UW Biostatistics Working Paper Series

3-16-2016

Evaluation of multiple interventions using a stepped wedge design

Vivian H. Lyons
University of Washington - Seattle Campus, vlyons@uw.edu

Lingyu Li
University of Washington - Seattle Campus, lli6@uw.edu

James Hughes
University of Washington - Seattle Campus, jphughes@u.washington.edu

Ali Rowhani-Rahbar
University of Washington, rowhani@uw.edu

Evaluation of multiple interventions using a stepped wedge design

Vivian H Lyons[1,2], Lingyu Li[3], James Hughes[3], Ali Rowhani-Rahbar[1,2]

1: Department of Epidemiology, University of Washington, Seattle, WA
2: Harborview Injury Prevention & Research Center, Seattle, WA
3: Department of Biostatistics, University of Washington, Seattle, WA

Funding:
This work was supported by Grant #T76 MC00011 from the Maternal and Child Health Bureau, HRSA and Grant #AI29168 from the National Institute of Allergy and Infectious Diseases.

Abstract

Background: Stepped wedge cluster randomized trials are a class of unidirectional crossover studies that have historically been limited to evaluating a single intervention. This design is especially suitable for pragmatic trials where the study feasibility can be improved with a phased introduction of the intervention. We examined variations of stepped wedge designs that would support evaluation of multiple interventions. **Methods:** We propose four different design variants for implementing a stepped wedge trial with two interventions: concurrent design, supplementation, replacement, and factorial designs. Analyses were conducted comparing the precision of the estimated intervention effects for the different designs. **Results**: Concurrent, supplementation, and factorial variants provide equal precision for estimating the treatment effect within a design for each of the interventions. However, in the replacement design, the effect of the first introduced intervention is generally estimated more precisely than the second intervention. Surprising and nonintuitive changes in the precision of the intervention effect estimates are observed when additional observation time intervals are included in multiple intervention designs. **Conclusion**: These stepped wedge design variations offer alternative methods for studying two interventions using a cluster-randomized trial. The selection of the appropriate variants should be driven by the research question with consideration given to the trade-off in number of steps, number of clusters, restrictions for concurrent implementation based on intervention characteristics, lingering effects of each intervention, and desired ability to compare interventions within clusters or within the same steps.

Introduction

The use of stepped wedge designs (SWD) has been gaining popularity since the Gambia Hepatitis study.[1] The classic SWD uses a unidirectional crossover design with time until the implementation of the intervention randomized at the cluster level (See Figure 1). SWDs offer a pragmatic approach, where each cluster acts as a control at the first time step with one or more clusters receiving the intervention at each subsequent time step until all clusters receive the intervention by the end of the study.[2] This is a major departure from parallel cluster randomized trials (CRT) where assignments for intervention and control groups are decided at the beginning of the study and not changed for the study duration.[3]

In a SWD trial the intervention effect is determined using both within-cluster and a between-cluster information. Indeed, one of the strengths of a SWD is that it allows investigators to conduct a within-cluster analysis, thus avoiding time-invariant confounding on the cluster level and gaining precision in measurement.[3] The stepped rollout can also increase logistic feasibility for researchers when simultaneous intervention implementation in many clusters may be prohibitive. However, there are also inherent limitations in a SWD. By design, the treatment effect in a SWD is partially confounded by time so estimation of the intervention effect is model-dependent. Use of the SWD is also dependent on the amount of time necessary to begin rolling out the intervention for each cluster as well as any needed wash-in or wash-out period.[4] Other limitations have been discussed in Hughes, Granston and Haegerty[5] and elsewhere.[2]

SWDs have been historically limited to evaluating a single intervention. This paper describes design variations to support evaluation of multiple interventions using a SWD. Evaluating multiple interventions has the potential to decrease time until each cluster receives an intervention, thus improving participant engagement, and decrease total funding, subjects and study support needed by employing more than one intervention in a single study that otherwise

would have necessitated two separate investigations. Including multiple interventions in a single SWD can also allow for evaluation of the interaction between interventions.

Design Variations

There are several ways to introduce multiple interventions in a SWD. Here, we outline four major variants, each suitable for addressing certain scientific questions of interest regarding the effect of two interventions in a SWD.

Design 1 - Concurrent

This design is conceptually similar to the classic SWD, with two separate interventions evaluated at the same time and in the same setting (Figure 2). Intervention 1 and Intervention 2 are each only offered to some clusters in the study like random assignment used in other RCTs. The difference between this design and the classic SWD is that the same period of time can serve as a control period for both interventions, thereby increasing efficiency in estimating intervention effects.

A recent protocol published by Reuther, et al. in 2014 proposed a Concurrent design to study the impact of two types of case conferences [WELCOME-IdA or WELCOME-NEO] (Figure 3) on quality of life in dementia patients in German nursing homes.[6] With 12 nursing homes, they will complete two separate SWD studies with 6 nursing homes in each, where each nursing home receives only one of the case conference methodologies. Investigators who conduct studies such as this to examine the effect of multiple interventions may be able to employ the proposed Concurrent design if it is appropriate in the specific setting in which their investigation is conducted.

Design 2 - Replacement

In this design introduction of Intervention 1 follows the same pattern as the introduction of the intervention in the classic SWD. Subsequently, however, Intervention 2 replaces Intervention

1 in each cluster (See Figure 2). The study ends when all clusters have received Intervention 2. This variation is appropriate for assessing two interventions that will not or cannot be employed at the same time and when the number of clusters available for study is limited. Note also that we assume that there are no carryover effects of intervention 1 once it is removed. However, a washout period could be added to minimize the potential for any carryover effects.

While we were unable to find a study that employed this exact design, we did identify a few studies that could have employed this design. Patel, et al. 2005 implemented a variation of the traditional unidirectional crossover study comparing a corn/soy-blend and read-to-use food to as dietary supplements for malnourished children (Figure 4).[7] They did not collect data on children who visited their clinics before the implementation of their Intervention 1 (i.e. there was no control period data collection), and follow up for Intervention 2 lasted the same amount of time for all but one of the clinics (Cluster 1 had Intervention 2 the entire time).

Haines, et al. (2015) recently published a protocol for a single study of two nested SWD studies, where the second SWD study begins in the time period after the first SWD study ends with both studies taking place in the same two hospitals (Figure 5).[8] For investigators interested in rolling out two interventions back to back, using the Replacement Design may decrease funding needed as well as total study duration needed compared to conducting two separate SWD studies in series.

Design 3 - Supplementation

Design 3, introduces Intervention 1 and then adds Intervention 2 (Figure 2). This is an appropriate design for two interventions where the second intervention requires the presence of the first intervention. While it would be possible to use this design when the two interventions could be offered both individually and simultaneously, the Factorial design (discussed below)

would be the most appropriate in that case. The ability to estimate the effect of Intervention 2 in this study relies on the assumption of no interaction between the two interventions.

Chinbuah, et al, used two interventions in the treatment of childhood fever (See Figure 6) [artesunate amodiaquine (AAQ) and AAQ plus amoxicillin (AAQ+AMX)] where their Intervention 2 added amoxicillin to their Intervention 1.[9] While their study closely followed our concurrent design, they would have needed fewer clusters if they had used the Supplementation Design.

As another example, the Supplementation Design could be used to analyze the rollout of a technological support piece, like a near-real time support system, used during surgery to improve patient outcomes that requires a different surgical protocol. In that case, training the surgical staff on the new protocol would be Intervention 1 with the addition of the technological support piece employed as Intervention 2. For future research questions, use of the Supplementation Design would allow investigators to analyze both the effect of Intervention 1 and Intervention 2 within every cluster.

Design 4 - Factorial

The Factorial Design is characterized by concurrent rollout of Interventions 1 and 2 thereby allowing estimation of Intervention 1 and 2 separately as well as together (See Figure 2). This design is appropriate for two interventions that can be employed concurrently or separately, and allows the investigator to estimate the interaction between the two interventions. This approach may also improve recruitment as it provides an intervention at an earlier time step for clusters that receive interventions last in a classic SWD. The Factorial Design would provide an intervention to all clusters earlier than a classic SWD does.

While we were unable to find any study that had employed multiple interventions in this way, we identified several studies that could have used the Factorial Design to assess the effects of

multiple interventions in the same group, rather than randomizing each group to receive a single intervention. One of these studies by Whittingham et al. used two behavioral intervention programs [Stepping Stones Triple P (SSTP) and SSTP plus Acceptance and Commitment Therapy (SSTP+ACT)] to target behavioral problems in children with cerebral palsy.[10] Parents of children with cerebral palsy were assigned to one of three arms of the study, SSTP, SSTP+ACT or waitlist control. Due to ethical reasons, the control group was offered SSTP at the end of the primary phase of the study and enrolled in a secondary phase. Inclusion of an ACT-only arm (as used in other studies)[11–14] and use of the Factorial Design could allow investigators to provide all patients with an intervention sooner than with the classic SWD. It would also allow investigators to assess potential synergistic or antagonistic effects between two interventions.

For investigators interested in testing different interventions in the same population, the Factorial Design could allow for the roll out of multiple interventions concurrently, conceptually similar to running two separate stepped wedge studies in the same study population at the same time. Durovni and colleagues recently completed two separate SWD studies in Brazil. Their first study examined the impact of TB screening training at HIV clinics on TB incidence[15] and their second study examined the impact of changing smear microscopy on the notification rate of lab-confirmed TB.[16] While these studies took place in different settings (HIV clinics vs. laboratories), it is possible that these studies could have been run in concert with each other using a Factorial design if they had chosen HIV clinics with attached laboratories for their study. A factorial design could measure the impact of the interventions on a single outcome or, as in the Physician's Health Study,[17] a SWD Factorial Design can also estimate Intervention 1 specific outcomes and Intervention 2 specific outcomes.

Comparing Designs

To study the characteristics of the four proposed designs, we used the following model for multiple interventions, which extends the original model for SWD of Hussey and Hughes,[4]

$$\eta_{ij} = \mu + a_i + b_j + \sum_{m=1}^{M} X_{mij}\theta_m \quad (1)$$

where a_i is a random effect for cluster i such that $a_i \sim N(0, \tau^2)$ (i in 1, …, I), b_j is a fixed effect corresponding to time interval j (j in 1, …, T - 1, $b_T = 0$ for identifiability), X_{mij} is an indicator of the treatment mode for the m^{th} intervention in cluster i at time j (1 = intervention; 0 = control), θ_m is the treatment effect for the m^{th} intervention, I is the number of clusters, T is the number of time points, and M is the number of interventions. Note that this model assumes no interaction (effect modification) between the interventions, although it could be readily extended to include such a term. A model for the observed cluster means is:

$$Y_{ij} = \eta_{ij} + e_{ij} \quad (2)$$

where $e_{ij} \sim N(0, \sigma^2)$. Therefore, we have $Y_{ij} \sim N(\eta_{ij}, \sigma^2)$. For binary outcomes, we use $\sigma^2 = \eta_{ij}(1 - \eta_{ij})/N$ where N is the number of individuals per cluster per time interval. The dependence between observations in the same cluster is often parameterized in terms of the intracluster correlation (ICC) defined as $ICC = \tau^2/(\tau^2 + \sigma_e^2)$ where $\sigma^2 = \sigma_e^2/N$.

Let $\theta = (\mu, \beta_1, \ldots, \eta_1, \ldots, \eta_m)$. For each study design, we can derive a corresponding design matrix[18], Z, and the variance of $\hat{\theta}$ (which is used for power calculations[4]) can be calculated by:

$$\mathrm{cov}(\hat{\theta}) = (Z' V^{-1} Z)^{-1} \quad (3)$$

where V is an $IT \times IT$ block diagonal matrix. Each $T \times T$ block is of the form:

$$\begin{pmatrix} \sigma^2+\tau^2 & \tau^2 & \cdots & \tau^2 \\ \tau^2 & \sigma^2+\tau^2 & \cdots & \tau^2 \\ \vdots & \vdots & \ddots & \vdots \\ \tau^2 & \tau^2 & \cdots & \sigma^2+\tau^2 \end{pmatrix}$$

We used this approach to evaluate the variance of η_1 and η_2 for various designs (see the appendix of Hughes, Granston and Haegerty[3] for more details).

Comparing efficiency of Designs 1-4.

Across-design comparison

Figure 7 shows the variance of the estimated intervention effects for the four designs shown in figure 2. In each case we assume 100 subjects in each cluster, prevalence equal to 0.05, a nonparametric time effect and no interaction between Interventions 1 and 2. All variances were calculated using equation 3. Note that direct comparison of the 4 designs is complicated by the variation in number of clusters and/or time periods between the designs. However, we do note a few interesting findings. For example, for any value of ICC, the variance of $\widehat{\eta_1}$ and $\widehat{\eta_2}$ are equal in the Concurrent, Supplementation and Factorial designs. That is, these three designs provide equal precision for estimating the treatment effect for Interventions 1 and 2. However, in the Replacement Design, Intervention 1 is generally estimated more precisely than Intervention 2 except for extreme designs with many steps and limited time spent in intervention 1. We also note that, when comparing the specific designs shown in figure 2, the Concurrent Design gives the most precise estimates for both intervention effects, followed by the Factorial Design and then the Supplementation and Replacement Designs. However, these differences are partly due to variation in the number of clusters and/or time periods between the designs so may not generalize.

Concurrent Design

As noted, the variances of the two treatment effects are equal in the concurrent design. Figure 8 illustrates the effect of additional time intervals added at the end of the study period once all clusters have received the interventions. For each additional time interval, the variance of $\hat{\eta}$ decreases, which is not surprising since we have more and more information to estimate the treatment effect. Interestingly, the variance of the intervention effect decreases with additional time periods even at ICC=0, which is not true with the standard, single intervention, stepped wedge design (See Supplemental Figure A1).

Replacement Design

As noted previously (Figure 7), the variance of $\widehat{\eta_1}$ is generally lower than the variance of $\widehat{\eta_2}$ in the replacement design. To better understand this, consider the independent case (ICC=0). When the ICC = 0 and model (1) is used for analysis, only between-cluster comparisons contribute to the estimation of treatment effect. There are direct between-cluster comparisons of intervention 1 to the control, while for intervention 2, we need to first compare it to intervention 1, and then use the information on the treatment effect of intervention 1 to indirectly compare the treatment effect for 2 compared to control. Therefore, we have less power to detect the treatment effect for intervention 2 .

We investigated the influence of the number of intervention 1 time intervals in this design (Figure 9) (for example, in figure 2, intervention 1 remains in place for 2 time intervals before being replaced by intervention 2). The design with one intervention 1 time interval has the highest variance for intervention 1 (because this option has the least information for intervention 1) but the lowest variance for intervention 2 (because, for a 3 cluster replacement design, this is the only option that includes a direct comparison of intervention 2 to the control). Interestingly, as the number of intervention 1 time intervals increases there is no consistent pattern on the

variance of the intervention effects. We see a similar phenomenon with the supplementation design and discuss reasons for this behavior in that section below.

We could also add 1, 2 or 3 intervention 2 time intervals to the end of the study. The plot illustrating differences in variance due to additional Intervention 2 time intervals is presented in Figure 10.We see that the variances are unchanged when ICC = 0 because adding more Intervention 2 time intervals does not include any between-cluster comparisons (any mean change at these added time points is completely absorbed in the additional time parameter, β_j, in model (1)). But more time intervals do contribute to the within-cluster comparisons and reduce the variance of the intervention effects when ICC > 0, even for intervention 1.

Supplementation

In supplementation designs, the variances of the intervention 1 and 2 effects are equal. We investigate the effect of the number of time intervals of intervention 1 in these designs (for example, figure 2 has two intervention 1 time intervals) in Figure 11.

Similar to the replacement design the observed pattern is not intuitive – adding more intervention 1 time intervals (more data) sometimes decreases precision, depending on the value of the ICC. The reason for this anomalous behavior appears to be related to the time parameterization in model (3). As more Intervention 1 time periods are added, the number of time periods also increases. Each additional time period requires estimation of an addition time parameter in model 1. Depending on the ICC, the additional information gained from more data might be outweighed by the information lost through the need to estimate additional time effects. Note that if time effects in model 1 are modeled parametrically (e.g. linear trend), then this anomalous behavior disappears and adding more Intervention 1 time periods increases precision as expected (Figure A2).

Factorial Designs

For a Factorial design with 6 clusters, we see that, similar to our findings for the Replacement and Supplementation designs, adding more single intervention time periods does not necessarily increase the precision of the intervention estimates (See Figure 12). However, when we have more clusters (N=8), the precision of the intervention effects increases with more time periods for all ICC values. Adding more 1+2 time intervals at the end of the study increases the precision of the intervention estimates, except at ICC = 0.

Discussion

These design variations offer alternative methods for studying more than one intervention using a stepped wedge cluster-randomized trial and can provide an opportunity for additional data collection and improved efficiency. The selection of the appropriate design variant should be primarily driven by the research question and nature of the interventions, with additional consideration given to ethical and logistical issues, the trade-off between number of time periods required, minimum number of clusters, restrictions for concurrent implementation based on intervention characteristics, and desired ability to compare interventions within clusters or between clusters within the same time period. Indeed, as noted by Hargreaves et al., ethical issues may preclude use of a stepped wedge design at all.[19]

The information gained from a study using one of these design variants could be especially useful for communities, especially those with finite resources, who can only implement a single intervention. For example, a school assessing interventions to reduce obesity in their students that could only support implementing a single intervention could examine the relative impact of either Intervention 1 or Intervention 2 in a SWD study and chose the Intervention with the largest impact. If that same school already had a program in place to reduce obesity that was similar to

one of the Interventions used in the study they are referencing, they would be able to weigh the potential additive effect of adding in the other Intervention against the cost of implementation. While each design could be preferred depending on the various study constraints (number of clusters, time steps, additively of interventions), some studies could have the ability to choose between different designs. Although we have equal precision to estimate the effects of Intervention 1 and Intervention 2 within each design among the Concurrent, Supplementation and Factorial Designs, the Replacement Design generally provides increased precision to estimate the effect of Intervention 1 versus Intervention 2. Choosing between designs when the study constraints could allow two different designs should also take into account which comparisons are of greatest interest, which assumptions are most plausible, and which design achieves the highest power.

There are some nonintuitive results in this study. In the Replacement design the variances of the two intervention effects are different, even when they are implemented for the same number of time periods. In contrast, in the Supplementation design, the two intervention effect variances are equal even when there are unequal periods of the two interventions.

It is also nonintuitive that in the Replacement, Supplementation, and Factorial designs, adding more time periods does not always increase the precision of the intervention effect estimates. As mentioned before, we explain this anomalous behavior by the fact that as more time periods of interventions are added, the number of time periods also increases, which means more time effects must be estimated. Depending on the ICC, when the additional information gained by getting more data appears to be outweighed by the information lost through the need to estimate additional time effects, we sometimes observe lower precision by adding more time periods. Using a parametric model for time (i.e. linear trend) reduces or eliminates these nonintuitive behaviors, albeit at the cost of an additional assumption.

Limitations

All of the calculations done comparing efficiency of models were done assuming no interaction of the interventions. However, an interaction effect could be validly estimated in our Factorial Design. The variance and ability to estimate the treatment effect of both Interventions was calculated with linear regression and normal errors, but this is reasonable for the analysis of cluster-level results. Finally, valid estimation of the Intervention 2 effect in the Replacement Design assumes that there is no carry-over effect of Intervention 1; however, if necessary, a washout period could be added between the interventions in the Replacement Design.

Further Variations

Investigators could employ these basic models flexibly with a number of small variations. Depending on the time needed to begin Intervention 2 after implementing Intervention 1 may lead to a design with additional time steps for Intervention 1 included per cluster. Haines, et al. 2014 discussed the benefits of using a SWD study to remove an intervention and test the potential lingering effect of an intervention and is currently testing that variation in a SWD study discussed above.[8],[20] There are likely other instances where the residual effect of removing an intervention or policy may be of scientific interest to researchers, in which case these variations could be employed with our design variations (See Supplemental Figures A3 and A4).

Other variations could include employing more than 2 interventions. Concurrent design could add two additional clusters and complete a third concurrent SWD study. For Replacement and Supplementation this would be a simple variation where Intervention 3 could be rolled out following the same pattern as Intervention 2. Adding a third intervention to the Factorial design would be more complex and likely necessitate enough additional clusters and time steps to limit any advantage gained with using a stepped wedge study design. Alternatively, variations could

include multiple-level designs as proposed by Hemming, et al. 2015.[21] Due to the highly flexible nature of these basic models, the variances and precision for any design variants (adding time steps, adding interventions, etc.) would need to be re-calculated. As the next step, we are currently working on assessing further variations of SWD studies and their relative strengths.

References:

1. Group GHS. The Gambia hepatitis intervention study. *Cancer Res*. 1987:5782-5787. http://cancerres.aacrjournals.org/content/47/21/5782.short.
2. Mdege ND, Man M-S, Taylor Nee Brown CA, Torgerson DJ. 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(9):936-948. doi:10.1016/j.jclinepi.2010.12.003.
3. Hughes JP, Granston TS, Heagerty PJ. Current issues in the design and analysis of stepped wedge trials. *Contemp Clin Trials*. 2015. doi:10.1016/j.cct.2015.07.006.
4. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials*. 2007;28(2):182-191. doi:10.1016/j.cct.2006.05.007.
5. Hughes JP, Granston TS, Heagerty PJ. Current issues in the design and analysis of stepped wedge trials. *Contemp Clin Trials*. 2015;45(Pt A):55-60. doi:10.1016/j.cct.2015.07.006.
6. Reuther S, Holle D, Buscher I, et al. Effect evaluation of two types of dementia-specific case conferences in German nursing homes (FallDem) using a stepped-wedge design: study protocol for a randomized controlled trial. *Trials*. 2014;15:319. doi:10.1186/1745-6215-15-319.
7. Patel MP, Sandige HL, Ndekha MJ, Briend A, Ashorn P, Manary MJ. Supplemental feeding with ready-to-use therapeutic food in Malawian children at risk of malnutrition. *J Health Popul Nutr*. 2005;23(4):351-357.
8. Haines TP, O'Brien L, Mitchell D, et al. Study protocol for two randomized controlled trials examining the effectiveness and safety of current weekend allied health services and a new stakeholder-driven model for acute medical/surgical patients versus no weekend allied health services. *Trials*. 2015;16:133. doi:10.1186/s13063-015-0619-z.
9. Chinbuah MA, Kager PA, Abbey M, et al. Impact of community management of fever (using antimalarials with or without antibiotics) on childhood mortality: a cluster-randomized controlled trial in Ghana. *Am J Trop Med Hyg*. 2012;87(5 Suppl):11-20. doi:10.4269/ajtmh.2012.12-0078.
10. Whittingham K, Sanders M, McKinlay L, Boyd RN. Interventions to reduce behavioral problems in children with cerebral palsy: An RCT. *Pediatrics*. 2014;133(5):e1249-e1257. doi:10.1542/peds.2013-3620.
11. Johns LC, Oliver JE, Khondoker M, et al. The feasibility and acceptability of a brief Acceptance and Commitment Therapy (ACT) group intervention for people with psychosis: The "ACT for life" study. *J Behav Ther Exp Psychiatry*. 2016;50:257-263. doi:10.1016/j.jbtep.2015.10.001.
12. Kohle N, Drossaert CHC, Schreurs KMG, Hagedoorn M, Verdonck-de Leeuw IM,

Bohlmeijer ET. A web-based self-help intervention for partners of cancer patients based on Acceptance and Commitment Therapy: a protocol of a randomized controlled trial. *BMC Public Health*. 2015;15:303. doi:10.1186/s12889-015-1656-y.
13. Mak YW, Loke AY. The acceptance and commitment therapy for smoking cessation in the primary health care setting: a study protocol. *BMC Public Health*. 2015;15:105. doi:10.1186/s12889-015-1485-z.
14. Weineland S, Arvidsson D, Kakoulidis TP, Dahl J. Acceptance and commitment therapy for bariatric surgery patients, a pilot RCT. *Obes Res Clin Pract*. 2012;6(1):e1-e90. doi:10.1016/j.orcp.2011.04.004.
15. 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 Infect Dis*. 2013;13(10):852-858. doi:10.1016/S1473-3099(13)70187-7.
16. Durovni B, Saraceni V, van den Hof S, et al. Impact of replacing smear microscopy with Xpert MTB/RIF for diagnosing tuberculosis in Brazil: a stepped-wedge cluster-randomized trial. *PLoS Med*. 2014;11(12):e1001766. doi:10.1371/journal.pmed.1001766.
17. Christen WG, Gaziano JM, Hennekens CH. Design of Physicians' Health Study II--a randomized trial of beta-carotene, vitamins E and C, and multivitamins, in prevention of cancer, cardiovascular disease, and eye disease, and review of results of completed trials. *Ann Epidemiol*. 2000;10(2):125-134.
18. Seber GA, Lee AJ. *Linear Regression Analysis*. Vol Vol. 936. John Wiley & Sons; 2012.
19. Hargreaves JR, Copas AJ, Beard E, et al. Five questions to consider before conducting a stepped wedge trial. *Trials*. 2015;16:350. doi:10.1186/s13063-015-0841-8.
20. Haines T, O'Brien L, McDermott F, et al. A novel research design can aid disinvestment from existing health technologies with uncertain effectiveness, cost-effectiveness, and/or safety. *J Clin Epidemiol*. 2014;67(2):144-151. doi:10.1016/j.jclinepi.2013.08.014.
21. Hemming K, Lilford R, Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *Stat Med*. 2015;34(2):181-196. doi:10.1002/sim.6325.

Figures

Figure 1: The classic stepped wedge study

	Time 0	Time 1	Time 2	Time 3
Cluster 1	0	1	1	1
Cluster 2	0	0	1	1
Cluster 3	0	0	0	1

Figure 2: Proposed SWD Variations

Design 1 – Concurrent

	Time 0	Time 1	Time 2
Cluster 1	0	1	1
Cluster 2	0	0	1
Cluster 3	0	2	2
Cluster 4	0	0	2

Design 2 - Replacement

	Time 0	Time 1	Time 2	Time 3	Time 4	Time 5
Cluster 1	0	1	1	2	2	2
Cluster 2	0	0	1	1	2	2
Cluster 3	0	0	0	1	1	2

Design 3 – Supplementation

	Time 0	Time 1	Time 2	Time 3	Time 4	Time 5
Cluster 1	0	1	1	1+2	1+2	1+2
Cluster 2	0	0	1	1	1+2	1+2
Cluster 3	0	0	0	1	1	1+2

Design 4 – Factorial

	Time 0	Time 1	Time 2	Time 3	Time 4
Cluster 1	0	1	1	1+2	1+2
Cluster 2	0	0	1	1	1+2
Cluster 3	0	0	2	2	1+2
Cluster 4	0	2	2	1+2	1+2

Figure 3: Timeline of the Reuther, et al. 2014 study

Cluster	Time 1	Time 2	Time 3	Time 4	Time 5
Cluster 1	1	1	1	1	1
Cluster 2	0	1	1	1	1
Cluster 3	0	0	1	1	1
Cluster 4	0	0	0	1	1
Cluster 5	0	0	0	0	1
Cluster 6	0	0	0	0	1
Cluster 7	2	2	2	2	2
Cluster 8	0	2	2	2	2
Cluster 9	0	0	2	2	2
Cluster 10	0	0	0	2	2
Cluster 11	0	0	0	0	2
Cluster 12	0	0	0	0	2

Figure 4: Timeline of the Patel, et al. 2005 study

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

Figure 5: Timeline of the Haines, et al. 2015 study

Cluster	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14
Cluster 1	1	0	0	0	0	0	0	0	2	2	2	2	2	2
Cluster 2	1	1	0	0	0	0	0	0	0	2	2	2	2	2
Cluster 3	1	1	1	0	0	0	0	0	0	0	2	2	2	2
Cluster 4	1	1	1	1	0	0	0	0	0	0	0	2	2	2
Cluster 5	1	1	1	1	1	0	0	0	0	0	0	0	2	2
Cluster 6	1	1	1	1	1	1	0	0	0	0	0	0	0	2

Figure 6: Timeline of the Chinbuah, et al.[9] study

Cluster	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7	Time 8
Cluster 1	0	0	0	1	1	1	1	1
Cluster 2	0	0	0	1+2	1+2	1+2	1+2	1+2
Cluster 3	0	0	0	0	1	1	1	1
Cluster 4	0	0	0	0	1+2	1+2	1+2	1+2

Figure 7: Variances of the estimates of intervention effects in 4 designs a) variances of $\widehat{\eta_1}$ b) variances of $\widehat{\eta_2}$

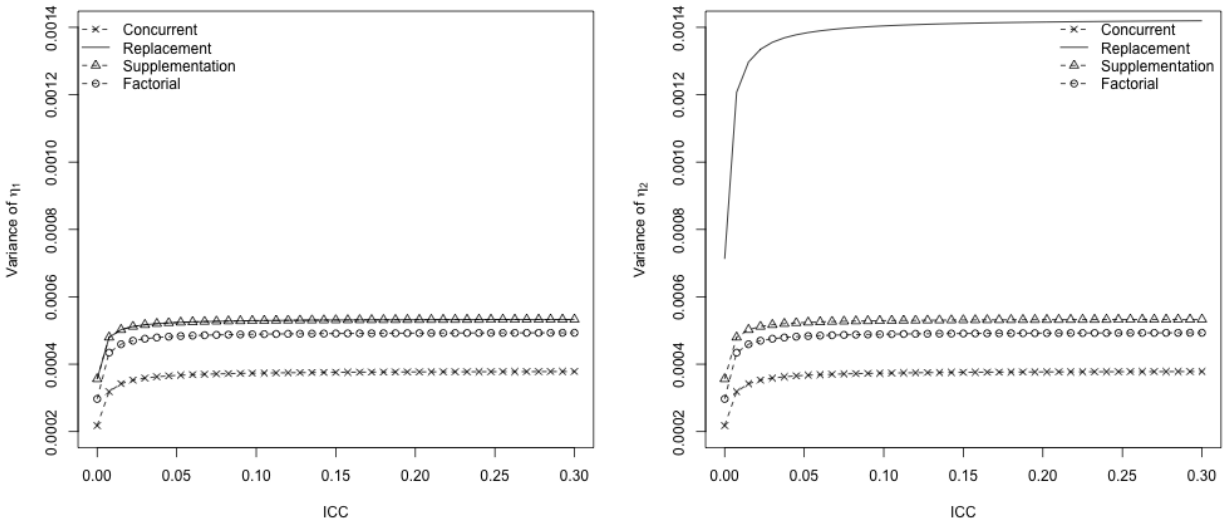


Figure 8: Variance of (either) treatment effect in concurrent design with different numbers of time intervals added at the end of the study period once all clusters have received Interventions.

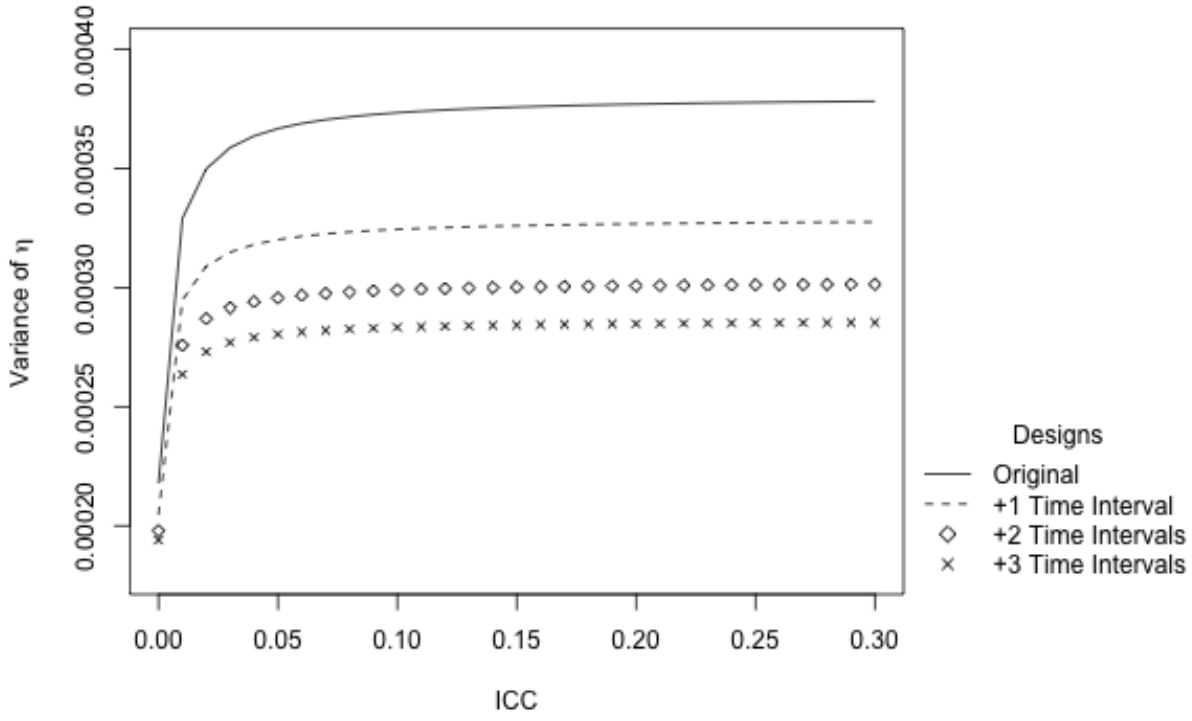
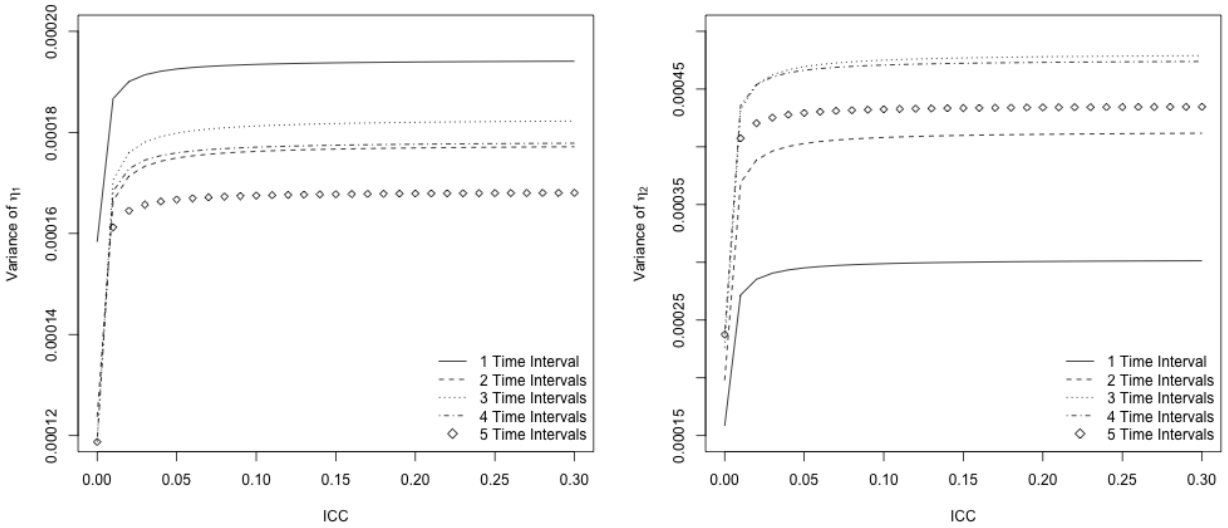
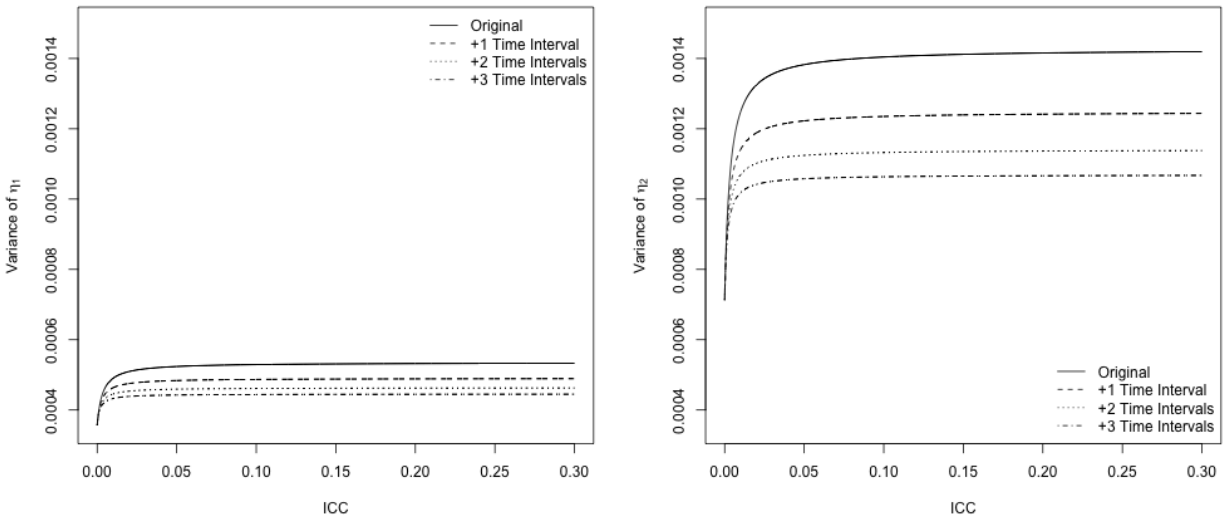


Figure 9: Comparing variances of the estimates of intervention effects for Replacement designs with different numbers of time intervals for intervention 1



a) variances of $\widehat{\eta_1}$ b) variances of $\widehat{\eta_2}$

Figure 10: Comparing variances of the estimates of intervention effects in designs with different numbers of time intervals for intervention 2 added to the end of the original design.



a) variances of $\widehat{\eta_1}$ b) variances of $\widehat{\eta_2}$

Figure 11: Comparing variances of the estimates of intervention effects for Supplementation designs with different numbers of time intervals for intervention 1

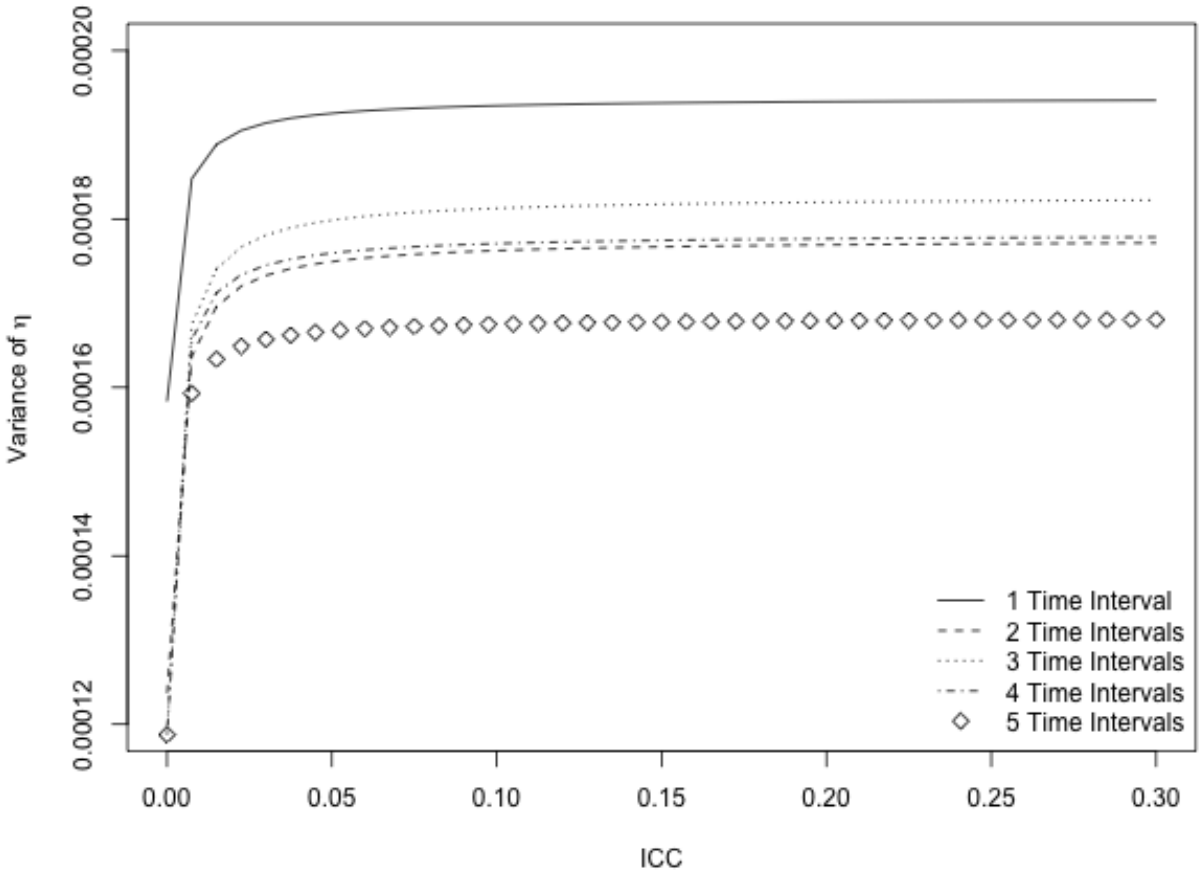
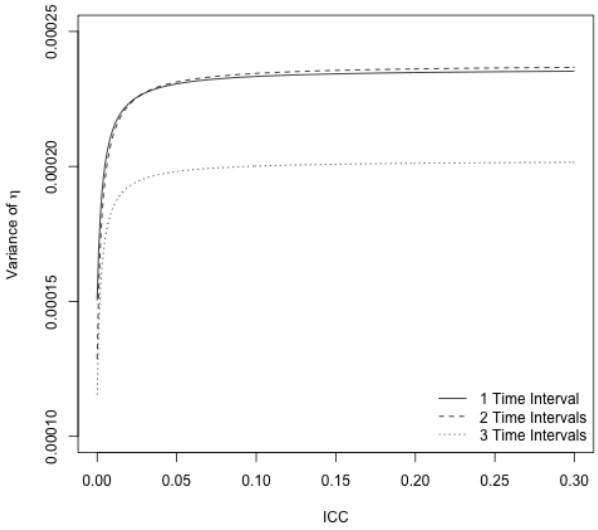
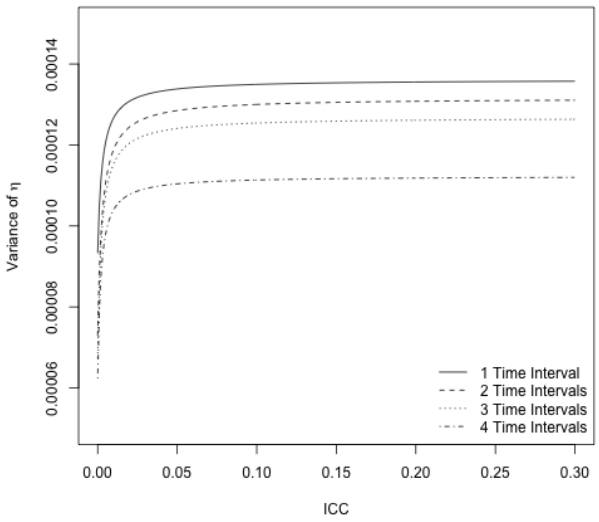


Figure 12: Factorial design with 6 vs 8 clusters



(a) For 6 clusters



(b) For 8 clusters

Supplemental Material:
Figure A1: Effect of additional time periods added to the end of the standard stepped wedge designs with single intervention.

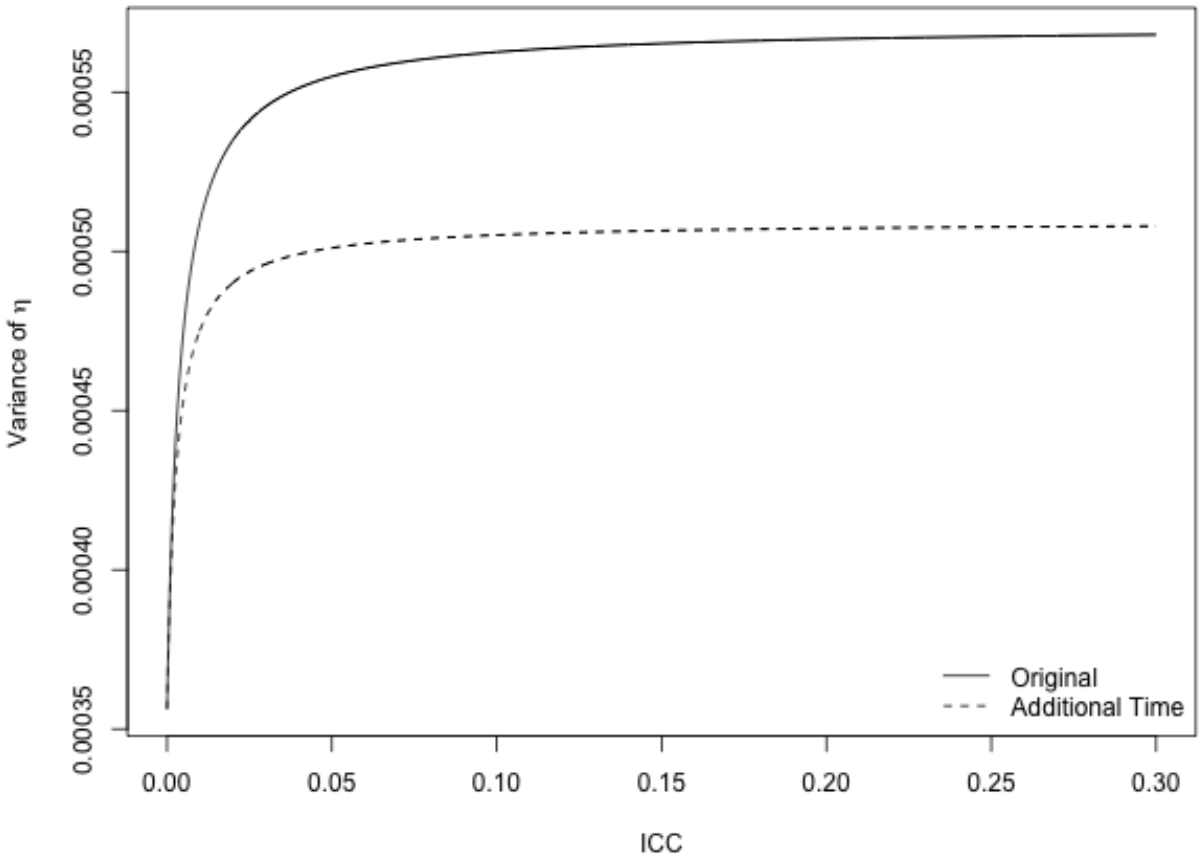
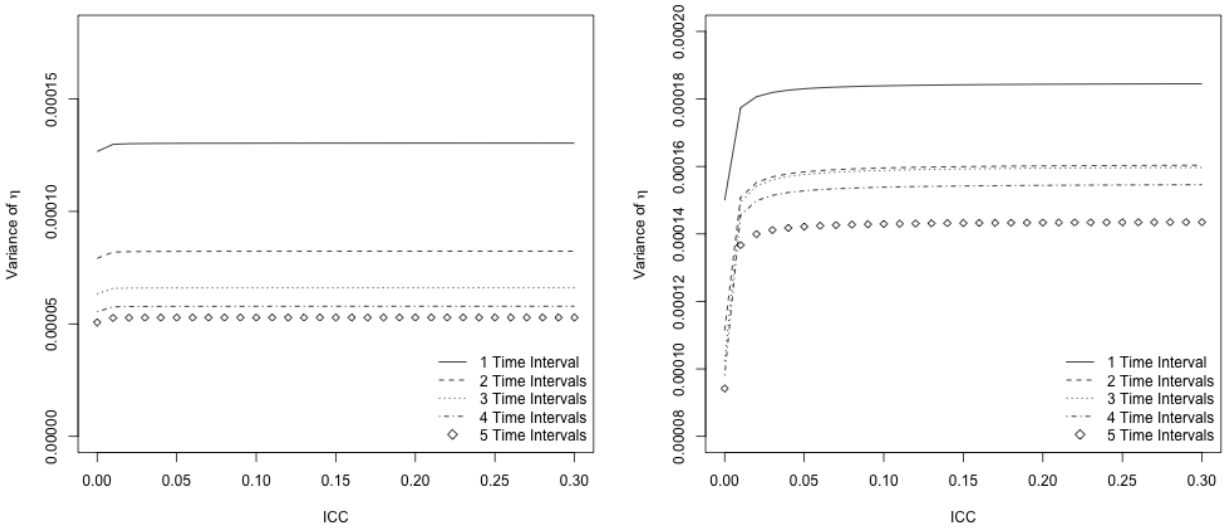


Figure A2: Investigation of the differences in designs with different types of time effect.



a) Design without time effect b) Design with linear time effect

Figure A3: Variation of Model 3 with phased study ending

	Time 0	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7	Time 8	Time 9	Time 10
Cluster 1	0	1	1	1	1	1	1	1+2	1+2	1	0
Cluster 2	0	0	1	1	1	1	1+2	1+2	1	0	0
Cluster 3	0	0	0	1	1	1+2	1+2	1	0	0	0
Cluster 4	0	0	0	0	1+2	1+2	1	0	0	0	0
Cluster 5	0	0	0	0	1+2	1+2	1	0	0	0	0
Cluster 6	0	0	0	2	2	1+2	1+2	2	0	0	0
Cluster 7	0	0	2	2	2	2	1+2	1+2	2	0	0
Cluster 8	0	2	2	2	2	2	2	1+2	1+2	2	0

FigureA4: Variation of Model 2 with phased study ending*

	T0**	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
Cluster 1	0	1	1	1	1+2	1+2	1+2	1	1	1	0	0	0	0	0	0
Cluster 2	0	0	1	1	1	1+2	1+2	1+2	1	1	1	0	0	0	0	0
Cluster 3	0	0	0	1	1	1	1+2	1+2	1+2	1	1	1	0	0	0	0
Cluster 4	0	0	0	0	1	1	1	1+2	1+2	1+2	1	1	1	0	0	0
Cluster 5	0	0	0	0	0	1	1	1	1+2	1+2	1+2	1	1	1	0	0
Cluster 6	0	0	0	0	0	0	1	1	1	1+2	1+2	1+2	1	1	1	0

*This phased study end could also be used with Model 1 if the 1+2 intervention was replaced with Intervention 2 from Model 1.

**T = Time