Received: 12 February 2021 | Revised: 30 November 2021 | Accepted: 22 December 2021

DOI: 10.1002/sim.9317

RESEARCH ARTICLE

Statistics in Medicine WILEY

Response adaptive intervention allocation in stepped-wedge cluster randomized trials

Michael J. Grayling[1] | **James M. S. Wason**[1] | **Sofía S. Villar**[2]

[1]Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK

[2]MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK

Correspondence
Michael J. Grayling, Population Health Sciences Institute, Ridley Building 1, Queen Victoria Road, Newcastle upon Tyne NE1 7RU, UK.
Email: michael.grayling@newcastle.ac.uk

Funding information
Medical Research Council, Grant/Award Numbers: MC_UU_00002/15, MC_UU_00002/6

Background: Stepped-wedge cluster randomized trial (SW-CRT) designs are often used when there is a desire to provide an intervention to all enrolled clusters, because of a belief that it will be effective. However, given there should be equipoise at trial commencement, there has been discussion around whether a pre-trial decision to provide the intervention to all clusters is appropriate. In pharmaceutical drug development, a solution to a similar desire to provide more patients with an effective treatment is to use a response adaptive (RA) design.

Methods: We introduce a way in which RA design could be incorporated in an SW-CRT, permitting modification of the intervention allocation during the trial. The proposed framework explicitly permits a balance to be sought between power and patient benefit considerations. A simulation study evaluates the methodology.

Results: In one scenario, for one particular RA design, the proportion of cluster-periods spent in the intervention condition was observed to increase from 32.2% to 67.9% as the intervention effect was increased. A cost of this was a 6.2% power drop compared to a design that maximized power by fixing the proportion of time in the intervention condition at 45.0%, regardless of the intervention effect.

Conclusions: An RA approach may be most applicable to settings for which the intervention has substantial individual or societal benefit considerations, potentially in combination with notable safety concerns. In such a setting, the proposed methodology may routinely provide the desired adaptability of the roll-out speed, with only a small cost to the study's power.

KEYWORDS
adaptive design, clinical trial, interim analysis, multi-stage, sequential allocation

1 | INTRODUCTION

Stepped-wedge cluster randomized trials (SW-CRTs) roll an intervention out over several time periods, with all clusters typically ending the trial in the intervention condition.[1] SW-CRTs have been favored for several reasons, including that sequential roll-out may assist with logistical constraints. However, SW-CRTs have not been without criticism. In

particular, there has been much discussion of another reason commonly given for using an SW-CRT: a strong belief that the intervention will do more good than harm, which implies its allocation to all clusters is advantageous. Kotz et al[2] argued this makes SW-CRTs troubling, because a decision to provide the intervention to all clusters should not be made when its effectiveness remains unproven.

It has been pointed out, however, that the design has typically been used when the intervention has been "shown to be effective in more controlled...settings."[3] This raises a further important issue though around whether there can be equipoise in an SW-CRT if there is a strong belief, perhaps emboldened by previous studies, that the intervention will be effective. Given that it has been argued "genuine uncertainty...about the preferred treatment" is a prerequisite for conducting a randomized trial,[4] this calls in to question when an SW-CRT could be conducted.

Prost et al[5] suggested a constructive solution to this question is to consider whether the evidence in favor of the intervention is sufficient to suggest equipoise is truly disturbed. While there may be a consensus that the intervention will be beneficial, there may still be true uncertainty about its effectiveness in a given context. Thus equipoise may still apply. Ultimately, it has been argued SW-CRTs in which equipoise is disturbed should not be undertaken.[6] Given equipoise, though, we return to the scenario above where there may then be concern around a decision to provide the intervention to all clusters. This could be particularly true of closed-cohort SW-CRT designs, where all participants would then receive the intervention, or when the intervention is associated with substantial safety considerations.

In drug development, response adaptive (RA) design has been suggested as way to address deviations from equipoise that could arise from data collected during a trial. To introduce RA design, consider a parallel two-arm individually randomized trial. With RA design, the trial incorporates interim analyses at which the allocation ratio can be modified, with the standard being to increase allocation to the best performing treatment. The number of patients expected to receive the best treatment is then increased. If the endpoint used to evaluate the treatments is related to patient benefit, then on average this provides an advantage to patients enrolled on the trial compared to fixed 1:1 randomization. Importantly, any decision to increase allocation to a particular treatment is made using concurrent study data; unlike in an SW-CRT, which makes this decision pre-trial. For an overview of RA trial design, see one of several recent monographs[7-9] or the recent review by Robertson et al.[10]

It is interesting therefore to ask whether and/or how a conventional SW-CRT could be modified to incorporate RA intervention allocation, enabling an intervention to be provided to more participants when it is effective, but its roll-out slowed, or stopped, when ineffective. In this article, we describe a flexible framework for modifying an SW-CRT allocation matrix at a series of interim analyses. To evaluate the framework, we present the results of an extensive simulation study. To conclude, we describe several practical issues associated with utilizing an RA SW-CRT design and discuss when it may be useful.

2 | METHODS

2.1 | Design setting

We suppose an SW-CRT will be used to compare an intervention to a control; aiming to contrast an RA SW-CRT with its conventional fixed-sample analog. We suppose this fixed-sample SW-CRT has been designed, omitting discussion on how this can be achieved as it is has been covered elsewhere.[11-14] Thus, we assume the number of clusters $C > 1$, time periods $P > 1$, and measurements $m > 1$ per cluster-period, have been specified. We consider designs where the m measurements from each cluster-period are from the same (closed-cohort design) or from different (cross-sectional design) participants. We comment on application to open-cohort designs in Section 4. We also suppose a treatment allocation matrix has been nominated, $X = \{X_{ij}\}$, $i = 1, \dots, C$, $j = 1, \dots, P$, with $X_{ij} = 1$ implying cluster i receives the intervention in time period j, and $X_{ij} = 0$ otherwise. We refer to this as the initially planned allocation matrix.

We denote the responses to be accrued up to time period p, $1 \leq p \leq P$, by $\boldsymbol{Y}_p$. Specifically, suppose measurement $k = 1, \dots, m$ from cluster $i = 1, \dots, C$ in period $j = 1, \dots, P$ is denoted by Y_{ijk}. Then

$$\begin{aligned}\boldsymbol{Y}_p = (&Y_{111}, \dots, Y_{11m}, Y_{121}, \dots, Y_{12m}, \dots, Y_{1p1}, \dots, Y_{1pm}, \dots,\\ &Y_{C11}, \dots, Y_{C1m}, Y_{C21}, \dots, Y_{C2m}, \dots, Y_{Cp1}, \dots, Y_{Cpm})^\top.\end{aligned}$$

We suppose that at the design stage a particular linear mixed model has been designated for data analysis, and thus it has been assumed $\boldsymbol{Y}_p \sim N(D_{p|X}\boldsymbol{\theta}, \Sigma_{p|X})$, for known nonsingular covariance matrix $\Sigma_{p|X}$, design matrix $D_{p|X}$, and fixed effects $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)^\top$. As we see in our simulation study later, $\boldsymbol{\theta}$ would typically be expected to include an intercept term,

factors to adjust for time effects, and an effect for the intervention relative to the control. Furthermore, we note that a large number of possible analysis models have been proposed for SW-CRTs; see Li et al[15] for an overview of many of these. To emphasize, our designations above allow for any of these that work within a linear mixed model framework, including those assuming a decaying correlation structure. Finally, we note that we explicitly state the dependence of $\Sigma_{p|X}$ and $D_{p|X}$ upon X since X will later be treated as a variable.

We assume the goal is to make inference on the intervention's effect relative to the control. We suppose this is estimated through θ_q and refer to this from here as θ for brevity. We assume that the one-sided hypothesis $H_0 : \theta \leq 0$ will be tested, with a type-I error-rate of $\alpha \in (0, 1)$ desired when $\theta = 0$. Later, we also compare designs in terms of their power when $\theta = \delta$, for specified $\delta > 0$, with the target to achieve power of $1 - \beta \in (0, 1)$.

Note the generalized least squares estimate of $\boldsymbol{\theta}$ after time period p is $\hat{\boldsymbol{\theta}}_{p|X} = (D_{p|X}^\top \Sigma_{p|X}^{-1} D_{p|X})^{-1} D_{p|X}^\top \Sigma_{p|X}^{-1} \boldsymbol{Y}_p$. Extracting the last element, $\hat{\theta}_{p|X}$, the following Wald test statistic can be calculated

$$Z_{p|X} = \frac{\hat{\theta}_{p|X}}{\{\text{Var}(\hat{\theta}_{p|X})\}^{1/2}} = \frac{\hat{\theta}_{p|X}}{[\{(D_{p|X}^\top \Sigma_{p|X}^{-1} D_{p|X})^{-1}\}_{qq}]^{1/2}} := \hat{\theta}_{p|X} I_{p|X}^{1/2}.$$

A conventional SW-CRT would proceed by enrolling C clusters, accruing m measurements per cluster in time periods $1, \ldots, P$, and allocating treatments according to X. Its final analysis could be conducted by assessing whether $Z_{P|X} > \Phi^{-1}(1 - \alpha)$. Our aim, as discussed, is to describe methodology through which X may be altered mid-trial. Note that it is only X we modify; to provide a fairer comparison to the corresponding conventional fixed-sample SW-CRT we assume the initial values of m, C, and P are not altered at the interim analyses.

2.2 | Response adaptive stepped-wedge cluster randomized trials

First, a set of integers $\{p_1, \ldots, p_L\}$, with $1 \leq p_{l_1} < p_{l_2} \leq P - 1$ for $1 \leq l_1 < l_2 \leq L$, are specified. Then, L interim analyses at which the allocation matrix may be altered are conducted; after time periods $p \in \{p_1, \ldots, p_L\}$. Accordingly, we denote by $X_p = \{X_{pij}\}$, $1 \leq p \leq P$, the matrix containing the allocations used in time periods $1, \ldots, p$ and those planned for time periods $p + 1, \ldots, P$. We set $X_1 = \cdots = X_{p_1} = X$, using the initially planned allocation matrix X.

Next, sets $\mathcal{X}_{X_p}$ are specified, giving the possible allocation matrices to be chosen from at the analysis following time period p, dependent on the value of X_p. That is, $X_{p+1} \in \mathcal{X}_{X_p}$. Arbitrary restrictions can be placed on the $\mathcal{X}_{X_p}$ as are desired. In all instances though, $\mathcal{X}_{X_p}$ must consist of $C \times P$ binary matrices whose elements in columns $1, \ldots, p$ match those from X_p (as past allocations cannot be changed), and whose elements are such that if $X_{pip} = 1$ then $X_{pip+1} = \cdots = X_{piP} = 1$ (as clusters cannot switch back to the control). Thus, formally, we must always have that

$$\mathcal{X}_{X_p} \subseteq \mathcal{M}_{X_p} = \big[M = \{M_{ij}\} \in \mathbb{M}_{CP}(\{0,1\}) : \forall (i,j) \in \{1, \ldots, C\} \times \{1, \ldots, p\}\ M_{ij} = X_{pij},$$
$$\forall i \in \{1, \ldots, C\}\ X_{pip} = 1 \Rightarrow M_{ip+1} = \cdots = M_{iP} = 1\big].$$

Note that $X_p \in \mathcal{M}_{X_p}$, so it is always possible to ensure $\mathcal{X}_{X_p} \neq \emptyset$.

To illustrate the possible specification of $\mathcal{X}_{X_p}$ more clearly, consider an example with $C = P = 4$ and an interim analysis conducted after time period 2. Suppose that

$$X_1 = X_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Placing no restrictions on $\mathcal{X}_{X_2}$ beyond those which are always required (ie, $\mathcal{X}_{X_2} = \mathcal{M}_{X_2}$)

$$\mathcal{X}_{X_2} = \left\{ \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \right\}.$$

If we wished to ensure that all clusters receive the intervention by the trial's completion, we would modify the above to

$$\mathcal{X}_{X_2} = \left\{ \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \right\}.$$

Note that we order the sequences in the allocation matrices such that a nonincreasing proportion of time is spent in the intervention condition. This removes any degeneracy in the choice of possible allocation matrices.

The remaining component required is a function $s(\cdot)$ such that $s(X_{p+1})$ provides a *score* associated with a choice of $X_{p+1} \in \mathcal{X}_{X_p}$. Our approach is then to set

$$X_{p+1} = \underset{X \in \mathcal{X}_{X_p}}{\text{argmax}} s(X).$$

In practice, $s(\cdot)$ could be defined in any way that reasonably evaluates the suitability of X_{p+1}. Our approach is to specify $s(\cdot)$ to permit a balance to be sought between desires to (i) maximize allocation to the most effective arm and (ii) maximize power. We set

$$s(X_{p+1}) = w \frac{I_{P|X_{p+1}}}{\max_{X' \in \mathcal{X}_{X_p}} I_{P|X'}} + (1-w) \frac{b(X_{p+1})}{\max_{X'' \in \mathcal{X}_{X_p}} b(X'')}, \quad w \in [0,1].$$

Here, $b(\cdot)$ assesses the performance of X_{p+1} in terms of whether it allocates clusters to the most effective arm (ie, it monitors *patient benefit* considerations). Note the term involving the information levels $I_{P|X_{p+1}}$ evaluates X_{p+1} in terms of the power it likely provides. Thus, w is an explicit weight balancing (i) and (ii) above. Note the two factors are rescaled because they exist on different scales. Furthermore, $w \in \{0, 1\}$ should usually be avoided as a means to breaking ties between designs with identical values for $I_{P|X_{p+1}}$ or $b(X_{p+1})$.

The above formulation has been used previously in RA design, for example, for sequence specification in individually randomized crossover trials.[16] Nonetheless, specifying $b(\cdot)$ is complex for SW-CRTs because allocation is to be adapted for clusters already in the trial. In practice, there may be good reason to make $b(\cdot)$ a complex function that, for example, incorporates penalties for the speed or cost of the intervention roll-out if its availability is limited. Here, we consider a function of arguably more general utility, using only current evidence of effectiveness to guide allocation.

It is logical to insist that as $Z_{p|X_p}$ increases, $b(\cdot)$ should score designs switching a larger number of clusters to the intervention more highly. It is thus desirable to ensure that when $Z_{p|X_p} \to \infty$ the allocation matrix that switches all clusters to the intervention immediately is recommended. Similarly, as $Z_{p|X_p} \to -\infty$, the matrix that switches no additional clusters to the intervention should be recommended. Many functions will have these properties. In the Supplementary Material, we describe a form for $b(\cdot)$ that could be useful if the desire is to only alter the design for extreme intervention effects. To more clearly describe the benefits of RA SW-CRTs, we focus here on a probabilistic form for $b(\cdot)$ that can recommend a broader range of designs, taking

$$b(X_{p+1}) = \mathbb{P}\left(S = \sum_{i=1}^{C} \mathbb{I}(X_{pip} = 0) \sum_{j=p+1}^{P} X_{p+1ij} \right),$$
$$S \sim \text{Bin}\left[(P-p)\left(C - \sum_{i=1}^{C} X_{pip} \right), \Phi\left\{ \frac{Z_{p|X_p} - \eta}{\gamma(1 - p/P)} \right\} \right].$$

To understand this formulae, note that $C - \sum_{i=1}^{C} X_{pip}$ is the number of clusters in the control condition after time period p. Thus $(P-p)(C - \sum_{i=1}^{C} X_{pip})$ is the number of cluster-periods for which the roll-out could be modified. Similarly, $\sum_{i=1}^{C} \mathbb{I}(X_{pip} = 0) \sum_{j=p+1}^{P} X_{p+1ij}$ is the number of the modifiable cluster-periods matrix X_{p+1} spends in the intervention condition. The form for the success probability, $\Phi[(Z_{p|X_p} - \eta)/\{\gamma(1 - p/P)\}]$, is chosen to provide the sought after qualities of the function $b(\cdot)$ and to provide flexibility such that a search can be conducted for an RA design that has desirable

operating characteristics. First, $\Phi(\cdot)$ is used to map the continuous Wald test statistic to $[0, 1]$, enabling its value to serve as a probability that controls the speed of the roll-out conditional on the interim effectiveness. In addition, $\eta \in \mathbb{R}$ is a parameter that can be chosen to influence the value of $b(X_{p+1})$; larger values of η result in smaller values of $\Phi(\cdot)$, favoring designs slowing the roll-out of the intervention. Similarly, parameter $\gamma > 0$ influences how extreme the values of $\Phi(\cdot)$ are, with larger γ shifting the success probability toward 0.5, which should translate to a more balanced intervention roll-out. Finally, the denominator includes the factor $1 - p/P$ to scale the success probabilities, allowing them to be more extreme for larger p (ie, when more information is available to base the decision upon). This form for $b(\cdot)$ is also discussed further in the Supplementary Material.

The above fully describes the proposed framework for incorporating RA intervention allocation in an SW-CRT, with the final analysis conducted here analogously to a conventional SW-CRT by assessing whether $Z_{P|X_P} > \Phi^{-1}(1-\alpha)$. We comment in the discussion on potential alternatives to this rejection rule that may be useful in practice. An algorithm on the conduct of an RA SW-CRT is provided in the Supplementary Material.

2.3 | Simulation study

We assess the performance of the proposed framework through an extensive simulation study that considers three trial design scenarios (TDSs). Each TDS assumes the following model for data generation and analysis[14,17,18]

$$Y_{ijk} = \beta_j + \theta X_{ij} + c_i + \pi_{ij} + s_{ik} + \epsilon_{ijk}.$$

Here, Y_{ijk} is the response from individual $k = 1, \dots, m$, in cluster $i = 1, \dots, C$, in period $j = 1, \dots, P$, β_j is a fixed effect for time period j, $c_i \sim N(0, \sigma_c^2)$ is a random cluster effect, $\pi_{ij} \sim N(0, \sigma_\pi^2)$ is a random cluster-period effect, $s_{ik} \sim N(0, \sigma_s^2)$ is a random individual effect, and $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$ is the residual error. Thus, in this case, $\boldsymbol{\theta} = (\beta_1, \dots, \beta_P, \theta)$. Primary results for TDS1 are presented here, where TDS2 is also used to provide a simple illustration of the method's use. Additional findings for TDS1 are given in the Supplementary Material, where the results for TDS2 and TDS3 are also presented.

TDS1 is a cross-sectional SW-CRT ($\sigma_s^2 = 0$) that has been considered previously.[19-21] It is based on the average characteristics of SW-CRTs according to Grayling et al,[22] setting $C = 20$ and $P = 9$. In X, three clusters switch to the intervention in each of time periods 2 to 5, and two clusters switch in each of time periods 6 to 9. To give a larger value for the intra-cluster correlation than TDS2, it has $\sigma_c^2 = 1/9$ and $\sigma_e^2 = 1$. Additionally, $\alpha = 0.05$, $\beta = 0.2$, and $\delta = 0.24$. Using the sample size calculation method from Hussey and Hughes[11] (ie, $\sigma_\pi^2 = 0$), $m = 7$ is chosen. For the RA designs, we consider conducting a single interim analyses after time period $\{3\}$, $\{4\}$, or $\{5\}$, and conducting two interim analyses after time periods $\{3, 6\}$.

TDS2 is a cross-sectional SW-CRT ($\sigma_s^2 = 0$) based upon the trial presented in Bashour et al;[23] a study assessing the effect of training doctors in communication skills on women's satisfaction with doctor-woman relationship during labor and delivery. In this case, $C = 4$ and $P = 5$, with X switching one cluster to the intervention in each of time periods 2 to 5. The final analysis estimated that $\sigma_c^2 = 0.02$ and $\sigma_e^2 = 0.51$. We use these values in all simulations. Following the approach of Hussey and Hughes[11] ($\sigma_\pi^2 = 0$), for these variance components the trial would have required 70 patients per cluster-period for its desired type-I error-rate of 5% and its desired type-II error-rate of 10% when $\theta = 0.2$. Thus, we fix $m = 70$, $\alpha = 0.05$, $\beta = 0.1$, and $\delta = 0.2$. We consider conducting interim analyses after time periods $\{3\}$ and $\{2, 3, 4\}$.

TDS3 is a closed-cohort SW-CRT scenario, based on the "Girls on the Go!" program to improve self-esteem in young women in Australia,[24] following the calculation in Hooper et al.[14] Thus, we consider a case where $C = 12$ and $P = 4$, with X switching four clusters to the intervention in time periods 2 to 4. Measurements from $m = 10$ individuals are assumed to be collected in each cluster and the primary outcome measure (Rosenberg Self-esteem Scale) is assumed to have $\sigma_c^2 = 7.425$, $\sigma_\pi^2 = 0.825$, $\sigma_s^2 = 11.725$, and $\sigma_\epsilon^2 = 5.025$. The conventional design achieves $\beta = 0.2$ for $\delta = 2$ with $\alpha = 0.025$. We consider conducting interim analyses after time periods $\{2\}$ and $\{2, 3\}$.

In all three TDSs, we consider performance when $\theta \in \{-\delta, -0.5\delta, \dots, 2\delta\}$, $w \in \{1/1000, 1/4, 1/3, 1/2, 2/3, 3/4, 999/1000\}$, $\eta \in \{-1, 0, 1, 2, 3, 4\}$, and $\gamma \in \{1, 2.5, 5\}$. These values were chosen by factoring in what was computationally feasible and through an initial grid search to identify a range for the parameters beyond which the operating characteristics did not appear to vary substantially. In all cases, we place no restrictions on the $\mathcal{X}_{X_p}$ beyond those required (ie, we always set $\mathcal{X}_{X_p} = \mathcal{M}_{X_p}$).

For each combination of design parameters, 100 000 replicate simulations are used to estimate several key quantities. These are

- The empirical rejection probability (ERP) for H_0, with the values for $\theta = 0$ and $\theta = \delta$ referred to as the empirical type-I error-rate and power.
- The empirical average, standard deviation, and probability mass function of the proportion of cluster-periods spent in the intervention condition. We refer to the average and standard deviations of this quantity for brevity as the EACP and ESDCP, respectively. The EACP and ESDCP together evaluate patient benefit, for example, larger (smaller) values of the EACP are desired for larger (smaller) treatment effects, while we would likely always prefer small ESDCP. Note that when evaluating these quantities, one must account for the fact that the choice of X_{p_1} imparts particular minimal and maximal values for the time spendable in the intervention condition; these will be indicated on all relevant plots.
- The empirical average value of X_P, denoted $\overline{X}_P$.
- The empirical bias (EB) and root-mean-square error (ERMSE) of the final point estimate of θ, $\hat{\theta}_{p|X}$. Previous work for individually randomized trials has explored the negative impacts of RA design on point estimation when it is performed in a manner that does not take in to account the interim analyses.[25]

Code to reproduce our results is available from https://github.com/mjg211/article_code.

3 | RESULTS

3.1 | Illustrative description: Trial design scenario 2

To make the proposed methodology more tangible, we illustrate its application to TDS2, where the low number of clusters ($C = 4$) and time periods ($P = 5$) makes the possible allocation matrices limited. As discussed, Bashour et al[23] utilized the following allocation matrix

$$X = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Suppose that there was concern around use of this allocation matrix, such that RA design was to be utilized. In practice, this could happen for one of numerous reasons, though principally it may often be because investigators wish to provide a larger number of participants with the intervention if it is effective (this is often especially true for disease settings in which the condition under investigation can be particularly harmful), or because downsides (eg, cost or harm/safety concerns) mean that they would want to limit roll-out if the intervention was ineffective. As discussed, the first step is then to specify the time periods after which interim analyses will be conducted. As a basic example, we suppose that this is after period $\{3\}$, such that $X_1 = X_2 = X_3 = X$.

Thus, the RA trial would proceed by conducting periods 1 to 3 and then computing $Z_{3|X_3}$ using the interim data. Placing no constraints on $\mathcal{X}_{X_3}$ beyond those required, we would have

$$X_4 = X_5 \in \mathcal{X}_{X_3} = \{M_1, M_2, \ldots, M_6\},$$

$$M_1 = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \; M_2 = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \; M_3 = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$M_4 = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \; M_5 = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \; M_6 = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

For the assumed Hussey and Hughes model and variance parameters ($\sigma_s^2 = 0$, $\sigma_\pi^2 = 0$, $\sigma_c^2 = 0.02$, $\sigma_e^2 = 0.51$), it can be shown that

$$I_{5|M_1} \approx 188.5,\ I_{5|M_2} \approx 224.5,\ I_{5|M_3} \approx 204.7,\ I_{5|M_4} \approx 222.2,\ I_{5|M_5} \approx 215.2,\ I_{5|M_6} \approx 169.8.$$

To determine the choice of the interim specified allocation matrix, X_4, we then must also calculate the values of the $b(\cdot)$. Suppose that $\eta = 0$ and $\gamma = 2.5$, and as an example assume $Z_{3|X_3} = 1$. Using our definition of S, we have

$$S \sim \text{Bin}\left[(5-3)\left(4 - \sum_{i=1}^{4} X_{3i3}\right), \Phi\left\{\frac{1-0}{2.5(1-3/5)}\right\}\right].$$

That is, $S \sim \text{Bin}\{4, \Phi(1)\}$. Using our definition of $b(\cdot)$, this gives

$$b(M_1) = \mathbb{P}(S=0) \approx 0.001,\ b(M_2) = \mathbb{P}(S=1) \approx 0.013,\ b(M_3) = \mathbb{P}(S=2) \approx 0.107,$$
$$b(M_4) = \mathbb{P}(S=2) \approx 0.107,\ b(M_5) = \mathbb{P}(S=3) \approx 0.378,\ b(M_6) = \mathbb{P}(S=4) \approx 0.501.$$

Finally, supposing that $w = 0.5$, we can use the above to show that

$$s(M_1) \approx 0.420,\ s(M_2) \approx 0.513,\ s(M_3) \approx 0.563,\ s(M_4) \approx 0.601,\ s(M_5) \approx 0.856,\ s(M_6) \approx 0.878.$$

Thus, M_6 is the matrix that maximizes $s(\cdot)$, and so we set $X_4 = X_5 = M_6$ and conduct periods 4 to 5 of the trial using its roll-out. At the end of the study, we have that the proportion of cluster-periods spent in the intervention condition is 55%, while the value of $Z_{5|X_5}$ determines whether H_0 is rejected.

This is of course description of one possible realization of carrying out an RA trial. Our key concerns revolve around what the expected performance of this approach would look like, in terms of our metrics the ERP, EACP, ESDCP, EB, and ERMSE. We present these evaluations in the Supplementary Materials, where we also consider conducting interim analyses after time periods $\{2, 3, 4\}$.

3.2 | Trial design scenario 1

Switching to TDS1, we commence our investigation of the expected performance of RA procedures. Note that additional results for TDS1 are given in the Supplementary Materials.

3.2.1 | Operating characteristics for $\eta = 0$ and $\gamma = 2.5$

Figure 1 displays the ERP, EACP, ESDCP, EB, and ERMSE of several RA SW-CRT designs as a function of w and θ when $\{p_1, \ldots, p_L\} = \{3, 6\}$. As an example, results for $\eta = 0$ and $\gamma = 2.5$ are displayed. Increasing the value of w results in increased power as would be expected, though the difference between the power curves for $w \neq 999/1000$ is small. For $w = 999/1000$ the priority given to maximizing power results in an empirical power of 83.0%; above the desired level. The EB is observed to be small, relative to the value of θ, regardless of the value of w. However, only for $w = 999/1000$ is the final point estimate unbiased. A slightly larger impact on the ERMSE is seen for $w \neq 999/1000$ compared to the impact on the EB, though arguably performance is surprisingly strong considering $w = 999/1000$ results in the design that minimizes the ERMSE.

For $w \in \{1/1000, 1/4, 1/3, 1/2\}$ the EACP is almost identical and increases monotonically in θ. For $w = 999/1000$ the EACP is constant, indicative of the same design being chosen to maximize power no matter the value of θ. For $w \in \{2/3, 3/4\}$, the EACP initially increases in θ, but the competing factors in $s(\cdot)$ eventually result in decreases for larger θ. The ESDCP is maximized for each w when $\theta = 0$. The precise values of the ESDCP are arguably small when considered in unison with the EACP. For example, for $w = 1/2$, the ESDCP for $\theta = \delta$ together with the corresponding EACP indicates that in the majority of cases we would expect the roll-out to be sped up, as would be desired.

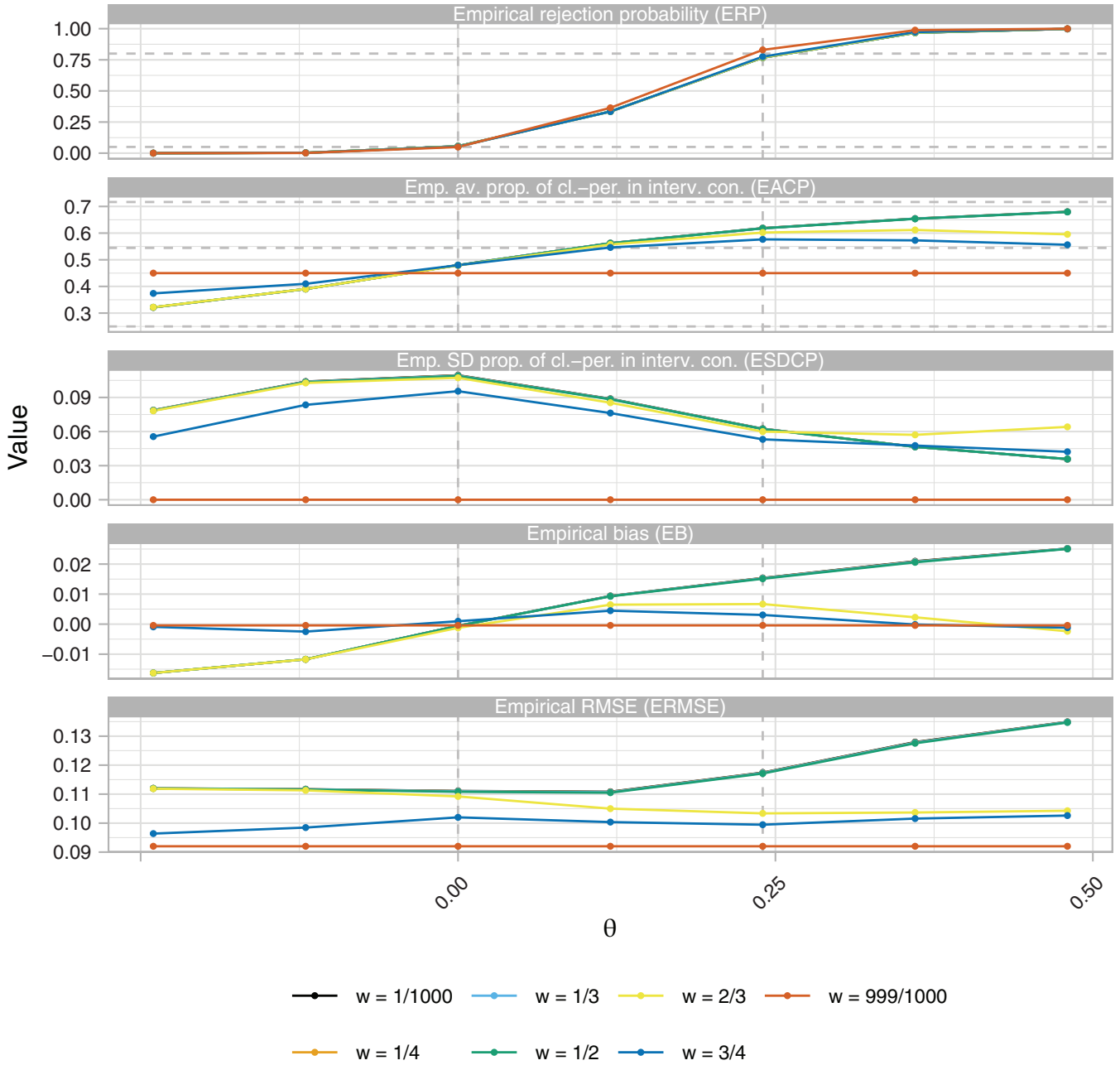

FIGURE 1 The empirical rejection probability (ERP) and empirical average proportion of cluster-periods spent in the intervention condition (EACP), as functions of w and θ, of the response adaptive (RA) stepped-wedge cluster randomized trial (SW-CRT) designs with $\eta = 0$, $\gamma = 2.5$, and $\{p_1, \ldots, p_L\} = \{3, 6\}$, in trial design scenario 1 (TDS1). The dashed lines in the ERP plot indicate the desired type-I and type-II error-rates. In the EACP plot they indicate the minimal, initially planned, and maximal values of the EACP based on $X = X_{p_1}$

For $w = 1/2$, the EACP ranges from 32.2% when $\theta = -\delta$ to 67.9% when $\theta = 2\delta$. Under the null and alternative hypotheses the corresponding figures are 48.0% and 61.8%, respectively. This contrasts to 54.4% for the fixed (initially-planned) design and 45.0% for $w = 999/1000$. Figure 2 displays this pictorially, giving the average value of X_P when $w = 1/2$. Similarly, Figure 3 presents the probability mass function of the proportion of time spent in the intervention condition. The probability of making an "incorrect" decision (eg, decreasing the roll-out speed for a large true intervention effect) is evidently small when the absolute value of θ is large. A potential downside of RA design is observed for, for example, $\theta = 0$, where the precise variation in the final proportion of participants who received the intervention

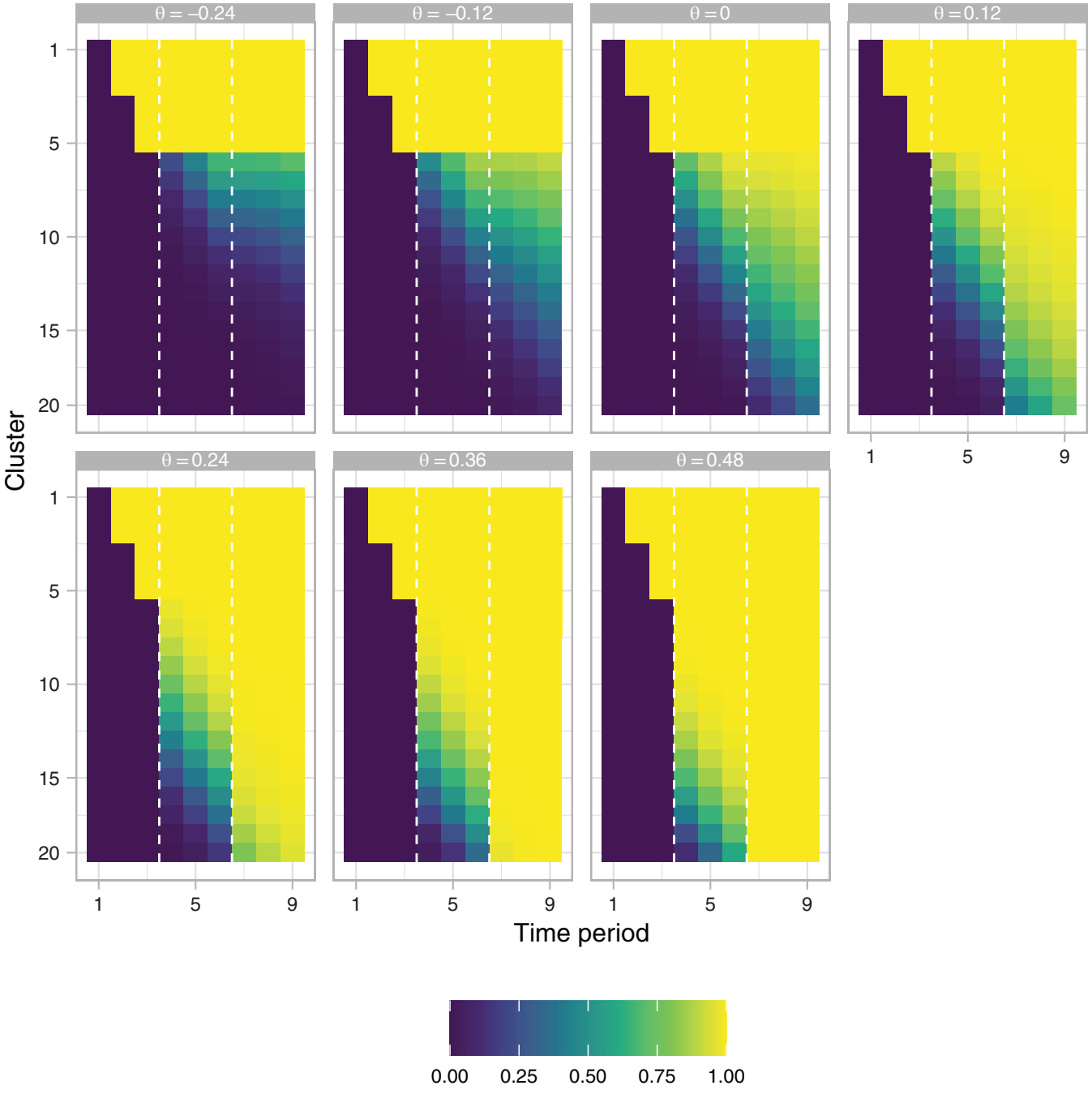

FIGURE 2 The empirical average final allocation matrix ($\overline{X}_P$), as a function of θ, of the response adaptive (RA) stepped-wedge cluster randomized trial (SW-CRT) design with $\eta = 0$, $\gamma = 2.5$, $w = 1/2$, and $\{p_1, \ldots, p_L\} = \{3, 6\}$, in trial design scenario 1 (TDS1). The dashed lines indicate the timing of the interim analyses

is evident, when in this case we may prefer some (fixed) value close to 50%. The empirical type-I error-rate and power are 5.6% and 76.8%, respectively, in this case.

3.2.2 | Operating characteristics as a function of η and γ

Figures 4 to 6, respectively, present the ERP, EACP, and ESDCP of the RA SW-CRT designs, as functions of w and θ, for different combinations of η and γ. Corresponding presentations for the EB and ERMSE are given in the Supplementary

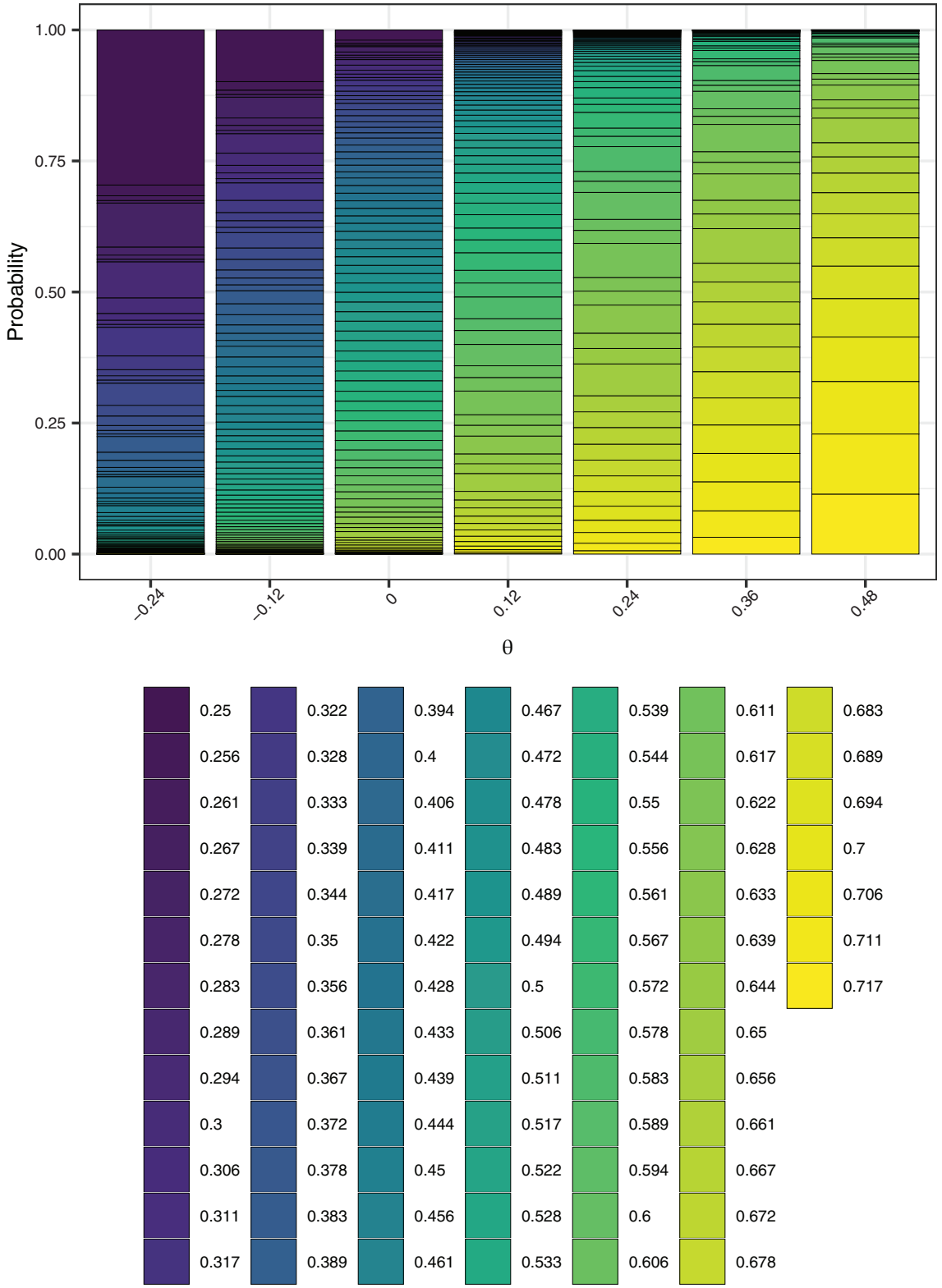

FIGURE 3 The empirical probability mass function of the proportion of cluster-periods spent in the intervention condition, as a function of θ, of the response adaptive (RA) stepped-wedge cluster randomized trial (SW-CRT) design with $\eta = 0$, $\gamma = 2.5$, $w = 1/2$, and $\{p_1, \ldots, p_L\} = \{3, 6\}$, in trial design scenario 1 (TDS1)

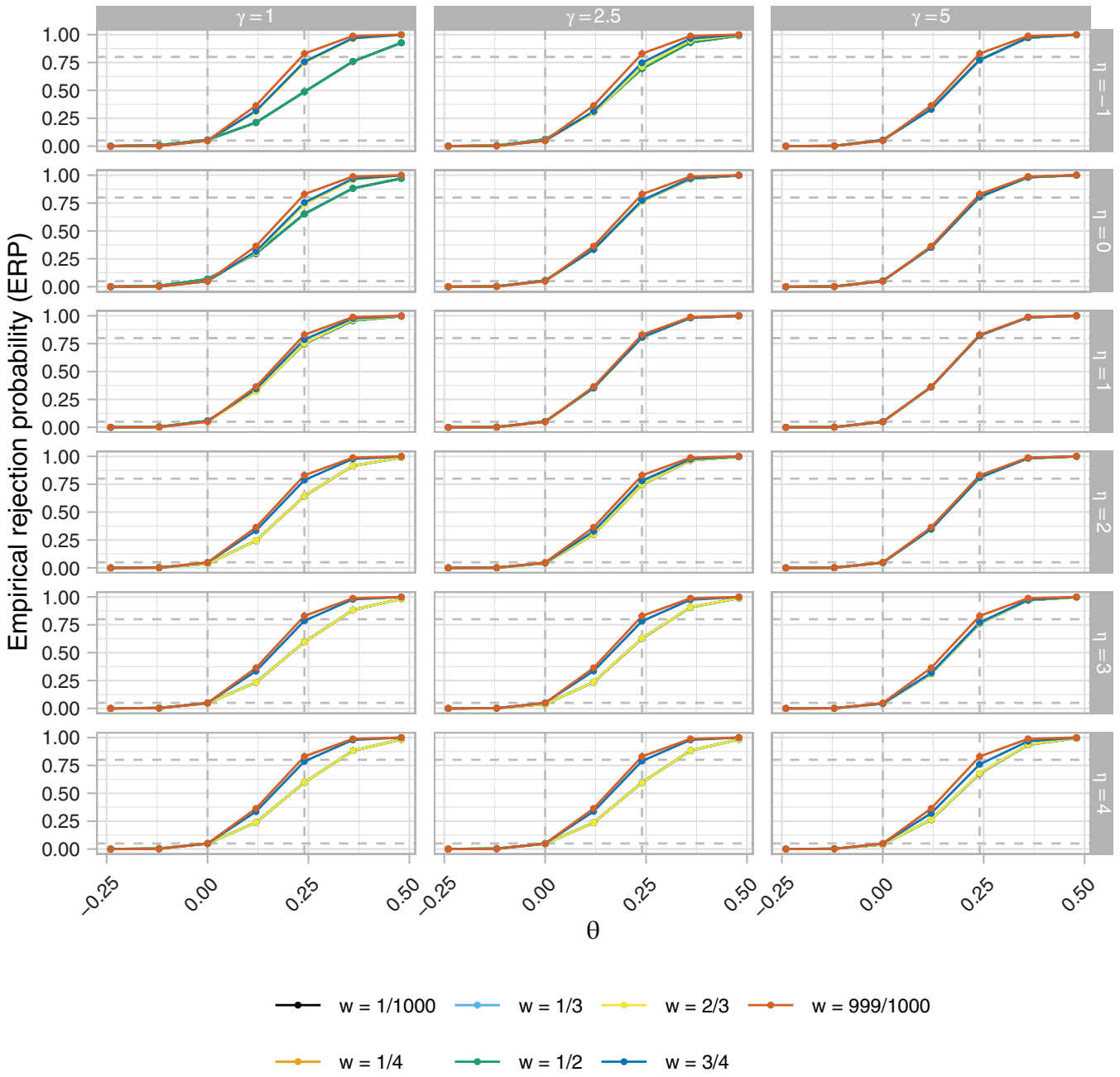

FIGURE 4 The empirical rejection probability (ERP), as a function of w and θ, of the response adaptive (RA) stepped-wedge cluster randomized trial (SW-CRT) designs with $\{p_1, \ldots, p_L\} = \{3, 6\}$ for different combinations of η and γ, in trial design scenario 1 (TDS1). The dashed lines indicate the desired type-I and type-II error-rates

Materials. For several combinations of η and γ the power curves are similar across θ for multiple values of w, attaining approximately the desired type-I error-rate and power. Larger differences are observed in some instances, however, typically for more extreme values of η and γ. For fixed η, increasing γ generally results in an increase in power. This should be anticipated as larger γ promotes a more steady roll-out, which will often correspond to allocation matrices with power closer to the desired level. Similarly, for fixed γ, increasing η initially results in power gains, but in many cases eventually leads to power loss as the procedure recommends those designs that terminate the roll-out.

These comments match the plots in Figure 5, with for example those designs with $\eta = 4$ having very low values for the EACP. Furthermore, it can be seen that for $w = 0.5$, for example, increasing γ generally results in a flattening of the EACP curve as a function of θ, as the more extreme roll-outs attain lower values for $b(\cdot)$.

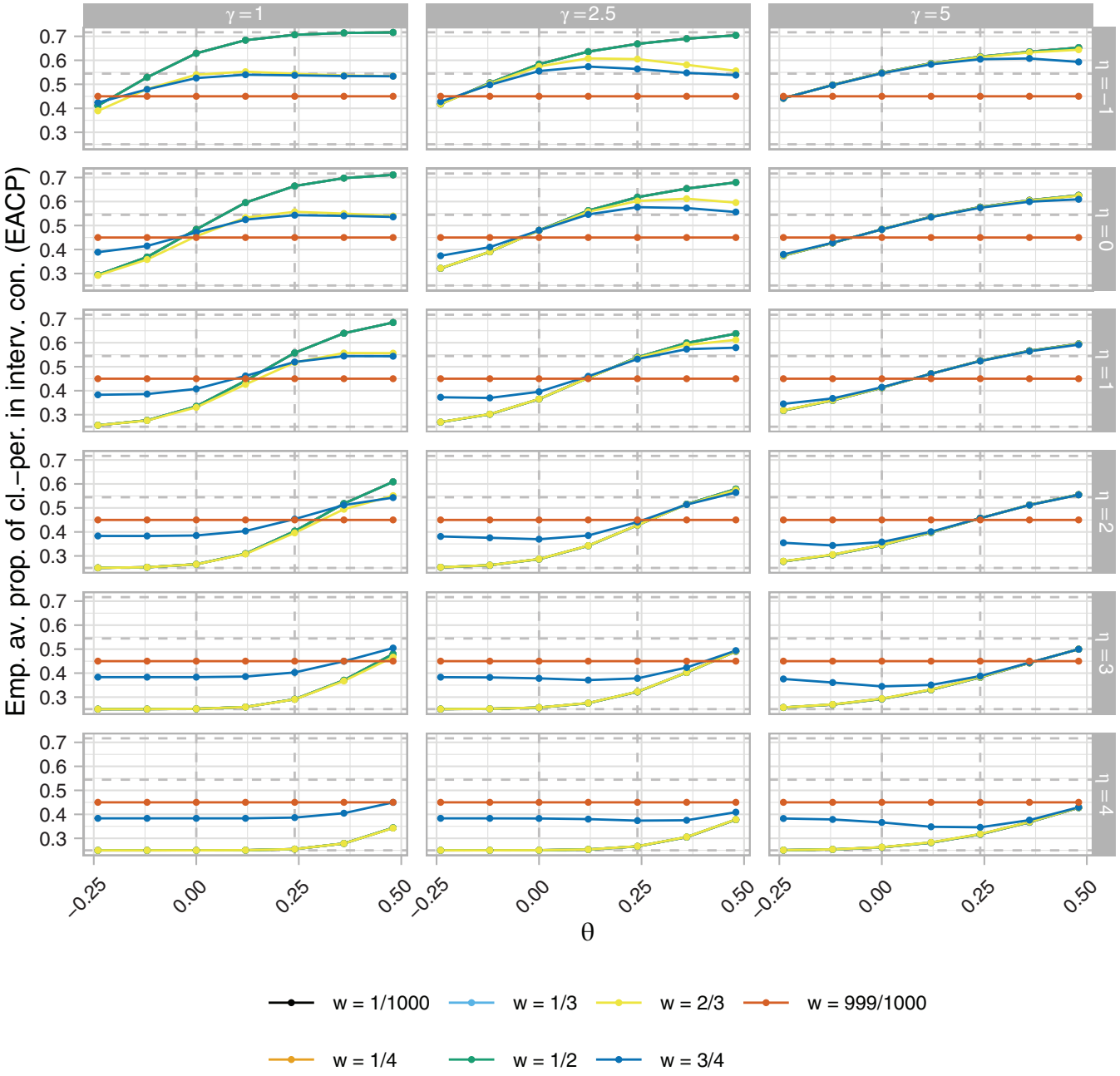

FIGURE 5 The empirical average proportion of cluster-periods spent in the intervention condition (EACP), as a function of w and θ, of the response adaptive (RA) stepped-wedge cluster randomized trial (SW-CRT) designs with $\{p_1, \dots, p_L\} = \{3, 6\}$ for different combinations of η and γ, in trial design scenario 1 (TDS1). The dashed lines indicate the minimal, initially planned, and maximal values of the EACP based on $X = X_{p_1}$

Qualitatively different findings are observed in Figure 6, however. For $\gamma = 5$, the ESDCP is similar for all $w \neq 999/10\,000$ and varies little as a function of θ or η. This is a consequence of large γ placing a high preference on approximately 50% of cluster-periods being spent in the intervention condition. For $\gamma = 2.5$, the ESDCP again varies little across values of $w \neq 999/1000$, but now varies substantially as a function of θ and η. The maximal values of the ESDCP for $\gamma = 2.5$ can often be considered low when viewed in combination with the corresponding EACP. This is not always the case for $\gamma = 1$, though, where for certain w (eg, $w = 1/2$) the ESDCP indicates variation in the roll-out speed such that performance may often be considered poor (eg, an increase in roll-out from that initially planned when $\theta < 0$).

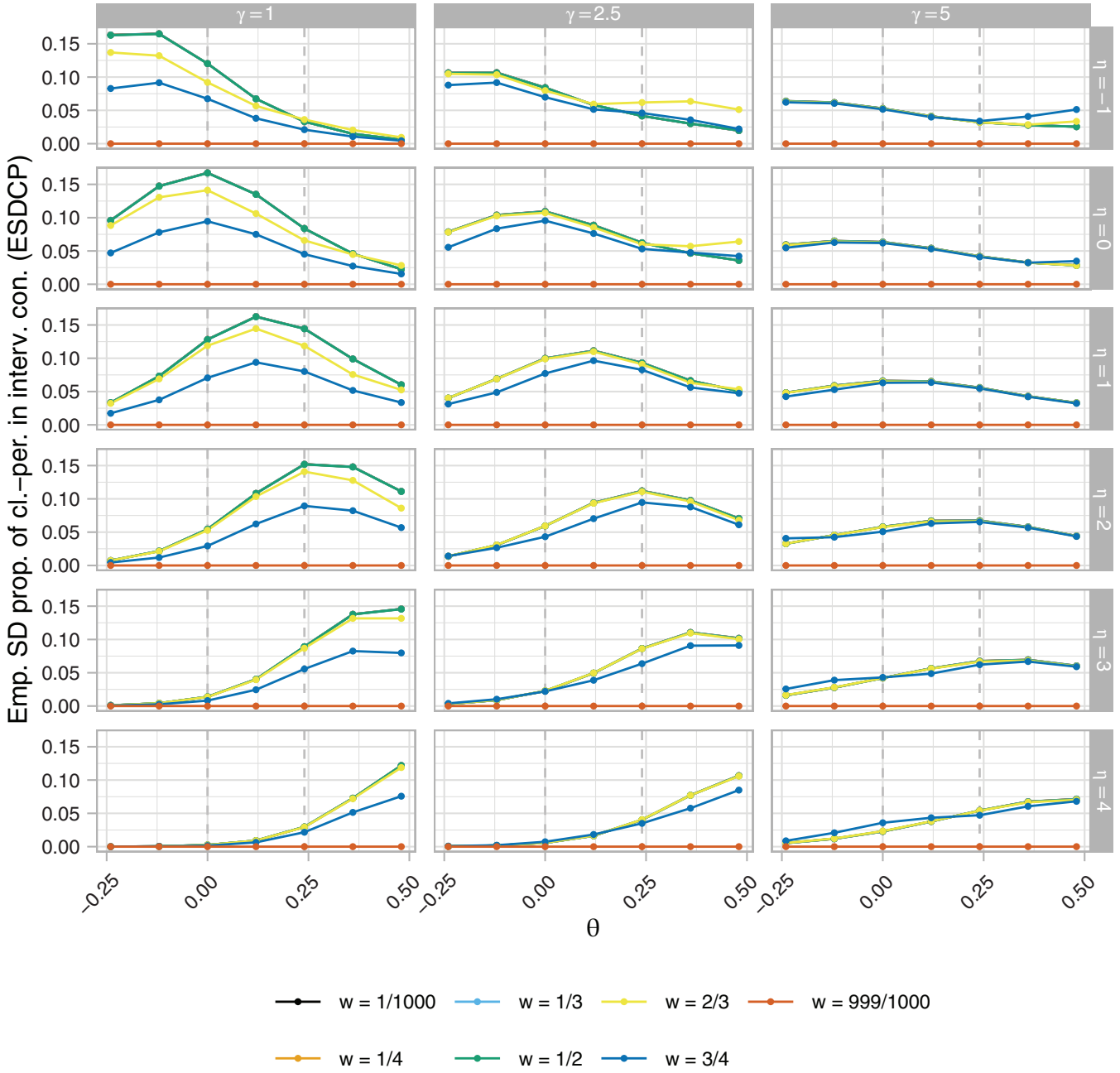

FIGURE 6 The empirical standard deviation of the proportion of cluster-periods spent in the intervention condition (ESDCP), as a function of w and θ, of the response adaptive (RA) stepped-wedge cluster randomized trial (SW-CRT) designs with $\{p_1, \dots, p_L\} = \{3, 6\}$ for different combinations of η and γ, in trial design scenario 1 (TDS1)

3.2.3 | Operating characteristics as a function of $\{p_1, \dots, p_L\}$

Figure 7 presents the ERP, EACP, ESDCP, EB, and ERMSE of the RA SW-CRT designs as functions of $\{p_1, \dots, p_L\}$ for $\eta = 0$ and $\gamma = 2.5$. It can be seen that there is little evidence changing the timing of a single-interim analysis from $\{3\}$ to $\{5\}$ carries a cost to the ERP. However, delaying the timing of a single-interim analysis inhibits the ability of the RA designs to offer a wider range of values for the EACP. In addition, the EACP curves are similar for an increasing number of values of w the later the timing of the interim analysis; this is a consequence of both the decrease in possible allocation matrices that can be chosen from and increasing precision in the value of the $Z_{p|X_p}$. Another consequence of this is that the ESDCP

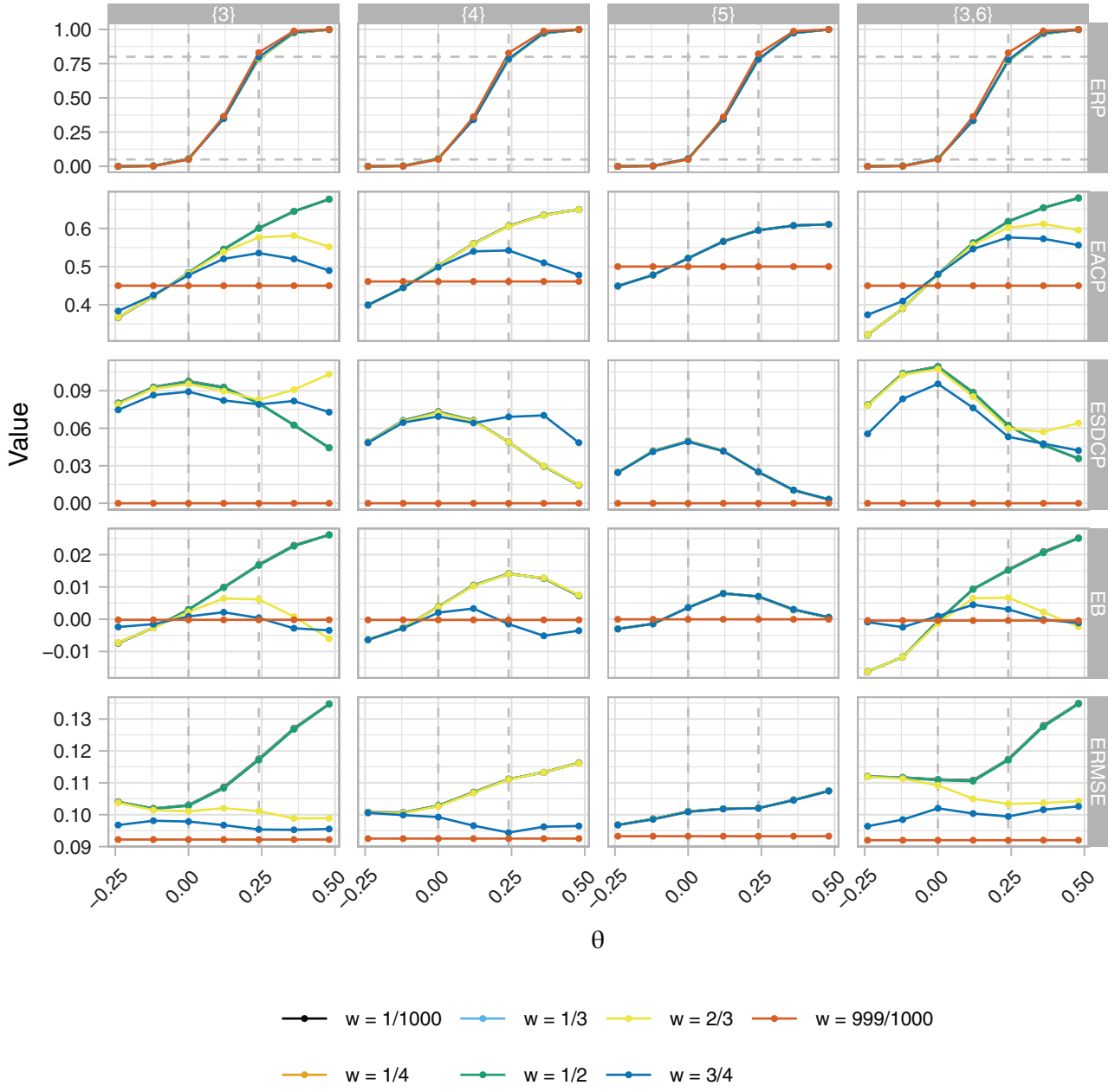

FIGURE 7 The empirical rejection probability (ERP), average proportion of cluster-periods spent in the intervention condition (EACP), standard deviation of cluster-periods spent in the intervention condition (EACP), bias (EB), and root-mean-square error (RMSE), as functions of w and θ, of the response adaptive (RA) stepped-wedge cluster randomized trial (SW-CRT) designs with $\eta = 0$ and $\gamma = 2.5$, for different values of $\{p_1, \ldots, p_L\}$, in trial design scenario 1 (TDS1). The dashed lines in the ERP plot indicate the desired type-I and type-II error-rates. In the EACP plot they indicate the minimal, initially planned, and maximal values of the EACP based on $X = X_{p_1}$

is smaller when the first interim analysis is conducted later in the trial, though the difference is only pronounced when {3} is contrasted with {5}.

There is a larger cost to the EB for certain w when an interim analysis is conducted earlier in the trial (ie, for {3} and {3,6}). However, the actual cost remains small relative to the value of θ. Similar statements are true for the ERMSE.

Compared to the designs with $\{p_1, \ldots, p_L\} = \{3\}$, those with $\{p_1, \ldots, p_L\} = \{3, 6\}$ incur a small cost to their empirical power. However, this is counterbalanced by them achieving a wider range of values for the EACP when $w \neq 999/1000$.

4 | DISCUSSION

Concerns have been expressed over the pre-trial decision of SW-CRTs to provide the intervention to all clusters. It may therefore be advantageous to allow the intervention roll-out to be sped-up or slowed-down according to information accrued during the trial. Accordingly, we have presented methodology through which this could be achieved. Our presented framework is flexible, allowing the design to be constructed to balance considerations on power and ethical allocation. Furthermore, while we focused on data analysis via a linear mixed model, the framework is dependent only on the availability of an interim estimate of effectiveness. It could therefore be readily modified, for example, for a generalized estimating equation analysis of noncontinuous data (see, eg, Li et al[26] or Ford and Westgate[27] for relevant methodology in the nonadaptive setting).

To examine the performance of the framework, we conducted a large simulation study. From this, several important observations can be made. Principally, it should not be assumed that any choice of values for η and γ will provide desirable operating characteristics. However, in all three TDSs it was possible to find combinations that provided monotonically increasing values for the EACP without major inflation of the type-I or type-II error-rate (eg, in TDS1 $w = 0.5$, $\eta = 0$, and $\gamma = 2.5$ provided such performance). Our recommendation would be therefore that these should be chosen carefully in practice, via a comprehensive simulation study. Nonetheless, it was clear that some small impact to the error-rates may be unavoidable if one is to attain a design with large variation in the EACP as a function of the intervention effect. The small power loss may be resolved in practice through a small increase to the sample size computed for the corresponding fixed sample design.

Addressing the observed type-I error-rate inflation poses an interesting question as to whether methodology developed to help attain a desired test size in small fixed-sample CRTs could find additional utility in adaptive design scenarios. Such methodology has been a topic of much recent interest. For example, Leyrat et al[28] considered the performance of numerous analysis methods (eg, weighted and unweighted cluster-level analyses, mixed-effects models with different degree-of-freedom corrections, GEEs with and without a small-sample correction) for parallel-group CRTs with a low number of clusters and a continuous outcome. Scott et al[29] and Ford and Westgate[27] examined possible correction methods for GEE analyses of SW-CRTs. Thompson et al[30] recently provided an extensive comparison of such small-sample correction methods and degrees-of-freedom corrections for GEE analyses of binary data in SW-CRTs, with Ren et al[31] previously conducting similar work in a continuous outcome setting. While the type-I error-rate inflation observed in our RA SW-CRTs was often small, if addressing such inflation was a priority then it is likely such methodology would offer a potential, albeit heuristic, solution. We note though that simulation would be required to ascertain which approach may be most appropriate, as there is no guarantee results in a fixed-sample setting would be directly transferable to RA design.

The advantageous performance of the RA designs is particularly noteworthy since only designs with a small number of interim analyses were evaluated. One may have anticipated that more interim analyses may have been required to realize benefits of RA randomization. A small number of interim analyses may be important in practice to reduce their logistical burden. It is also more computationally feasible to evaluate performance in this setting and it may be anticipated to be associated with smaller inflation of the type-I error-rate as the data is assessed less frequently.

The findings should perhaps not be surprising, given the large number of alternatives to the initially planned allocation matrix that will have similar power means there are often other choices available that can at least slightly alter the intervention's allocation without compromising on power. Furthermore, the timing of the first interim analysis provides a natural and effective means of protecting a degree of data accrual in the intervention and control conditions; this is similar to the typical use of a burn-in period for RA designs in individually randomized trials. The timing of the first interim analysis can also be seen to be crucial to enabling a wider range of EACP values to be possible; the one-directional switching of SW-CRTs means that RA design can offer far less later in a trial as the number of possible allocation schemes decreases. However, we note that even when only small changes in the EACP are achieved this can have a substantial impact on the number of patients who receive the intervention, depending on the value of the total trial sample size. Finally, there was substantial degeneracy in the operating characteristics for different values of w, particularly in those designs where the first interim analysis was timed later in the trial. In practice, only a small number of values for w may need to be considered, and in many instances the choice of $w = 1/2$ worked well.

It is important to acknowledge some limitations to our work. First, while our investigations reveal limited impact on the bias in the final point estimate from utilizing an RA design, we have not addressed potentially important characteristics of the asymptotic properties of the estimator (eg, consistency) or provided a way to remove any bias. We leave extending bias removal methodology for individually randomized RA trials[25] to this SW-CRT setting for future work. Nor have we examined the potential implications of model mis-specification on the utility of the proposed RA

procedure. Recent work has, as discussed, highlighted a range of possible analysis methods that make, for example, differing assumptions on the correlation between the outcome measurements.[15] It is possible that model mis-specification may impact RA design more starkly than it does a fixed-sample SW-CRT. While there is potential in an adaptive setting to adaptively update the chosen analysis model, which could help overcome such a problem, we have not addressed this here and no work to date is available to indicate whether this may be a fruitful approach. Each of these considerations may, in particular, impact the applicability of the proposed methodology in a regulated trial setting.

In addition, while we have provided examples on cross-sectional and closed-cohort designs, we have not directly addressed RA design of an open-cohort SW-CRT. Our methods could be applied to an open-cohort SW-CRT under the assumption of some particular sampling scheme.[32] However, the degree to which the assumed sampling scheme is "correct" would then likely influence the usefulness of RA design. Consequently, the approach to RA design for an open-cohort trial should arguably also attempt to re-estimate the "true" sampling scheme at the time of the interim analyses, which we have not presented methodology for here. Regardless of the approach used, thorough investigation of the utility of RA design for open-cohort designs would then require simulations to be performed under a variety of open-cohort sampling schemes, with exploration of the impact of these being correctly or incorrectly specified.

The practical considerations in relation to utilizing an RA SW-CRT design should also be recognized. Many of these are similar to those described in Grayling et al[22] within the context of early termination in SW-CRTs. In particular, while the time period structure of SW-CRTs may appear to lend itself naturally to sequential methodology, the interim analyses would be highly dependent upon the efficient collection, storage, and processing of data. Arguably the largest issue for RA intervention allocation, though, is whether logistical or practical constraints may inhibit the ability to modify the roll-out. While a roll-out could likely often be slowed down, it may be challenging to speed it up. Furthermore, allowing slow-down could be argued to disincentivize cluster participation.

Limitations above aside, our results indicate RA allocation of the intervention could potentially provide notable advantages. It is important to discuss therefore when such a design may be useful. In practice, RA design could be deemed useful in a wide variety of settings, where this conclusion may not be immediately apparent; a number of SW-CRTs have now incorporated interim evaluations of efficacy/futility,[33-39] and it is not always clear from published information why such adaptations were included. However, we note that RA could be particularly helpful when either the intervention itself or its evaluation is highly expensive, such that investigators would not wish to complete the roll-out unless it was effective. Most likely though, in our opinion, it may be helpful when there are substantial patient benefit considerations associated with the intervention, potentially in combination with notable safety concerns. This could be true, for example, of vaccine development during an epidemic.

Following the Ebola outbreak of 2014 to 2015, many authors discussed the applicability of SW-CRTs to evaluating vaccine effectiveness.[40-56] Importantly, this setting was one in which a short time was expected between intervention delivery and outcome accrual,[52] which is important for RA design. Furthermore, there was little data available about the safety or immunogenicity of the vaccine candidates.[44] Consequently, proposals to use SW-CRT designs were not based on preliminary data that the vaccine may do more good than harm and the safety considerations arguably amplify the need to prevent roll-out if a vaccine was ineffective. Indeed, van der Tweel and van der Graaf[56] noted their concerns that many clusters could end up being exposed to an inferior treatment, while Doussau and Grady[44] went as far as to state that interim analyses may be needed. It also seems reasonable to assume such a setting would be one in which resources would be made available to carry out interim adaptations efficiently, owing to the degree of the public health emergency.

The main limitation to utilizing an RA SW-CRT design of the type considered here would be the aforementioned resource availability to speed up a vaccines roll-out. It would be important to ensure that at the epidemic's onset manufacturing processes were put in place to scale up the development of any vaccine for which preliminary evidence of effectiveness was obtained. The other principal limitation, discussed extensively by Bellan et al,[40] is that SW-CRTs are not well equipped to handling spatiotemporal variation in a virus outbreak; much power can often be gained from prioritizing where to administer a vaccine. This issue cannot be handled by the type of RA SW-CRT proposed here. However, it indicates that an adaptive incomplete-block CRT may be worth considering in future studies of the efficient evaluation of a vaccine. Such a design could add new clusters during the course of the study, constraining the randomization to prioritize the speed of its delivery to specific hot-spots. We note it may also be important to consider incorporating other types of adaptation in to this type of design, including stopping rules[19,22] or sample size re-estimation,[21] in order to identify the most suitable CRT design.

In conclusion, when it is feasible to modify an intervention's allocation in an SW-CRT, RA design theory could help improve the trial's patient benefit characteristics. This may be particularly relevant to settings in which the intervention is expensive or could be associated with significant harm.

ACKNOWLEDGEMENTS

This work was supported by the Medical Research Council (grant number MC_UU_00002/6 to JMSW and grant number MC_UU_00002/15 to SSV).

DATA AVAILABILITY STATEMENT

Code to reproduce our results is available from https://github.com/mjg211/article_code.

ORCID

Michael J. Grayling https://orcid.org/0000-0002-0680-6668
James M. S. Wason https://orcid.org/0000-0002-4691-126X

REFERENCES

1. Hemming K, Taljaard M, McKenzie J, et al. Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. *BMJ*. 2018;363:k1614.
2. Kotz D, Spigt M, Arts I, Crutzen R, Viechtbauer W. Use of the stepped wedge design cannot be recommended: a critical appraisal and comparison with the classic cluster randomized controlled trial design. *J Clin Epidemiol*. 2012;65:1249-1252.
3. Mdege N, Man MS, Taylor nee Brown C, Torgerson D. There are some circumstances where the stepped-wedge cluster randomized trial is preferable to the alternative: no randomized trial at all. response to the commentary by Kotz and colleagues. *J Clin Epidemiol*. 2012;65:1253-1254.
4. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med*. 1987;317:141-145.
5. 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.
6. de Hoop E, van der Tweel I, van der Graaf R, et al. The need to balance merits and limitations from different disciplines when considering the stepped wedge cluster randomized trial design. *BMC Med Res Methodol*. 2015;15:93.
7. Hu F, Rosenberger W. *The Theory of Response-Adaptive Randomization in Clinical Trials*. Hoboken, NJ: John Wiley & Sons; 2006.
8. Atkinson A, Biswas A. *Randomised Response-Adaptive Designs in Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC Press; 2014.
9. Antognini A, Giovagnoli A. *Adaptive Designs for Sequential Treatment Allocation*. Boca Raton, FL: Chapman & Hall/CRC Press; 2015.
10. Robertson D, Lee K, Lopez-Kolkovska B, Villar S. Response-adaptive randomization in clinical trials: from myths to practical considerations; 2020. arXiv:2005.00564.
11. Hussey M, Hughes J. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials*. 2007;28:182-191.
12. Woertman W, de Hoop E, Moerbeek M, Zuidema S, Gerritsen D, Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J Clin Epidemiol*. 2013;66:752-758.
13. Hemming K, Taljaard M. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. *J Clin Epidemiol*. 2016;69:137-146.
14. Hooper R, Teerenstra S, de Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Stat Med*. 2016;35:4718-4728.
15. Li F, Hughes J, Hemming K, Taljaard M, Melnick E, Heagerty P. Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: an overview. *Stat Meth Med Res*. 2021;30:612-639.
16. Liang Y, Li Y, Wang J, Carriere K. Multiple-objective response-adaptive repeated measurement designs in clinical trials for binary responses. *Stat Med*. 2014;33:607-617.
17. Girling A, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Stat Med*. 2016;35:2149-2166.
18. Li F, Turner E, Preisser J. Optimal allocation of clusters in cohort stepped wedge designs. *Stat Probab Lett*. 2018;137:257-263.
19. Grayling M, Robertson D, Wason J, Mander A. Design optimisation and post-trial analysis in group sequential stepped-wedge cluster randomised trials; 2018. arXiv:1803.09691v1.
20. Grayling M, Wason J, Mander A. Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. *Trials*. 2017;18:33.
21. Grayling M, Mander A, Wason J. Blinded and unblinded sample size re-estimation procedures for stepped-wedge cluster randomized trials. *Biom J*. 2018;60:903-916.
22. Grayling M, Wason J, Mander A. Group sequential designs for stepped-wedge cluster randomised trials. *Clin Trials*. 2017;14:507-517.
23. Bashour H, Kanaan M, Kharouf M, Abdulsalam A, Tabbaa M, Cheikha SA. The effect of training doctors in communication skills on women's satisfaction with doctor-woman relationship during labour and delivery: a stepped wedge cluster randomised trial in damascus. *BMJ Open*. 2013;3:e002674.

24. Tirlea L, Truby H, Haines T. Investigation of the effectiveness of the "Girls on the Go!" program for building self-esteem in young women: trial protocol. *Springerplus*. 2013;2:683.
25. Bowden J, Trippa L. Unbiased estimation for response adaptive clinical trials. *Stat Meth Med Res*. 2017;26:2376-2388.
26. Li F, Turner E, Preisser J. Sample size determination for GEE analyses of stepped wedge cluster randomized trials. *Biometrics*. 2018;74:1450-1458.
27. Ford W, Westgate P. Maintaining the validity of inference in small-sample stepped wedge cluster randomized trials with binary outcomes when using generalized estimating equations. *Stat Med*. 2020;39:2779-2792.
28. Leyrat C, Morgan K, Leurent B, Kahan B. Cluster randomized trials with a small number of clusters: which analyses should be used? *Int J Epidemiol*. 2018;47:321-331.
29. Scott J, deCamp A, Juraska M, Fay M, Gilbert P. 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.
30. Thompson J, Hemming K, Forbes A, Fielding K, Hayes R. Comparison of small-sample standard-error corrections for generalised estimating equations in stepped wedge cluster randomised trials with a binary outcome: a simulation study. *Stat Meth Med Res*. 2021;30:425-439.
31. Ren Y, Hughes J, Heagerty P. A simulation study of statistical approaches to data analysis in the stepped wedge design. *Stat Biosci*. 2019;12:399-415.
32. Kasza J, Hooper R, Copas A, Forbes A. Sample size and power calculations for open cohort longitudinal cluster randomized trials. *Stat Med*. 2020;39:1871-1883.
33. Curley M, Gedeit R, Dodson B, et al. Methods in the design and implementation of the randomized evaluation of sedation titration for respiratory failure (RESTORE) clinical trial. *Trials*. 2018;19:687.
34. Dias M, De Oliveira L, Jeyabalan A, et al. PREPARE: protocol for a stepped wedge trial to evaluate whether a risk stratification model can reduce preterm deliveries among women with suspected or confirmed preterm pre-eclampsia. *BMC Preg Child*. 2019;19:343.
35. Hayes-Ryan D, Hemming K, Breathnach F, et al. PARROT Ireland: placental growth factor in assessment of women with suspected pre-eclampsia to reduce maternal morbidity: a stepped wedge cluster randomised control trial research study protocol. *BMJ Open*. 2019;9:e023562.
36. Huffman M, Mohanan P, Devarajan R, et al. Effect of a quality improvement intervention on clinical outcomes in patients in india with acute myocardial infarction: the ACS QUIK randomized clinical trial. *JAMA*. 2018;319:567-578.
37. Lundström E, Isaksson E, Wester P, Laska AC, Näsman P. Enhancing recruitment using teleconference and commitment contract (ERUTECC): study protocol for a randomised, stepped-wedge cluster trial within the EFFECTS trial. *Trials*. 2018;19:14.
38. Newman K, Rogers J, McCulloch D, et al. Point-of-care molecular testing and antiviral treatment of influenza in residents of homeless shelters in Seattle, WA: study protocol for a stepped-wedge cluster-randomized controlled trial. *Trials*. 2020;21:956.
39. Reeder R, Girling A, Wolfe H, et al. Improving outcomes after pediatric cardiac arrest – the ICU-resuscitation project: study protocol for a randomized controlled trial. *Trials*. 2018;19:213.
40. Bellan S, Pulliam J, Pearson C, et al. The statistical power and validity of Ebola vaccine trials in Sierra Leone: a simulation study of trial design and analysis. *Lancet Infect Dis*. 2015;15:703-710.
41. Chowell G, Viboud C. Ebola vaccine trials: a race against the clock. *Lancet Infect Dis*. 2015;15:624-626.
42. Dean N, Gsell P, Brookmeyer R, et al. Considerations for the design of vaccine efficacy trials during public health emergencies. *Sci Transl Med*. 2019;11:eaat0360.
43. Diakite I, Mooring E, Velasquez G, Murray M. Novel ordered stepped-wedge cluster trial designs for detecting Ebola vaccine efficacy using a spatially structured mathematical model. *PLoS Negl Trop Dis*. 2016;10:e0004866.
44. Doussau A, Grady C. Deciphering assumptions about stepped wedge designs: the case of Ebola vaccine research. *J Med Ethics*. 2016;42:797-804.
45. Edwards S. Response to open peer commentaries on "Ethics of clinical science in a public health emergency: drug discovery at the bedside". *Am J Bioethics*. 2013;13:9.
46. Eyal N, Lipsitch M. Vaccine testing for emerging infections: the case for individual randomisation. *J Med Ethics*. 2017;43:625-631.
47. Halloran M, Auranen K, Baird S, et al. Simulations for designing and interpreting intervention trials in infectious diseases. *BMC Med*. 2017;15:223.
48. Hitchings M, Grais R, Lipsitch M. Using simulation to aid trial design: ring vaccination trials. *PLoS Negl Trop Dis*. 2017;11:e0005470.
49. Kahn R, Rid A, Smith P, Eyal N, Lipsitch M. Choices in vaccine trial design in epidemics of emerging infections. *PLoS Med*. 2018;15:e1002632.
50. Lipsitch M, Eyal N. Improving vaccine trials in infectious disease emergencies. *Science*. 2018;357:153-156.
51. Nason M. Statistics and logistics: design of Ebola vaccine trials in West Africa. *Clin Trials*. 2016;13:87-91.
52. Piszczek J, Partlow E. Stepped-wedge trial design to evaluate Ebola treatments. *Lancet Infect Dis*. 2015;15:762-763.
53. Pulliam J, Bellan S, Gambhir M, Ancel Meyers L, Dushoff J. Evaluating Ebola vaccine trials: insights from simulation. *Lancet Infect Dis*. 2015;15:1134.
54. Tully C, Lambe T, Gilbert S, Hill A. Emergency Ebola response: a new approach to the rapid design and development of vaccines against emerging diseases. *Lancet Infect Dis*. 2015;15:1356-1359.
55. Vandebosch A, Mogg R, Goeyvaerts N, et al. Simulation-guided phase 3 trial design to evaluate vaccine effectiveness to prevent Ebola virus disease infection: statistical considerations, design rationale, and challenges. *Clin Trials*. 2016;13:57-65.

56. van der Tweel I, van der Graaf R. Issues in the use of stepped wedge cluster and alternative designs in the case of pandemics. *Am J Bioethics*. 2013;13:W1-W3.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Grayling MJ, Wason JMS, Villar SS. Response adaptive intervention allocation in stepped-wedge cluster randomized trials. *Statistics in Medicine*. 2022;41(6):1081-1099. doi: 10.1002/sim.9317