and Hughes model to be $(I-1)J$, which tends to be larger than its counterpart in parallel CRTs. It remains to be explored which degree of freedom approximation would be adequate in small stepped wedge designs.

The permutation test is another attractive tool for the inference in CRTs due to its robustness in controlling test size.[85] Under the strong null hypothesis of no intervention effect, Gail et al.[86] demonstrated that the type I error rate of the permutation test will not exceed the nominal level, even in CRTs with a limited number of clusters. Murray et al.[87] and Li et al.[88,89] also showed that the permutation test could achieve a similar level of power as the model-based F-test, but had better control of test sizes. Several authors have considered permutation-based inference for the analysis of stepped wedge trials; the general idea is to obtain the reference distribution of a given test statistic by permuting the intervention sequences across clusters. For example, Wang and DeGruttola[90] and Ji et al.[31] considered the estimated treatment effect and the corresponding z-score (Wald statistic) as the test statistic for testing H_0: $\delta = 0$ based on the Hussey and Hughes model (5) and the nested exchangeable correlation model (13); they obtained the exact distribution of the statistic from randomly shuffling the intervention sequences within the randomization space characterized by the design configuration. They found that the specification of the random-effects structure (or more generally the heterogeneity term $\boldsymbol{R}_{ik}(j,s)'\boldsymbol{\alpha}_i$) only affected the power of the test, but not the validity, and therefore demonstrated its superiority over the

model-based test. Ren et al.[91] considered permuting the estimated treatment effect and the corresponding *z*-score obtained from a random intervention model (22), but reported an inflated type **I** error rate even when the random intervention model is correctly specified. This phenomenon arises likely because the intervention sequence affects both the mean and covariance structures, and the exchangeability assumption fails to hold under the null hypothesis. Others have considered more nonparametric test statistics. For example, Thompson et al.[92] proposed a test statistic based on combining the optimally weighted within-period comparisons (i.e. the vertical comparisons defined in Davey et al.[21] and Matthews and Forbes[93]), and developed a permutation test with fewer modeling assumptions. Their test can be applied to both continuous and binary outcomes, and has demonstrated adequate control of type **I** error rate in simulations. Kennedy-Shaffer et al.[94] proposed an ensemble test statistic that combined the within-period and between-period contrasts via the Synthetic Control method (their SC method) and difference-in-differences (their crossover method). The corresponding permutation test based on the ensemble statistic demonstrated higher power than the permutation test in Thompson et al.[92] and the permutation test based on mixed-effects models[31,90] when those models were misspecified. Hughes et al.[95] provided a design-based test statistic and characterized the closed-form variances of the statistic under permutation; they showed that the resulting test carried the nominal size even under misspecification of both the mean and covariance structures. Furthermore, since the closed-form permutation variance is derived analytically, the permutation test in Hughes et al.[95] dispenses with intensive enumerations and is considered computationally more efficient than previous proposals. To date, there has not been a comprehensive simulation study that evaluates the comparative performance of all of the above permutation tests under different data generating processes, and more investigations are needed to offer practical recommendations on optimal ways to conduct randomization-based inference for stepped wedge designs.

5 Discussion

We have provided an overview of mixed-effects models that have been applied to the design and analysis of stepped wedge CRTs. We offered a unified perspective from a general model formulation and illustrated that existing models in the literature were its special cases with different assumptions about the secular trend, intervention effect and sources of heterogeneity. Our overview suggests that the current literature on stepped wedge designs has placed more emphasis on modeling the between-cluster and between-time heterogeneity, compared to modeling the secular trend or the intervention effect. We conjecture that this is because a number of discussions have focused on sample size calculation, which becomes convenient based on a scalar intervention effect but still remains sensitive to the assumptions for the random-effects structure. However, given the possibility of a time-varying intervention effect, it will be important for future work to address implications of the alternative methods reviewed in Section 3.3 on sample size planning and data analysis. In addition, there is currently limited guidance on how to select the most appropriate random-effects structure in the context of stepped wedge designs. Murray et al.[58] explored the use of information criteria to select appropriate mixed-effects models for the analysis of parallel longitudinal CRTs, but recommended against them due to their unreliable performance.

More research on identifying the appropriate random-effects structure in both the design and analysis stages would be of substantial interest.

We found that there is more development for continuous outcomes than for binary, count or time-to-event outcomes, likely due to the availability of closed-form expressions for variance and ICCs. Although these closed-form expressions have helped us generate knowledge and insights on the role of various design parameters and facilitated the application of these new designs, the generalizability of such knowledge to binary or count outcomes requires further exploration. Zhou et al.[76] pointed out that binary outcomes were fairly common in stepped wedge trials, especially in health care studies with an implementation endpoint. However, accurate sample size methods for binary outcomes have only been developed based on the risk difference scale and a single random cluster intercept, as considered in Zhou et al.[76] It would be important to extend such approaches for risk ratio and odds ratio measures, and to accommodate more complex assumptions on the heterogeneity, such as a model with a random cluster-by-time interaction or correlation decay.[56,57] Regarding the analysis of stepped wedge trials, Thompson et al.[77] conducted simulation studies with binary outcomes and suggested that the logistic extension of the nested exchangeable correlation model performed well in terms of bias and coverage across several data generating processes. To date, there has been little work on count or rate outcomes. We are only aware of a simulation study by Scott et al.,[96] who used a Poisson log-linear mixed-effects model to simulate outcomes and examine the operating characteristics of population-averaged models estimated by generalized estimating equations (GEE).[97]

Methods for designing and analyzing stepped wedge trials with time-to-event outcomes also need further attention. In the THRio study,[98] Moulton et al.[99] discussed a log-rank type analysis to compare the incidence between intervention and control clusters within each period, analogous to the vertical comparison methods[21,93] in non-survival settings. They used a simulation-based approach to estimate the design effect relative to parallel cluster randomization which was then used to compute sample size and power. Zhan et al.[100] developed a discrete-time survival model for analyzing stepped wedge CRTs with terminal endpoints and interval censoring (as the exact event time could be unknown within each discrete period). The key insight is to reformulate the likelihood using a generalized linear mixed model for the binary event history indicators. In this regard, considerations in Section 3 may still apply, but additional research is necessary. Importantly, closed-form sample size estimation procedures and optimal design configurations based on such discrete-time survival models remain unavailable and are open questions for future studies.[101]

As an alternative to mixed-effects models, population-averaged models have been proposed to design and analyze parallel CRTs.[6,102] While the conditional model we discussed requires the specification of a conditional mean structure with an association structure induced by random-effects, the population-averaged model counterpart requires the specification of a marginal mean and a separate correlation model for the association structure.[103,104] The conditional intervention effect from the mixed-effects model and the marginal intervention effect from the population-averaged model are identical with an identity link but could be different with a nonlinear link function.[104,105] Further, the interpretation of the marginal intervention effect remains the same regardless of the correlation model, while the

interpretation of the conditional intervention effect may change according to specifications of random effects.[106] Though the population-averaged model has several attractive features, it has not been as extensively studied in stepped wedge CRTs, with a few exceptions.[34,71,96,107] Specifically, Li et al.[34] and Li[71] used GEE to estimate the population-averaged intervention effect, coupled with the block exchangeable correlation structure (the correlation structure implied by model (13) and (26)), and the proportional decay structure (the correlation structure implied by model (27)). The GEE has been known to be prone to bias with a small number of clusters, both in terms of estimation of correlation parameters and variances,[108,109] therefore finite-sample corrections have been carefully studied and recommended.[34,71,96,107]

There are other aspects of the applications of models to stepped wedge designs that we have not reviewed. Above all, we have restricted the current article to models without cluster-level or individual-level covariates, although they could in principle be included in the analytical stage, especially when stratification or covariate-constrained randomization is carried out to minimize chance imbalance.[43,88,89,110] We have also only reviewed models applicable to stepped wedge trials with a single level of clustering, while Hemming et al.[25] and Teerenstra et al.[111] proposed extensions of the Hussey and Hughes model that accounted for multiple levels of clustering (e.g. patients nested in clinics and clinics nested in counties). Third, we have presented models assuming complete outcome information is available for all individuals and assumed away individual non-response. In practice, especially in closed-cohort designs, patient drop-out may occur given that the trial could last for a few years. When the drop-out mechanism can be considered as missing at random,[112] one may use inverse probability weighting or multilevel multiple imputation to reduce the bias due to missing outcomes. Turner et al.[113] recently studied the relative merits of these two mainstream missing data approaches for parallel CRTs, and it would be of interest to consider their extensions to stepped wedge CRTs.

Finally, in stepped wedge trials, reporting the values of various ICCs or variance components is also critically important to help inform the design of future studies with similar endpoints. Our experiences suggest that, although the correlations or variance components are essential input in virtually any sample size procedure derived from mixed-effects models in Section 3, only a very limited number of stepped wedge trials report such values. Accurate reporting of correlation estimates or variance components has been recommended in the CONSORT extension to stepped wedge designs,[23] and an example where the within-period and between-period correlations are reported can be found in Martin et al.[114] and Hemming et al.[47] We need more studies to report estimates of ICCs and variance components, in particular for the correlation decay and random intervention models, to facilitate the design of trials based on these more recent extensions of the Hussey and Hughes model.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported within the National Institutes of Health (NIH) Health Care Systems Research Collaboratory by the NIH Common Fund through cooperative agreement U24AT009676 from the Office of Strategic Coordination within the Office of the NIH Director and cooperative agreement UH3DA047003 from the

National Institute on Drug Abuse. This work is also supported by the administrative supplement 3-UH3-DA047003-02S2 from the NIH Office of Disease Prevention and PCORI contract ME-1507-31750. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health or PCORI. Karla Hemming is funded by the U.K. National Institute of Health Research Senior Research Fellowship SRF-2017-10-002. The authors are grateful to Dr. David Murray for participating in discussions leading up to this article as part of the work of the NIH Collaboratory Biostatistics and Study Design Working Group, for which Dr. Murray is the NIH representative. The authors thank the Editor and two anonymous reviewers for providing helpful comments.

References

1. Murray DM. Design and analysis of group-randomized trials. New York, NY: Oxford University Press, 1998.
2. Donner A and Klar N. Design and analysis of group-randomized trials in health research. New York, NY: Oxford University Press, 2000.
3. Hayes RJ and Moulton LH. Cluster randomised trials. Boca Raton, FL: Taylor & Francis Group, LLC, 2009.
4. Eldridge S and Kerry S. A practical guide to cluster randomised trials in health services research. Chichester, UK: John Wiley & Sons, 2012.
5. Turner EL, Li F, Gallis JA, et al. Review of recent methodological developments in group-randomized trials: part 1–design. Am J Public Health 2017; 107: 907–915. [PubMed: 28426295]
6. Turner EL, Prague M, Gallis JA, et al. Review of recent methodological developments in group-randomized trials: part 2–analysis. Am J Public Health 2017; 107: 1078–1086. [PubMed: 28520480]
7. Murray DM, Taljaard M, Turner EL, et al. Essential ingredients and innovations in the design and analysis of group-randomized trials. Ann Rev Public Health 2020; 41: 1–19. [PubMed: 31869281]
8. Hargreaves JR, Copas AJ, Beard E, et al. Five questions to consider before conducting a stepped wedge trial. Trials 2015; 16: 1–4. [PubMed: 25971836]
9. Hussey MA and Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemporary Clin Trials 2007; 28: 182–191.
10. Hughes JP. Stepped wedge design. In: Wiley encyclopedia of clinical trials. Hoboken, NJ: John Wiley & Sons, Inc, 2008.
11. Mdege ND, Man MS, Taylor CA, et al. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. J Clin Epidemiol 2011; 64: 936–948. [PubMed: 21411284]
12. Prost A, Binik A, Abubakar I, et al. Logistic, ethical, and political dimensions of stepped wedge trials: critical review and case studies. Trials 2015; 16: 351–361. [PubMed: 26278521]
13. Hemming K, Haines TP, Chilton PJ, et al. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ 2015; 350: 1–7.
14. Jarvik JG, Comstock BA, James KT, et al. Lumbar imaging with reporting of epidemiology (LIRE) – Protocol for a pragmatic cluster randomized trial. Contemporary Clin Trials 2015; 45: 157–163.
15. Zatzick DF, Russo J, Darnell D, et al. An effectiveness-implementation hybrid trial study protocol targeting posttraumatic stress disorder and comorbidity. Implement Sci 2016; 11: 1–16. [PubMed: 26727969]
16. Weinfurt KP, Hernandez AF, Coronado GD, et al. Pragmatic clinical trials embedded in healthcare systems: generalizable lessons from the NIH Collaboratory. BMC Med Res Methodol 2017; 17: 1–10. [PubMed: 28056835]
17. Martin J, Taljaard M, Girling A, et al. Systematic review finds major deficiencies in sample size methodology and reporting for stepped-wedge cluster randomised trials. BMJ Open 2016; 6: e010166.
18. Grayling MJ, Wason JMS and Mander AP. Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. Trials 2017; 18: 1–13. [PubMed: 28049491]
19. Campbell MK, Piaggio G, Elbourne DR, et al. Consort 2010 statement: extension to cluster randomised trials. BMJ 2012; 345: 1–21.

20. Brown CA and Lilford RJ. The stepped wedge trial design: a systematic review. BMC Med Res Methodol 2006; 6: 1–9. [PubMed: 16412232]
21. Davey C, Hargreaves J, Thompson JA, et al. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. Trials 2015; 16: 1–13. [PubMed: 25971836]
22. Barker D, McElduff P, D'Este C, et al. Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. BMC Med Res Methodol 2016; 16: 1–19. [PubMed: 26728979]
23. Hemming K, Taljaard M, McKenzie JE, et al. Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. BMJ 2018; 363: 1–26.
24. Hemming K, Taljaard M and Grimshaw J. Introducing the new CONSORT extension for stepped-wedge cluster randomised trials. Trials 2019; 20: 18–21. [PubMed: 30616653]
25. Hemming K, Lilford R and Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. Stat Med 2015; 34: 181–196. [PubMed: 25346484]
26. Kasza J and Forbes AB. Information content of cluster-period cells in stepped wedge trials. Biometrics 2019; 75: 144–152. [PubMed: 30051909]
27. Kasza J, Taljaard M and Forbes AB. Information content of stepped wedge designs when treatment effect heterogeneity and/or implementation periods are present. Stat Med 2019; 38: 4686–4701. [PubMed: 31321806]
28. Murray DM and Hannan PJ. Planning for the appropriate analysis in school-based drug-use prevention studies. J Consult Clin Psychol 1990; 58: 458–468. [PubMed: 2212183]
29. Feldman HA and Mckinlay SM. Cohort versus cross-sectional design in large field trials: precision, sample size and a unifying model. Stat Med 1994; 13: 61–78. [PubMed: 9061841]
30. Copas AJ, Lewis JJ, Thompson JA, et al. Designing a stepped wedge trial: Three main designs, carry-over effects and randomisation approaches. Trials 2015; 16: 1–12. [PubMed: 25971836]
31. Ji X, Fink G, Robyn PJ, et al. Randomization inference for stepped-wedge cluster-randomized trials: an application to community-based health insurance. Ann Appl Stat 2017; 11: 1–20.
32. Rubin DB. Estimating causal effects of treatment in randomized and nonrandomized studies. J Educ Psychol 1974; 66: 688–701.
33. Sitlani CM, Heagerty PJ, Blood EA, et al. Longitudinal structural mixed models for the analysis of surgical trials with noncompliance. Stat Med 2012; 31: 1738–1760. [PubMed: 22344923]
34. Li F, Turner EL and Preisser JS. Sample size determination for GEE analyses of stepped wedge cluster randomized trials. Biometrics 2018; 74: 1450–1458. [PubMed: 29921006]
35. Woertman W, De Hoop E, Moerbeek M, et al. Stepped wedge designs could reduce the required sample size in cluster randomized trials. J Clin Epidemiol 2013; 66: 752–758. [PubMed: 23523551]
36. Lawrie J, Carlin JB and Forbes AB. Optimal stepped wedge designs. Stat Probabil Lett 2015; 99: 210–214.
37. Thompson JA, Fielding K, Hargreaves J, et al. The optimal design of stepped wedge trials with equal allocation to sequences and a comparison to other trial designs. Clin Trials 2017; 14: 639–647. [PubMed: 28797179]
38. Girling AJ and Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. Stat Med 2016; 35: 2149–2166. [PubMed: 26748662]
39. Grayling MJ, Wason JMS and Mander AP. Group sequential designs for stepped-wedge cluster randomised trials. Clin Trial 2017; 14: 507–517.
40. Rhoda DA, Murray DM, Andridge RR, et al. Studies with staggered starts: Multiple baseline designs and group-randomized trials. Am J Public Health 2011; 101: 2164–2169. [PubMed: 21940928]
41. Hemming K and Girling AJ. The efficiency of stepped wedge vs. cluster randomized trials: stepped wedge studies do not always require a smaller sample size. J Clin Epidemiol 2013; 66: 1427–1428. [PubMed: 24035495]

42. Kristunas CA, Smith KL and Gray LJ. An imbalance in cluster sizes does not lead to notable loss of power in cross-sectional, stepped-wedge cluster randomised trials with a continuous outcome. Trials 2017; 18: 109–119. [PubMed: 28270224]
43. Martin JT, Hemming K and Girling A. The impact of varying cluster size in cross-sectional stepped-wedge cluster randomised trials. BMC Med Res Methodol 2019; 19: 1–11. [PubMed: 30611213]
44. Harrison LJ, Chen T and Wang R. Power calculation for cross-sectional stepped wedge cluster randomized trials with variable cluster sizes. Biometrics 2019.
45. Taljaard M, Teerenstra S, Ivers NM, et al. Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. Clin Trials 2016; 13: 459–463. [PubMed: 26940696]
46. Bond S. Theory of general balance applied to step wedge designs. Stat Med 2019; 38: 184–191. [PubMed: 30209821]
47. Hemming K, Taljaard M and Forbes A. Analysis of cluster randomised stepped wedge trials with repeated cross-sectional samples. Trials 2017; 18: 1–11. [PubMed: 28049491]
48. Diggle PJ, Heagerty PJ, Liang KY, et al. Analysis of Longitudinal Data. Oxford, UK: Oxford University Press, 2002.
49. Nickless A, Voysey M, Geddes J, et al. Mixed effects approach to the analysis of the stepped wedge cluster randomised trial Investigating the confounding effect of time through simulation. PLoS ONE 2018; 13: 1–22.
50. Golden MR, Kerani RP, Stenger M, et al. Uptake and population-level impact of expedited partner therapy (EPT) on Chlamydia trachomatis and Neisseria gonorrhoeae: the Washington State community-level randomized trial of EPT. Plos Med 2015; 12: 1–22.
51. Fitzmaurice GM, Laird NM and Ware JH. Applied longitudinal analysis. London, UK: John Wiley & Sons, 2012.
52. Grantham KL, Forbes AB, Heritier S et al. Time parameterizations in cluster randomized trial planning. Am Stat 2019; 0: 1–17.
53. Heo M, Kim N, Rinke ML, et al. Sample size determinations for stepped-wedge clinical trials from a three-level data hierarchy perspective. Stat Meth Med Res 2018; 27: 480–489.
54. Zhou X, Liao X and Spiegelman D. “Cross-sectional” stepped wedge designs always reduce the required sample size when there is no time effect. J Clin Epidemiol 2017; 83: 108–109. [PubMed: 28093263]
55. Hughes JP, Granston TS and Heagerty PJ. Current issues in the design and analysis of stepped wedge trials. Contemporary Clin Trials 2015; 45: 55–60.
56. Hooper R, Teerenstra S, de Hoop E, et al. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Stat Med 2016; 35: 4718–4728. [PubMed: 27350420]
57. Kasza J, Hemming K, Hooper R, et al. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. Stat Meth Med Res 2017, 10.1177/0962280217734981
58. Murray DM, Hannan PJ, Wolfinger RD, et al. Analysis of data from group-randomized trials with repeat observations on the same groups. Stat Med 1998; 17: 1581–1600. [PubMed: 9699231]
59. Teerenstra S, Lu B, Preisser JS, et al. Sample size considerations for GEE analyses of three-level cluster randomized trials. Biometrics 2010; 66: 1230–1237. [PubMed: 20070297]
60. Li F, Forbes AB, Turner EL, et al. Power and sample size requirements for GEE analyses of cluster randomized crossover trials. Stat Med 2019; 38: 636–649. [PubMed: 30298551]
61. Kasza J and Forbes AB. Inference for the treatment effect in multiple-period cluster randomised trials when random effect correlation structure is misspecified. Stat Meth Med Res 2018, 10.1177/0962280218797151 (accessed 11 December 2019).
62. Grantham KL, Kasza J, Heritier S, et al. Accounting for a decaying correlation structure in cluster randomized trials with continuous recruitment. Stat Med 2019, 38: 1918–1934. [PubMed: 30663132]
63. Hooper R and Copas A. Stepped wedge trials with continuous recruitment require new ways of thinking. J Clin Epidemiol 2019; 116: 161–166. [PubMed: 31272885]

64. Hemming K, Taljaard M and Forbes A. Modeling clustering and treatment effect heterogeneity in parallel and stepped-wedge cluster randomized trials. Stat Med 2018; 37: 883–898. [PubMed: 29315688]
65. Baio G, Copas A, Ambler G, et al. Sample size calculation for a stepped wedge trial. Trials 2015; 16: 354–368. [PubMed: 26282553]
66. Li F, Turner EL and Preisser JS. Optimal allocation of clusters in cohort stepped wedge designs. Stat Probabil Lett 2018; 137: 257–263.
67. Grayling MJ, Mander AP and Wason JM. Admissible multiarm stepped-wedge cluster randomized trial designs. Stat Med 2019; 38: 1103–1119. [PubMed: 30402914]
68. Lyons VH, Li L, Hughes JP, et al. Proposed variations of the stepped-wedge design can be used to accommodate multiple interventions. J Clin Epidemiol 2017; 86: 160–167. [PubMed: 28412466]
69. Girling AJ. Relative efficiency of unequal cluster sizes in stepped wedge and other trial designs under longitudinal or cross-sectional sampling. Stat Med 2018; 37: 4652–4664. [PubMed: 30209812]
70. van Breukelen GJP, Candel MJJM and Berger MPF. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials Gerard. Stat Med 2007; 26: 2589–2603. [PubMed: 17094074]
71. Li F. Design and analysis considerations for cohort stepped wedge cluster randomized trials with a decay correlation structure. Stat Med 2020; 39: 438–455. [PubMed: 31797438]
72. Liu A, Shih WJ and Gehan E. Sample size and power determination for clustered repeated measurements. Stat Med 2002; 21: 1787–1801. [PubMed: 12111912]
73. Lefkopoulou M, Moore D and Ryan L. The analysis of multiple correlated binary outcomes: application to rodent teratology experiments. J Am Stat Assoc 1989; 84: 810–815.
74. Kasza J, Hooper R, Copas A, et al. Sample size and power calculations for open cohort longitudinal cluster randomized trials. Stat Med 2020; (0): 1–13, 10.1002/sim.8519 [PubMed: 31663647]
75. Hemming K and Taljaard M. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. J Clin Epidemiol 2016; 69: 137–146. [PubMed: 26344808]
76. Zhou X, Liao X, Kunz LM, et al. A maximum likelihood approach to power calculations for stepped wedge designs of binary outcomes. Biostatistics 2018, 10.1093/biostatistics/kxy031
77. Thompson JA, Fielding KL, Davey C, et al. Bias and inference from misspecified mixed-effect models in stepped wedge trial analysis. Stat Med 2017; 36: 3670–3682. [PubMed: 28556355]
78. Jiang J. Linear and generalized linear mixed models and their applications. New York, NY: Springer, 2006.
79. McCulloch CE, Searle SR and Neuhaus JM. Generalized, linear, and mixed models. Hoboken, NJ: John Wiley & Sons, Inc., 2008. ISBN 9780471722076.
80. Pinheiro J and Bates D. Mixed-effects models in S and S-PLUS. New York, NY: Springer, 2009.
81. Wolfinger R and O'Connell M. Generalized linear mixed models: a pseudo-likelihood approach. J Stat Computat Simulat 1993; 48: 233–243.
82. Liu Q and Pierce DA. A note on Gauss-Hermite quadrature. Biometrika 1994; 81: 624–629.
83. Li P and Redden DT. Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. BMC Med Res Methodol 2015; 15: 38. [PubMed: 25899170]
84. Schluchter M and Elashoff JT. Small-sample adjustments to tests with unbalanced repeated measures assuming several covariance structures. J Stat Computat Simulat 1990; 19: 69–87.
85. Edgington E. Randomization tests. New York, NY: Marcel-Decker, 1987.
86. Gail MH, Mark SD, Carroll RJ, et al. On design considerations and randomization-based inference for community intervention trials. Stat Med 1996; 15: 1069–1092. [PubMed: 8804140]
87. Murray DM, Hannan PJ, Pals SP, et al. A comparison of permutation and mixed-model regression methods for the analysis of simulated data in the context of a group-randomized trial. Stat Med 2006; 25: 375–88. [PubMed: 16143991]
88. Li F, Lokhnygina Y, Murray DM, et al. An evaluation of constrained randomization for the design and analysis of group-randomized trials. Stat Med 2015; 35: 1565–1579. [PubMed: 26598212]

89. Li F, Turner EL, Heagerty PJ, et al. An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes. Stat Med 2017; 36: 3791–3806. [PubMed: 28786223]
90. Wang R and DeGruttola V. The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials. Stat Med 2017; 36: 2831–2843. [PubMed: 28464567]
91. Ren Y, Hughes JP and Heagerty PJ. A simulation study of statistical approaches to data analysis in the stepped wedge design. Stat Biosci 2019, 10.1007/s12561-019-09259-x (accessed 11 December 2019).
92. Thompson JA, Davey C, Fielding K, et al. Robust analysis of stepped wedge trials using cluster-level summaries within periods. Stat Med 2018; 37: 2487–2500. [PubMed: 29635789]
93. Matthews JN and Forbes AB. Stepped wedge designs: insights from a design of experiments perspective. Stat Med 2017; 36: 3772–3790. [PubMed: 28786236]
94. Kennedy-Shaffer L, de Gruttola V and Lipsitch M. Novel methods for the analysis of stepped wedge cluster randomized trials. Stat Med 2020; 39: 815–844. [PubMed: 31876979]
95. Hughes JP, Heagerty PJ, Xia F, et al. Robust inference for the stepped wedge design. Biometrics 2019, 10.1111/biom.13106 (accessed 11 December 2019).
96. Scott JM, DeCamp A, Juraska M, et al. Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. Stat Meth Med Res 2017; 26: 583–597.
97. Liang KY and Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986; 73: 13–22.
98. Durovni B, Saraceni V, Moulton LH, et al. Effect of improved tuberculosis screening and isoniazid preventive therapy on incidence of tuberculosis and death in patients with HIV in clinics in Rio de Janeiro, Brazil: a stepped wedge, cluster-randomised trial. Lancet Infectious Dis 2013; 13: 852–858.
99. Moulton LH, Golub JE, Durovni B, et al. Statistical design of THRio: A phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. ClinTrials 2007; 4: 190–199.
100. Zhan Z, deBock GH, Wiggers T, et al. The analysis of terminal endpoint events in stepped wedge designs. Stat Med 2016; 35: 4413–4426. [PubMed: 27311403]
101. Zhan Z, de Bock GH and van den Heuvel ER. Statistical methods for unidirectional switch designs: past, present, and future. Stat Meth Med Res 2018; 27: 2872–2882.
102. Preisser JS, Young ML, Zaccaro DJ, et al. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. Stat Med 2003; 22: 1235–1254. [PubMed: 12687653]
103. Heagerty PJ and Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. Biometrika 2001; 88: 973–985.
104. Li F and Harhay MO. Commentary: right truncation in cluster randomized trials can attenuate the power of a marginal analysis. Int J Epidemiol 2020. 10.1093/ije/dyaa037.
105. Heagerty PJ and Zeger SL. Marginalized multilevel models and likelihood inference. Stat Sci 2000; 15: 1–19.
106. Heagerty PJ. Marginally specified logistic-normal models for longitudinal binary data. Biometrics 1999; 55: 688–698. [PubMed: 11314994]
107. Ford WP. Improved standard error estimation for maintaining the validities of inference in small-sample cluster randomized trials and longitudinal studies. Technical report, PhD Thesis, University of Kentucky, 2018.
108. Lu B, Preisser JS, Qaqish BF, et al. A comparison of two bias-corrected covariance estimators for generalized estimating equations. Biometrics 2007; 63: 935–941. [PubMed: 17825023]
109. Preisser JS, Lu B and Qaqish BF. Finite sample adjustments in estimating equations and covariance estimators for intracluster correlations. Stat Med 2008; 27: 5764–5785. [PubMed: 18680122]
110. Lew RA, Miller CJ, Kim B, et al. A method to reduce imbalance for site-level randomized stepped wedge implementation trial designs. Implement Sci 2019; 14: 1–9. [PubMed: 30611302]

111. Teerenstra S, Taljaard M, Haenen A, et al. Sample size calculation for stepped-wedge cluster-randomized trials with more than two levels of clustering. Clin Trials 2019; 16: 225–236. [PubMed: 31018678]

112. Little R and Rubin D. Statistical analysis with missing data. Hoboken, NJ: Wiley, 2002.

113. Turner EL, Yao L, Li F and Prague M.. Properties and pitfalls of weighting as an alternative to multilevel multiple imputation in cluster randomized trials with missing binary outcomes under covariate-dependent missingness. Stat Meth Med Res 2020; 29: 1338–1353.

114. Martin J, Girling A, Nirantharakumar K, et al. Intra-cluster and inter-period correlation coefficients for cross-sectional cluster randomised controlled trials for type-2 diabetes in UK primary care. Trials 2016; 17: 402–413. [PubMed: 27524396]

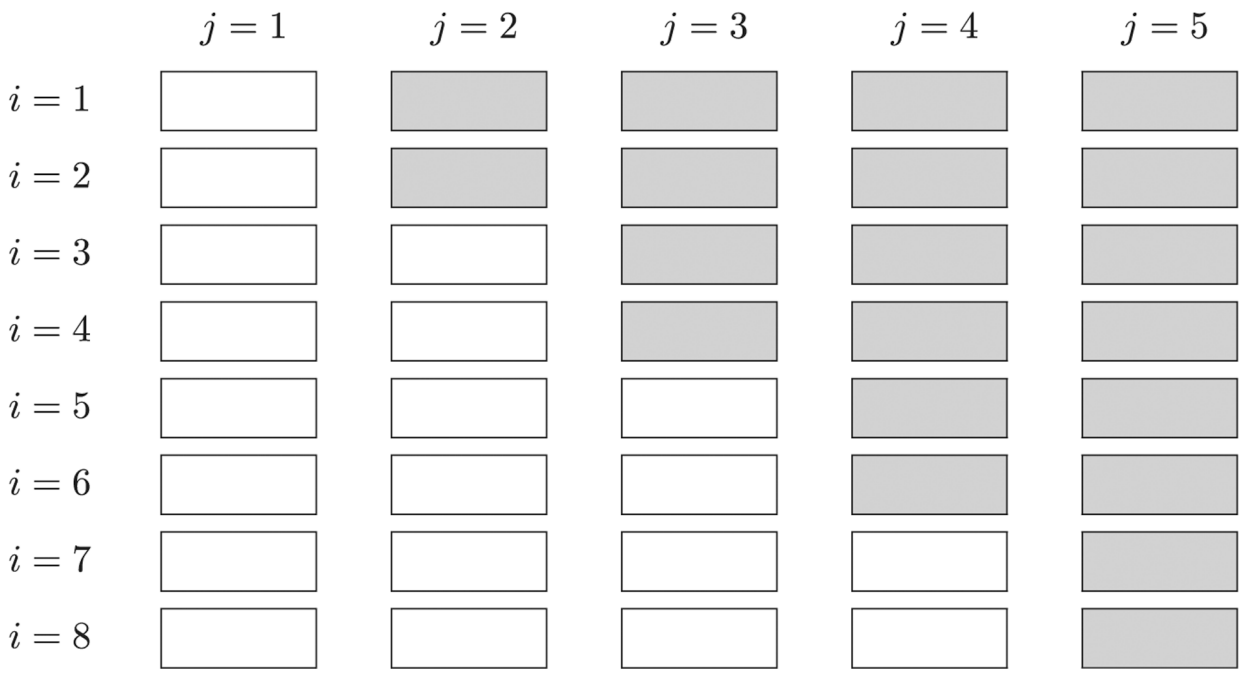

Figure 1.

A schematic illustration of a stepped wedge CRT with $I = 8$ clusters and $J = 5$ periods. Each white cell indicates a cluster-period under the control condition and each gray cell indicates a cluster-period under the intervention condition. There are in total $S = 4$ distinct intervention sequences.

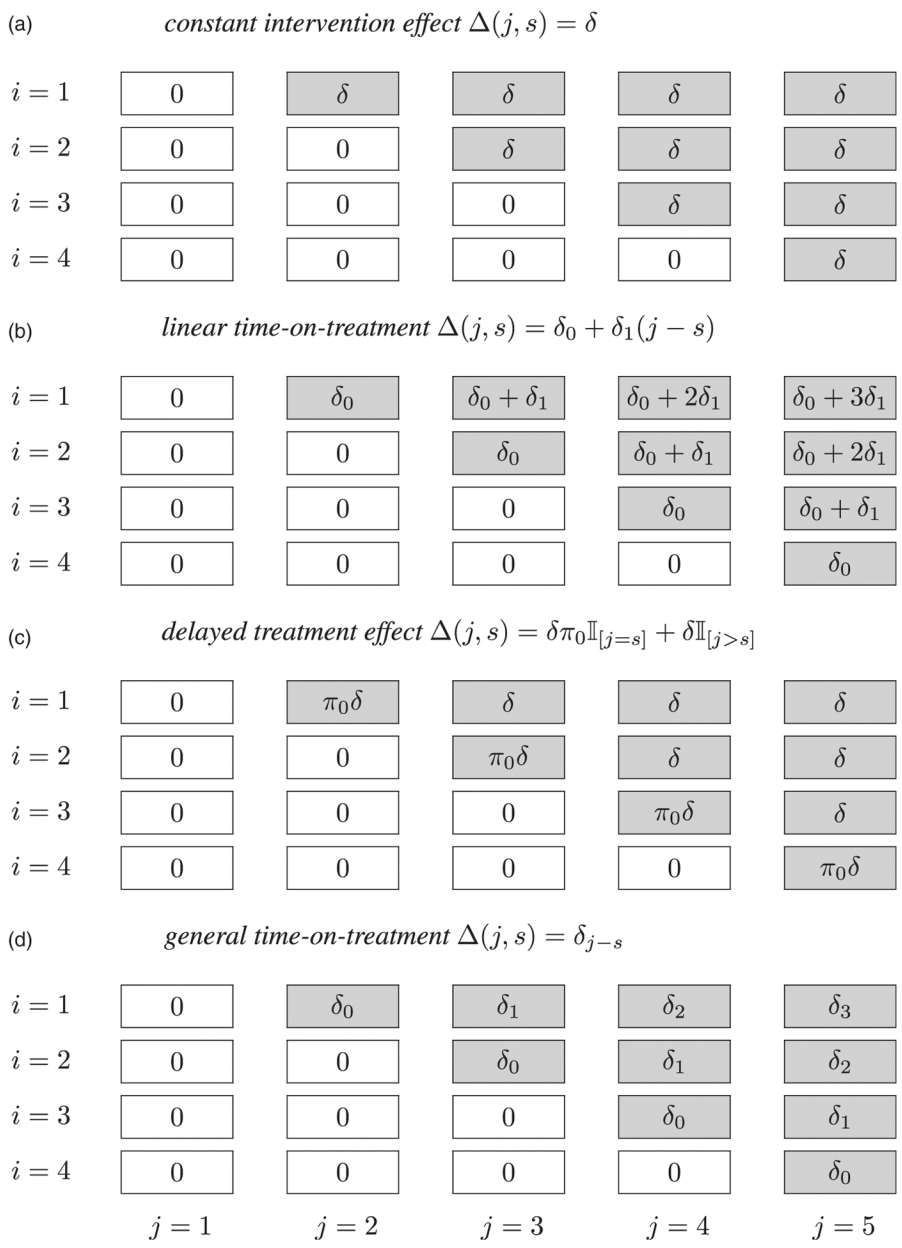

Figure 2.

Schematic illustrations of four intervention effect representations in a stepped wedge design with $I = 4$ clusters and $J = 5$ 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.

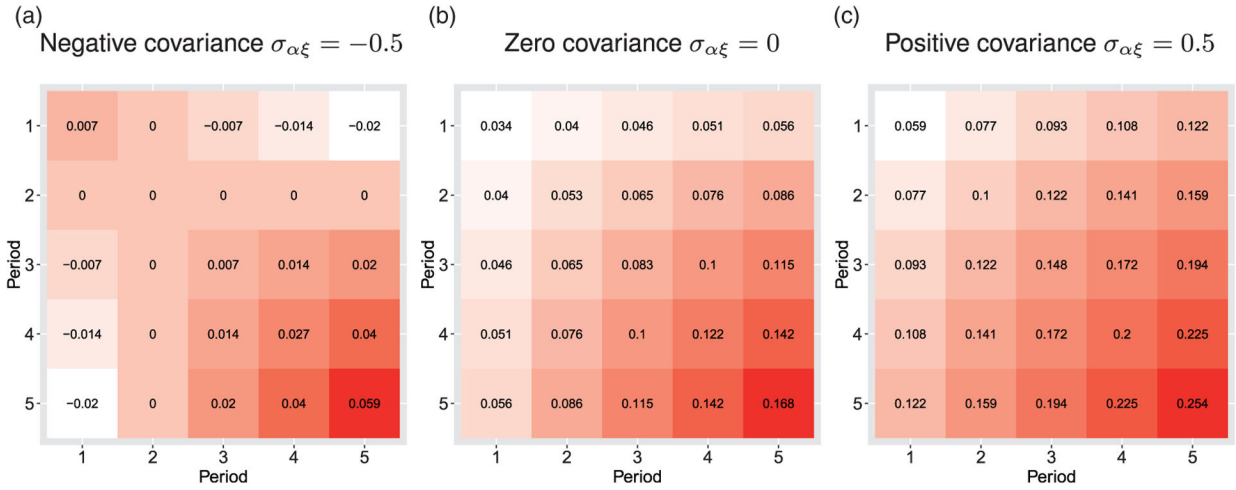

Figure 3.

Three examples of within-cluster correlation patterns implied by the random coefficient model. A trial with $J = 5$ is assumed throughout; the diagonal cells present the within-period ICC values, while the off-diagonal cells present the between-period ICC values. White color indicates a smaller ICC value while red color indicates a larger ICC value. The variance components parameters are assumed as $\sigma_\epsilon = 6$, $\tau_a = 1$, $\tau_\xi = 0.5$ and the covariance parameter (a) $\sigma_{a\xi} = -0.5$, (b) $\sigma_{a\xi} = 0$, (c) $\sigma_{a\xi} = 0.5$. (a) Negative covariance $\sigma_{a\xi} = -0.5$; (b) Zero covariance $\sigma_{a\xi} = 0$; (c) Positive covariance $\sigma_{a\xi} = 0.5$.

Table 1.

Example extensions to the Hussey and Hughes model for stepped wedge cluster randomized trials in cross-sectional and closed-cohort designs; all models assume a continuous outcome and an identity link function.

Design	Extension	Feature	Example references
Cross-sectional	Nested Exchangeable*	Distinguish between within-period and between-period ICCs	Hooper et al.;[56] Girling and Hemming[38]
	Exponential Decay*	Allow the between-period ICC to decay at an exponential rate over time	Kasza et al.[57] Kasza and Forbes[61]
	Random Intervention	Include random cluster-specific intervention effects, and ICC depends on intervention status	Hughes et al.[55] Hemming et al.[47]
	Random Coefficient	Include random cluster-specific time slopes; ICC tends to be an increasing function of distance in time	Murray et al.[58]
Closed-cohort	Basic	Include cluster-level and subject-level random effects to separate between-individual ICC and within-individual ICC	Baio et al.[65]
	Block Exchangeable*	Include three random effects to distinguish between within-period ICC, between-period ICC, and within-individual ICC	Hooper et al.[56] Girling and Hemming[38]
	Proportional Decay*	Allow the between-period ICC and within-individual ICC to decay over time at the same exponential rate	Li[60]
	Random Intervention	Include random cluster-specific intervention effects, and ICC depends on intervention status	Kasza et al.[27]

Note: The choice of terminology with the '*' symbol is based on the following. The nested exchangeable correlation model was defined in Teerenstra et al.[59] and Li et al.[60] in the context of three-level CRTs and crossover CRTs. Li et al.[34] introduced the block exchangeable correlation model for closed-cohort design and pointed out the nested exchangeable correlation model is a special case. The exponential decay correlation model is proposed in Kasza et al. and Kasza and Forbes.[57,61] The proportional decay correlation model is introduced in Li[60] and dates back to the earlier work of Liu et al.[72] in the context of longitudinal parallel CRTs.

Table 2.

Illustration of the non-decaying (exchangeable) and decaying within-cluster correlation structure implied by the random-effects model in cross-sectional, closed-cohort, and open-cohort designs.

	Nested exchangeable structure	Exponential decay structure
Cross-sectional (Section 3.4)		
	Block exchangeable structure	Proportional decay structure
Closed-cohort (Section 3.5)		
	Blended exchangeable structure	Blended correlation decay structure
Open-cohort (Section 3.6)		

Note: In each correlation matrix, each block represents the correlation structure in a given cluster-period or between two cluster-periods, and the total number of periods is $T = 3$. The cluster-period sizes are assumed to be equal ($N_{ij} = 2$). In the open-cohort design, we assume only one individual is followed through all periods, and a new individual will be supplemented in each period. Each correlation matrix is defined for the vector of observations collected across all periods in the same cluster.