Contemporary Clinical Trials 45 (2015) 55–60

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Current issues in the design and analysis of stepped wedge trials

James P. Hughes [a,b,*], Tanya S. Granston [a], Patrick J. Heagerty [a]

[a] Department of Biostatistics, University of Washington, Seattle, WA, United States
[b] Fred Hutchinson Cancer Research Center, Seattle, WA, United States

ARTICLE INFO

Article history:
Received 10 April 2015
Received in revised form 24 June 2015
Accepted 5 July 2015
Available online 3 August 2015

Keywords:
Stepped wedge design
Cluster randomized trial
Implementation research

ABSTRACT

The use of stepped wedge designs in cluster-randomized trials and implementation studies has increased rapidly in recent years but there remains considerable debate regarding the merits of the design. We discuss three key issues in the design and analysis of stepped wedge trials — time-on-treatment effects, treatment effect heterogeneity and cohort studies.

1. Introduction

The use of stepped wedge designs in cluster-randomized trials and implementation studies has become increasingly common in recent years. In the first comprehensive literature review on this topic, [4] found 12 references to papers incorporating stepped wedge designs. Four years later, [21] reported 25 citations. More recently, an informal PubMed search encompassing just the 7 month period August, 2014–February, 2015 found 18 studies reporting use of the stepped wedge design, as well as additional methodological papers. In spite of this rapid increase in use, there remains considerable debate regarding the merits of the stepped wedge design.

Fig. 1 illustrates a classic stepped wedge cluster randomized design compared to a standard parallel cluster randomized trial design. In general, a stepped wedge cluster randomized trial is any design in which the clusters cross over unidirectionally (in a randomized order) from the control or standard of care condition to the intervention condition in a staggered fashion such that the intervention effect is partly, but not completely, confounded with time. Outcome data collection must be synchronized in time between clusters. See [14] Figures 1–3 for some variations of this design. Note that this relatively broad definition of the stepped wedge design includes some designs that [19] include as variations of parallel designs. Nonetheless, we choose this definition to emphasize the unique features of this design — the unidirectional crossover and the partial confounding of intervention effect and time. Note that a before–after trial does not qualify as a stepped wedge design since the intervention effect is completely confounded with time in that design.

An informative series of papers and letters [21,19,22,18,13,20,14] lays out many of the strengths and weaknesses of the stepped wedge design. Table 1 summarizes key issues raised in these papers and others. In this manuscript we provide more detailed comments on three key issues relevant to the design and analysis of stepped wedge trials — delayed treatment effects, heterogeneous treatment effects, and cohort studies.

2. Power calculation and analysis of stepped wedge trials

In this section we introduce notation and briefly review standard approaches for analyzing data and computing power in stepped wedge trials based on repeated cross-sectional samples (see Section 5 for remarks on cohort-based trials).

Any analysis of data from a stepped wedge design must account for the intentional confounding of time and treatment as well as the correlation between repeated observations in the same cluster. Model based approaches include generalized linear mixed models (glmm) [16] or generalized estimating equations (gee) [25].

[16] propose the following model for data from a stepped wedge trial:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + X_{ij}\theta + e_{ijk} \quad (1)$$

where Y_{ijk} denotes the response corresponding to individual k at time j from cluster i ($i = 1 \ldots I, j = 1 \ldots T, k = 1 \ldots m_{ij}$), β_j are fixed time effects corresponding to interval j ($\beta_T = 0$ for identifiability), X_{ij} is an indicator of the treatment mode in cluster i at time j, θ is the treatment effect, μ is the mean of Y_{ijk} in the control or standard of care condition at

* Corresponding author at: Department of Biostatistics, University of Washington, Seattle, WA, United States.
E-mail address: jphughes@u.washington.edu (J.P. Hughes).

http://dx.doi.org/10.1016/j.cct.2015.07.006
1551-7144/

	Parallel	Stepped Wedge				
	Time	Time				
	1	1	2	3	4	5
1	C	C	T	T	T	T
2	C	C	T	T	T	T
3	C	C	C	T	T	T
Cluster 4	C	C	C	T	T	T
5	T	C	C	C	T	T
6	T	C	C	C	T	T
7	T	C	C	C	C	T
8	T	C	C	C	C	T

Fig. 1. Schematic representation of parallel versus stepped wedge cluster randomized designs with 8 clusters. Each row represents a cluster. C = Control condition, T = Treatment condition.

time T, α_i is a random intercept for cluster i such that $\alpha_i \sim N(0, \tau^2)$, and $e_{ijk} \sim N(0, \sigma_e^2)$. Typically, $X_{ij} = 1$ if the intervention is provided in cluster i at time j and 0 otherwise. If $m_{ij} = m$ for all i and j then an analysis of the cluster-level means ($\overline{Y}_{ij}$) is possible; however, if the cluster sizes vary then an analysis of individual-level data provides increased efficiency.

An important assumption of model (1) is that the underlying time effect is the same for each cluster. This assumption could be relaxed somewhat to allow groups of clusters (strata) to have different underlying time trends by adding a stratum main effect and a strata by time interaction. However, it is not possible to allow each cluster to have a separate time trend; such a model would be unidentifiable (unless one assumes a parametric form for the time trend).

The approximate power for testing the hypothesis $H_o : \theta = 0$ versus $H_a : \theta = \theta_A$ may be determined from the formula

$$power = \Phi\left(\frac{\theta_A}{\sqrt{Var\left(\hat{\theta}\right)}} - Z_{1-\alpha/2}\right). \quad (2)$$

Table 1
Summary of key issues regarding stepped wedge designs.

Issue	Comments
Feasibility	• No need to provide intervention to many communities at once. • Allows intervention to be rolled out in systematic manner.
Social/political	• Provides intervention to all clusters/communities. • Intervention never removed once introduced Randomized. • Order of introduction of the intervention perceived as "fair".
Duration	• If step length fully adjustable, the trial duration can be fixed and steps added as convenient. • Likely longer duration than parallel design trial if steps have minimum length due to measurement of outcome data or need to wait for intervention to take effect. • Longer duration trials are more susceptible to contamination, external events.
Delayed rollout/effects	• Delayed rollout and/or intervention effect can reduce power. • Accounting for delayed rollout and estimation of time-on--treatment effects possible (see Section 3).
Size	• Requires fewer clusters than a parallel design but more measurements per cluster. • See discussion in [27,11] and [15].
Participant burden	• High, if longitudinal data collection on a cohort of participants; low, if repeated cross-sectional samples.
Power	• Insensitive to intracluster correlation. • Greater sensitivity to treatment heterogeneity (see Section 4).
Analysis	• Use modeling to separate time and intervention effects (see Section 2). • Typical methods are generalized linear mixed models and generalized estimating equations.

For the special case were the X_{ij} are all 0 or 1 and $m_{ij} = m$, a closed form formula for $Var(\hat{\theta})$ based on model (1) is given in [16]. For a "balanced" stepped wedge design (one with T time periods, I clusters (all of equal size m), all clusters starting in the control condition and $h = I/(T-1)$ (an integer) clusters crossing over from control to intervention at each time) [24] show that

$$Var\left(\hat{\theta}\right) = \frac{6(T-1)(1-\rho)\left(\sigma_e^2+\tau^2\right)(1+(mT-1)\rho)}{mIT(T-2)\left(1+\left[\frac{m(T+1)}{2}-1\right]\rho\right)} \quad (3)$$

where $\rho = \tau^2/(\sigma_e^2 + \tau^2)$. For binary outcomes, one may set $\sigma_e^2 = \mu^*(1-\mu)$. [27] provide a formula for direct calculation of sample size by deriving a design effect for balanced stepped wedge cluster randomized trials (see also [11] and [15]).

More generally, for situations were the above do not apply (i.e. X_{ij} not 0 or 1, or all m_{ij} not equal), $Var(\hat{\theta})$ may be computed using basic results from generalized least squares [5]. Specifically, let D be the design matrix from model (1) and V be the covariance matrix of Y. Then the appropriate diagonal element of $(D'V^{-1}D)^{-1}$ gives the variance of $\hat{\theta}$ (see appendix A for details).

3. Delayed treatment effects

An important consideration in designing stepped wedge studies is the potential for loss of power associated with delayed intervention effects. Specifically, if the intervention is not fully effective in the time interval in which it is introduced then substantial reductions in power are possible, even for minor delays [16]. Such delays can occur due to a slower than expected intervention rollout or due to an intrinsic lag between introduction of the intervention and its effect on the outcome.

The most effective approach to avoiding this loss of power is through careful study design. For instance, in a stepped wedge design the time intervals between steps should be long enough for the intervention to be rolled out and become fully effective (and for the outcome to be measured) within the time step in which the intervention is introduced. In some cases this may require a "wash-out" period between steps that allows enrolled individuals to provide outcome data but during which no new individuals are enrolled in the trial [14]. For example, in a clinic-based trial of approaches to providing antiretroviral (ART) medications to HIV-infected women to prevent mother-to-child transmission of HIV, [17] enrolled women during pregnancy but the endpoint (ART uptake) could occur anytime until delivery. A transition period was needed between the stepped wedge time steps to allow the women enrolled in the control condition to deliver their infants before introducing the intervention in the clinic. Women initiating clinic care during the transition period were not included in the analysis.

In other cases, there may be an intrinsic delay between provision of the intervention and realization of its effect and this may lead to an unacceptably long trial duration. For example, suppose one were interested in implementing a program of male circumcision to reduce HIV incidence in a community. Previous randomized trials [2,9,3] have shown that male circumcision reduces the risk of acquiring HIV by approximately 60% in heterosexual men. However, the effect of implementing a large-scale male circumcision program on community-wide HIV incidence (i.e. incidence among both women and men, circumcised or not) is unknown. Since implementation of a mass circumcision campaign would take time and be extremely resource intensive, an argument might be made to use a stepped wedge design to measure the program effect during rollout. Nonetheless, a stepped wedge would be a poor design choice in such a situation. The effect of male circumcision on community-level HIV incidence depends not only on direct effects (protecting the circumcised men) but also on indirect effects (protecting partners, and partners of partners, of the circumcised men by breaking chains of infection). Modeling predicts that the full effect