Journal of Clinical Epidemiology 69 (2016) 137–146

Journal of Clinical Epidemiology

Sample size calculations for stepped wedge and cluster randomised trials: a unified approach

Karla Hemming[a,*], Monica Taljaard[b,c]

[a]School of Health and Population Sciences, University of Birmingham, Birmingham B15 2TT, UK
[b]Clinical Epidemiology Program, Ottawa Hospital Research Institute, 1053 Carling Avenue, Ottawa, Ontario K1Y4E9, Canada
[c]Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada

Accepted 28 August 2015; Published online 5 September 2015

Abstract

Objectives: To clarify and illustrate sample size calculations for the cross-sectional stepped wedge cluster randomized trial (SW-CRT) and to present a simple approach for comparing the efficiencies of competing designs within a unified framework.

Study Design and Setting: We summarize design effects for the SW-CRT, the parallel cluster randomized trial (CRT), and the parallel cluster randomized trial with before and after observations (CRT-BA), assuming cross-sectional samples are selected over time. We present new formulas that enable trialists to determine the required cluster size for a given number of clusters. We illustrate by example how to implement the presented design effects and give practical guidance on the design of stepped wedge studies.

Results: For a fixed total cluster size, the choice of study design that provides the greatest power depends on the intracluster correlation coefficient (ICC) and the cluster size. When the ICC is small, the CRT tends to be more efficient; when the ICC is large, the SW-CRT tends to be more efficient and can serve as an alternative design when the CRT is an infeasible design.

Conclusion: Our unified approach allows trialists to easily compare the efficiencies of three competing designs to inform the decision about the most efficient design in a given scenario.

Keywords: Stepped wedge; Cluster randomized trial; Power; Sample size; Efficiency; Study design

1. Introduction

The parallel cluster randomized trial (CRT) is an established design for the evaluation of interventions delivered at the level of the cluster or where risk of contamination inhibits individual randomization [1,2]. In the conventional parallel CRT at the beginning of the trial, half of the clusters are randomized to the intervention and half to the control. This design may be augmented by the addition of baseline measures before randomization. We refer to this design as the parallel cluster randomized trial with before and after observations (CRT-BA) [3].

The stepped wedge cluster randomized trial (SW-CRT) is a relatively new type of cluster randomized design, but rapidly increasing in popularity [4–6]. There is usually a period of baseline data collection, in which no clusters are exposed to the intervention. Subsequently, at periodic time points called ‘‘steps,’’ one or several clusters are randomized to cross from control to intervention, whereas the remaining clusters remain in the control condition. The study continues until all clusters have crossed to the intervention arm, and there is usually a period at the end of the study in which all clusters are exposed to the intervention [7]. The SW-CRT can be viewed an extension of the cluster trial with baseline and repeated measures, but with the addition that clusters are randomized sequentially to cross from control to intervention [8].

The Devon Active Villages study [9] was a stepped wedge trial to evaluate whether a 12-week tailored community-level physical activity intervention increased the activity levels of rural communities. A total of 128 rural villages in England were randomized to receive the intervention in one of four steps. Random samples of 50 participants, assuming that 10 would respond, were taken in each village at each of five data collection periods using a postal survey. The primary outcome of interest was the proportion of adults reporting sufficient physical activity to meet internationally recognized guidelines, whereas minutes spent in moderate-and-vigorous activity per week was analyzed as a secondary outcome. The study found no effect of the

Conflict of interest: None.

* Corresponding author. Tel.: 01214142955.
E-mail address: k.hemming@bham.ac.uk (K. Hemming).

http://dx.doi.org/10.1016/j.jclinepi.2015.08.015

What is new?

- Sample size calculations for stepped wedge cluster trials are complex and design effects have been misapplied in the literature.
- We set out a coherent unified framework for determining the sample size for stepped wedge and parallel cluster trials.
- We present new formula to allow trialists to determine the cluster size needed where other design parameters are fixed.

intervention on the proportions of adults meeting guidelines, but a trend toward an increase in weekly duration of activity.

The advantages and disadvantages of the SW-CRT design have been debated in the literature using ethical, practical, and logistical considerations [10−13]. The SW-CRT is often considered the design of choice when it is logistically impractical to simultaneously rollout the intervention to half of the clusters; when stakeholders have a strong desire for all clusters to receive the intervention, perceiving it to be beneficial; and sometimes (although perhaps contentiously) when the intervention is believed to be more likely effective than ineffective. Because of the longitudinal nature of the SW-CRT, the design might be considered particularly suitable when there is a need to include time-varying covariates.

The consideration of statistical efficiency is another important factor when deciding between the designs. Although sample size methodology for parallel CRT designs is well established, reporting and methodological quality of the CRT design in general has been inadequate [14], whereas appropriate methodology for determining sample size needed in stepped wedge studies in particular is still in development. In the review of 12 stepped wedge studies between 1987 and 2005 [4], sample size calculations were reported in only five. It was not reported whether these sample size calculations allowed for the stepped wedge design. In another review of 25 stepped wedge studies [5], sample size calculations were clearly reported in only 8 of the 25 studies, and only 3 took into account clustering; again, it was not clear whether the stepped wedge design was accounted for.

One approach to determining the sample size needed under a cluster randomized design involves multiplying the sample size needed under an individually randomized trial by a ‘‘design effect’’ or variance inflation factor [15]. The design effect essentially represents the inflation over the sample size needed under individual randomization. Initial developments in sample size methodology for the SW-CRT focused on methods to determine power only [7]. Recently, a design effect for the SW-CRT was published; however, there has been some confusion over its implementation. Moreover, there has been a debate about the efficiency of this design relative to the parallel design, with some researchers claiming that the SW-CRT is more efficient [16], with others disputing this [17−19].

Hemming et al. [20] recently proposed that power calculations for the CRT and the SW-CRT be carried out using a single generic framework. Moreover, they expanded the framework to allow for designs with transition periods and multiple levels of clustering. In this article, we illustrate the application of the generic framework and present simple formulas that allow calculation of both the required number of clusters given a specified cluster size, as well as the required cluster size, given a specified number of clusters. Our specific objectives are to (1) illustrate, by example, how to implement design effects in the SW-CRT to ensure correct sample size calculations under a variety of scenarios; (2) demonstrate that the SW-CRT does not always require a smaller total sample size or smaller number of clusters than the parallel CRT; and (3) provide novel sample size methodology to allow designers to determine required cluster size in the SW-CRT, as current published design effects allow computation of the number of clusters, but not number of subjects per cluster.

2. Methods

2.1. A unified framework for designing both the stepped wedge and parallel cluster trial

Hemming et al. [20] present a unified framework for comparing the efficiencies of the SW-CRT and the parallel CRT. We adopt a similar approach here as is illustrated schematically in Figure 1. Using this framework, the relative efficiencies of the parallel CRT design, the CRT-BA, and the SW-CRT may be more easily compared. Note that, in our framework, the total cluster sizes are fixed across the designs. In the parallel CRT design, half of the clusters are randomized to the intervention and half to the control and all clusters remain in the arm to which they had been allocated throughout the duration of the study. In studies with prospective recruitment, the width of the diagram may represent the time over which the observations are accrued (or patients recruited); otherwise, it represents the total number of observations sampled from each cluster. In the CRT-BA, the design includes a period of time in which no clusters are exposed to the intervention and then a randomization point in which half of the clusters are randomized to cross to the intervention. The period of time in which no clusters are exposed (sometimes referred to as a baseline period) might be of shorter length (or contain fewer observations) than the period of time during which half of the clusters are