Key Design Considerations Using a Cohort Stepped-Wedge Cluster Randomised Trial in Evaluating Community-Based Interventions: Lessons Learnt from an Australian Domiciliary Aged Care Intervention Evaluation

Mohammadreza Mohebbi[1,*], Masoumeh Sanagou[2] and Goetz Ottmann[3]

[1]*Biostatistics Unit, Deakin University, Geelong, Australia*

[2]*Australian Radiation Protection and Nuclear Safety Agency, Lower Plenty Road, Yallambie, Victoria 3085, Australia*

[3]*Australian College of Applied Psychology, Sydney, Australia*

Abstract: The 'stepped-wedge cluster randomised trial' (SW-CRT) harbours promise when for ethical or practical reasons the recruitment of a control group is not possible or when a staggered implementation of an intervention is required. Yet SW-CRT designs can create considerable challenges in terms of methodological integration, implementation, and analysis. While cross-sectional methods in participants recruitment of the SW-CRT have been discussed in the literature the cohort method is a novel feature that has not been considered yet. This paper provides a succinct overview of the methodological, analytical, and practical aspects of cohort SW-CRTs. We discuss five issues that are of special relevance to SW-CRTs. First, issues relating to the design, secondly size of clusters and sample size; thirdly, dealing with missing data in the fourth place analysis; and finally, the advantages and disadvantages of SW-CRTs are considered. An Australian study employing a cohort SW-CRT to evaluate a domiciliary aged care intervention is used as case study. The paper concludes that the main advantage of the cohort SW-CRT is that the intervention rolls out to all participants. There are concerns about missing a whole cluster, and difficulty of completing clusters in a given time frame due to involvement frail older people. Cohort SW-CRT designs can be successfully used within public health and health promotion context. However, careful planning is required to accommodate methodological, analytical, and practical challenges.

Keywords: Clinical trials, Stepped wedge design, missing data, sample size, Cluster randomized trial.

1. INTRODUCTION

The 'stepped-wedge cluster randomised trial' (SW-CRT) is a form of unidirectional cross-over design. The randomisation occurs before the start of the trial. Usually all clusters start the trial in a control phase then sequentially cross over from the control condition to the intervention condition, until all clusters are receiving the intervention. It is a pragmatic study design which can reconcile the need for robust evaluations with contextual or logistical constraints. It has been used for the evaluation of service delivery interventions, it is particularly suited to evaluations that do not rely on individual participant recruitment.

The stepped wedge design is especially useful when the intervention is thought to do more good than harm [1-3]. The parallel designs withdraw or withhold the intervention from a proportion of participants (i.e. half on the clusters) which could be unethical. In many situation (because of the limited capacity of the research team, or logistical, financial issues or other contextual reasons) it is not possible to deliver the interventions to all participants simultaneously. In such cases a staggered implementation of the intervention making use of a SW-CRTs could be the optimal solution.

In the stepped wedge designs the intervention effect can be estimated from both between- and within-cluster comparisons, this means that the clusters act as their own controls because they are exposed to both control and intervention conditions. This can result in more statistical power compared to a parallel cluster design with the same number of measurements, which can be considered as another advantages of the stepped wedge design [4].

We identified two recent systematic reviews of SW-CRTs [5, 6]. These systematic reviews were broadly concerned with identifying the scope of stepped wedge studies, rather than being systematic reviews of quality of design, analysis and potential advantages and disadvantages of SW-CRT. The latest of these reviews [6] identified 10 protocols for SW-CRTs and 15 completed study publications. The systematics reviews highlighted the diverse areas of application of SW-CRTs as interventions focusing on public health

*Address correspondence to this author at the Building BC, Room BC4.206, 221 Burwood Highway, Burwood, VIC 3125, Australia; Tel: +61 3 9246 8993; E-mail: m.mohebbi@deakin.edu.au

E-ISSN: 1929-6029/17

promotion in developing countries, education, and improvements in housing.

When planning a SW-CRT with repeat measures within a community-based setting, one is faced with three choices in terms of study design: (1) the cohort design, (2) the cross-sectional design, and (3) a mixed cohort/cross-sectional design. In a cohort design, the same subjects within the clusters will be measured repeatedly over time. In a cross-sectional design, different subjects will be measured at specified time points. A mix of the cohort and cross-sectional designs is also possible. In the case of a mixed cohort/cross-sectional design, some participants will be followed over the course of the study period, whereas others will be replaced during the study. The chosen design has implications for the precision, sample size, and potential bias [7]. The design, sample size calculation, analysis and challenges of the cross-sectional SW-CRTs have been previously reported [1, 5, 6, 8]. But to the best of our knowledge there is no publication that focuses on these issues in the context of a cohort SW-CRT design. Based on a study employing a cohort SW-CRT undertaken by the authors, this article outlines key design considerations and challenges of employing a SW-CRT within the context of an evaluation of a domiciliary social care intervention involving older Australians.

There are a number of challenges that arise from SW-CRT designs. Some of these challenges are common to all SW-CRTs and some are specific to the cohort design. In this paper we will consider five areas in particular: design consideration, sample size and statistical power, missing data, analysis method and challenges and facilitators regarding the implementation of cohort SW-CRTs. The article is organised in the following fashion: In Sections two to five design, analytical issues related to SW-RCTs are described. In section six we illustrate some of the practical implications of applying a cohort SW-CRT within an aged care context making use of a recent study conducted by the authors to highlight enablers and challenges. The study was based on a cohort SW-CRT to evaluate an intervention designed for older people receiving domiciliary aged care. A detailed description of this study is beyond the scope of this article and has been published elsewhere [9]. Section seven provides a discussion and conclusion of the main points raised in the article.

2. DESIGN CONSIDERATIONS

In general terms, there are two categories of cluster trials: parallel and staggered cluster designs (see Figure **1**). In the parallel cluster randomised trial, clusters are randomised either to the intervention or control arm at the beginning of the trial and remain in that arm for the duration of the study (Figure **1a**). Control and intervention conditions are implemented concurrently. This design may be expand to include a control condition period for the cluster receiving the intervention (Figure **1b**).

In the SW-CRT, this is extended so every cluster switches from control to become exposed to the intervention there for each cluster provides pre-intervention and post-intervention, but not at the same point in time (Figure **1c**). When designing a stepped wedge cluster-randomised trial, the number of clusters, number and length of steps, and number of clusters randomised at each step need to be determined. The complete stepped-wedge design thus assumes that at each step at which a cluster switches from control to intervention condition data will be collected for all clusters. In some circumstances there are transition periods where the cluster cannot be considered as either exposed or not exposed. Some designs allow for such a transition period by not collecting data during the intervention period (Figure **1d**). The SW-CRT characteristic such as number of clusters and steps, length of steps and number of clusters at each step are usually influenced by logistical considerations. For example, the availability of eligible clusters may limit the number of clusters included. When the motivation for using the stepped wedge design is the flexibility in implementing the intervention in a staggered manner an important factor in the study design is the system's to implement the service change (Figure **1e**). The chosen design can be illustrated schematically, as in the case of the case study mentioned below (Figure **2a**).

The Case Study

The control condition consisted of community aged care as usual based on conventional case management, care facilitation, and the assistance of paid carers. The intervention consisted of a self-directed care model allowing participants to take greater control of the care they received. It was designed as an 'incremental capacity building' model (hereafter the CHOICES model) where self-direction begins at a lower level with participants taking on the development of their care plan (Level 1). To achieve this, participants are mentored by case managers. As participants become comfortable with designing their own care plan, they can assume control of care

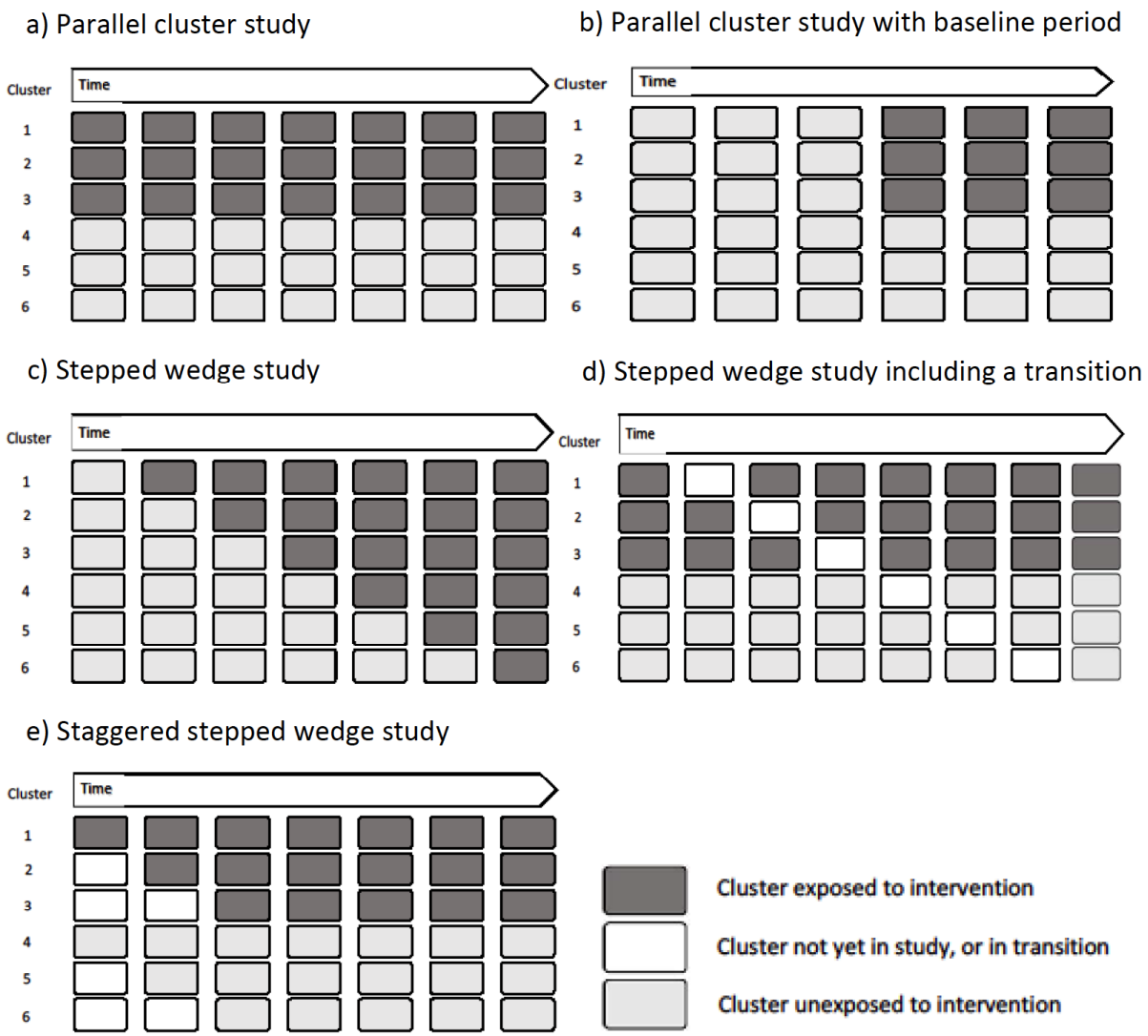
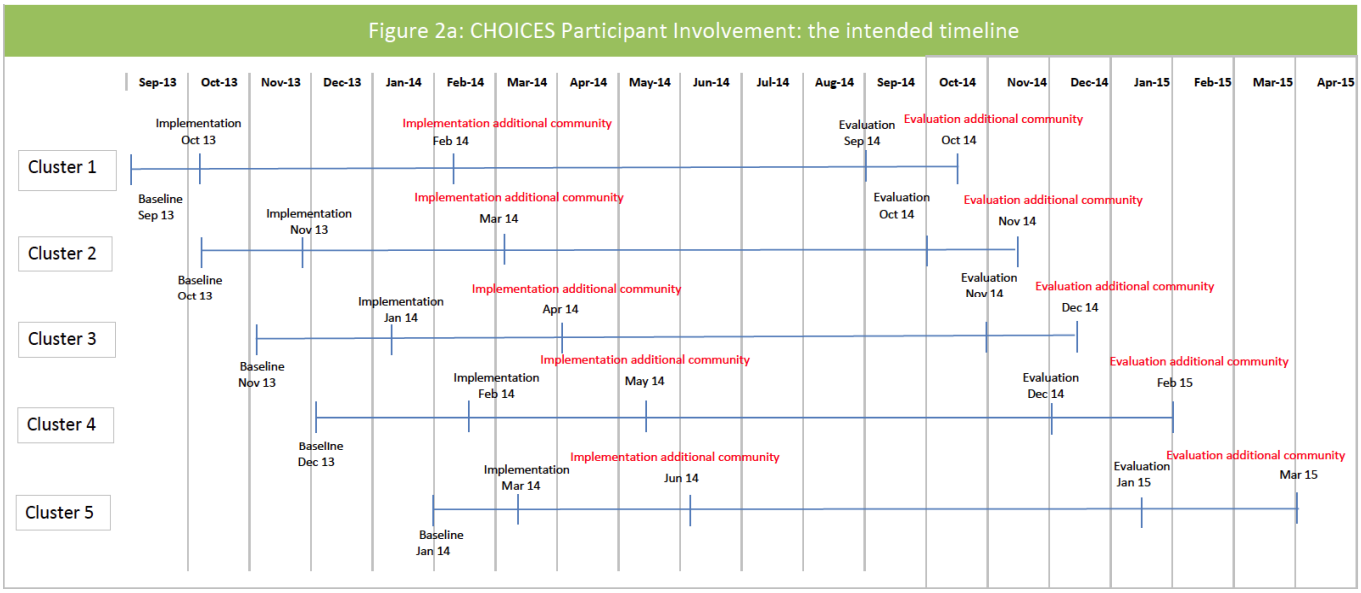


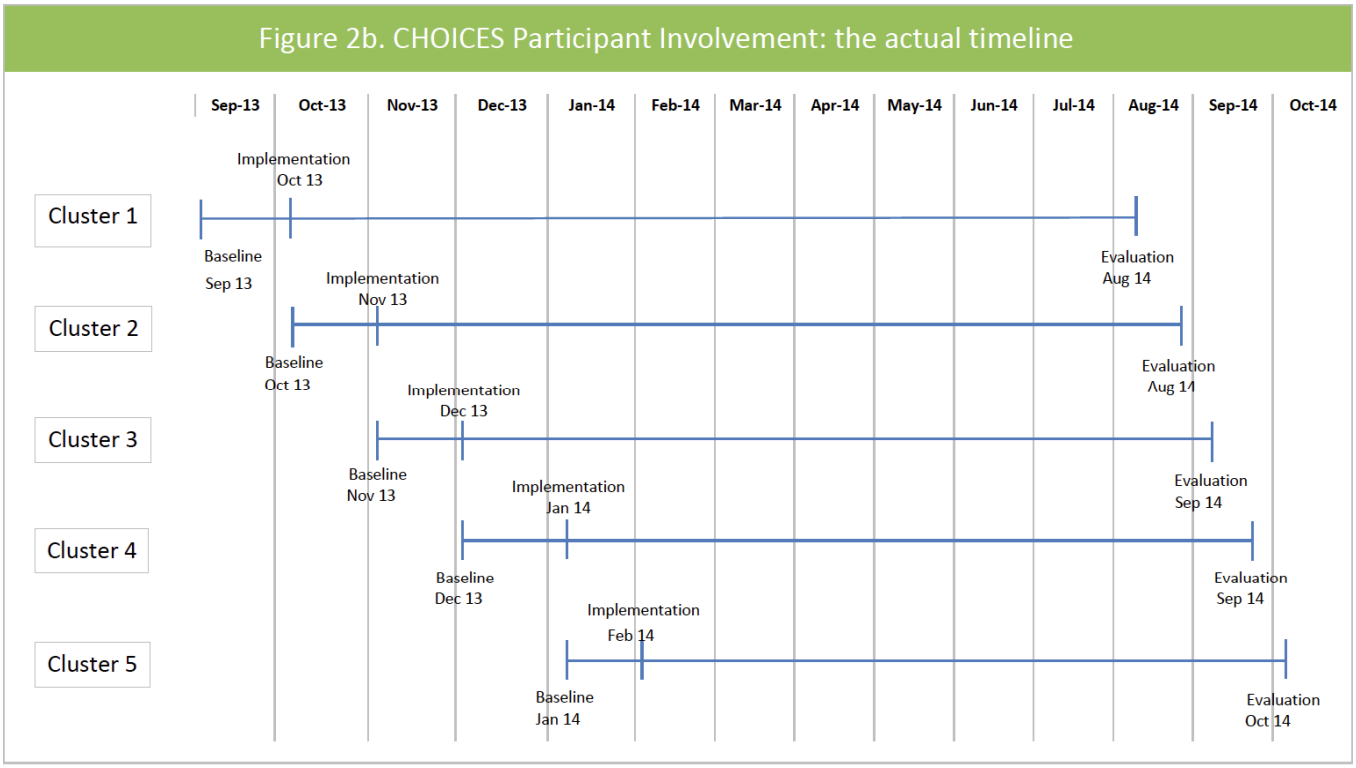
Figure 1: Comparing SW-CRTs to other cluster designs.

coordination responsibilities (Level 2). Again, participants receive the support of case managers until they feel comfortable to manage service providers. At Level 2, participants have access to comprehensive lists of service providers, their hourly rates, and the scope of services provided. Moreover, participants have access to an information pack outlining the most important services in their municipal region. Once comfortable with Level 2, they can elect to manage care services more directly assuming financial, administrative, and bookkeeping responsibilities (Level 3). In the CHOICES project, Level 3 takes the form of a voucher option with a minor cash component made available in the form of a debit card. Core services such as home and personal care are paid through a broker agency rather than directly by the client. Only peripheral services such as massages or complementary therapies were paid directly by clients.

In order to evaluate the impact of this self-directed domiciliary model of aged care intervention, we undertook a cluster-randomised, stepped-wedge controlled trial over 19 months. The design was a cluster-randomised trial in a form of cohort design with unidirectional cross-over (from control to experimental) on the same participants. The evaluation involved the measurement of the effects of the CHOICES project on clients' satisfaction with community aged care services, quality of life, subjective health, and perceived quality of care.

We partnered with seven community aged care organisations that provide services to people living in regional/rural, Greek, and indigenous communities in Victoria, Australia. Participants from each organisation were randomly assigned across five clusters. This meant that each cluster had an equal number of





Participants from one collaborating organisation experienced a shortened intervention period. The organisation supplied 38 clients.
As Figure **2b** indicates the period was shortened between 3 (cluster 1) and 2 months (most other clusters). All these participants included in the primary outcome analysis.

Figure 2: The CHOICES trial profile: **a**) the intended, and **b**) the actual timeline.

participants from each of the participating service providers. Managers provided an introduction to the study and invited eligible clients to participate in the trial. The contact detail of interested clients were, with their permission, forwarded to the research team. Clients were contacted by a team member and informed consent was obtained. Clients interested in participating in the trial from these organisations were randomly assigned to the five clusters contained 198 participants at Baseline and 137 participants at the end of the trial (Figure **2a**). The full protocol for the trial has been reported previously [9]. In order to be eligible, clients had to be in receipt of a commonwealth aged care package. At the time of implementation, aged care packages were available in the form of low-care Community Aged Care Packages (CACPs) and high-

care Extended Aged Care in the Home (EACH) packages and EACH packages with a dementia supplement (EACH-Ds). At baseline, the mean age of Regional/rural clients was 79.8 (SD 2.43), that of Greek clients 80.0 (SD 3.00) and that of Indigenous Elders 67.8 (SD 2.50). These age differences were statistically significant. The study was approved by Deakin University's ethics committee.

Randomisation Process

A total of seven organisations participated in the study. Since the study involved three distinct ethnic groups, we decided to randomly assign participants from each of the participating organisations to five clusters. We aimed to recruit a total of 200 frail older people and 198 were assigned to the five clusters. Randomisation took place around two month before the beginning of the trial. Computer-generated random numbers were used to make this allocation.

Procedures

Participating care facilitators were trained and mentored in the intervention condition starting approximately 2 months before commencement of the trial. Each cluster contained a five week data collection period (the study included one cluster with a six week data collection period taking into account delays resulting from absences due to Christmas and New Year) during which informed consent and baseline data was collected. After the initial five week period, each of the clusters was switched to intervention mode. The point in time at which the switch to intervention occurred, was the starting point for the data collection phase of the next cluster.

Changes to the Study Design

The design was subject to one major change. One service provider decided to withdraw 80% of its eligible clients at the beginning of the trial. As a result, a seventh aged care provider was recruited. Clients associated with this service provider were randomly assigned to the clusters as the intervention had already commenced. These clients experienced a shortened intervention period (See Figure **2b**). Moreover, other delays impacted on the way a number of clients experienced the intervention period. The reasons for these delays are outlined below.

3. SAMPLE SIZE

Sample sizes calculation is more complicated in clustered RCT compared with RCTs based on individual level data because instead of a single sample size the number of participants per cluster and the number of clusters should be determined. The number of clusters is an important factor since there are certain designs for which increasing the number of subjects per cluster will never achieve a required power. A major difficulties in sample size estimation is that for a continuous outcome, the variance of the outcome measure may be uncertain, and for binary data the incidence of the outcome event may be unknown. This is more difficult in cluster trials because we also do not know the intra-class correlation (ICC). The ICC defined as the ratio of the between-cluster variance to the total variance of an outcome variable, and the design effect (DE) defined as the ratio of the variance of an outcome measure when clustering is accounted for to the variance of the outcome measure when clustering is not accounted for, are two alternate approaches for presenting the clustering effect on an outcome variance. A consequence of the correlations between individuals in the same cluster is that a cluster trial will require a larger sample size than a corresponding individually randomised trial. Unlike the parallel design in a step wedge design the number of clusters exposed to the intervention increases as the study progresses unlike the parallel design in number of the exposed clusters are fixed at all-time points. This means that that in stepped wedge studies DE is no longer applicable [1]. On the other hand, pre-and-post intervention comparison within each cluster is possible in the step wedge designs which tends to reduce the impact of cluster effects. Sample size and power calculations have only been described only for cross-sectional stepped wedge designs [2, 3]. It has been showed that cross-sectional step wedge designs is always more efficient in regard to the required number of clusters [3]. The total number of required participants in stepped wedge trials depend on ICC, the number of clusters in the study, the number of observations in each cluster, and structure of the design [10]. Methods for calculation required sample size or statistical power for a cross-sectional stepped wedge have been implemented in the statistical software package Stata [11]. The methods assume equal numbers of observations per period in each cluster. Similar considerations could apply to cohort designs. In general, cohort studies are more powerful than cross-sectional studies [7]. In the SW-RCTs with cohort design both within-cluster and within-subject comparisons can be used to estimate the treatment effect as a result of repeated measurements of the same subjects. Therefore, the required sample size for

a cohort study will be smaller than for a cross-sectional study. Yet, there is no algorithm available for calculating the power or sample size in a cohort stepped wedge trial, nor implementation in a statistical package for this design.

4. ANALYSIS

Participants and clusters' characteristics should be summarised by intervention status to evaluate selection biases and lack of balance. If possible these characteristics can be compared by randomisation group per steps. This should include the numbers analysed, the average cluster size, cluster characteristics, and important participant characteristics.

Both conditional and subject-specific models has been used to account for clustering effect in the SW-RCTs [5, 6]. Conditional models use random effects to reflect the correlation among individuals within the same cluster; the individuals are assumed independent conditional on those random effects [12]. Marginal or population-averaged models define the marginal expectation of the response variable as a function of the predictor variables and assume that the variance is a known function of the mean; a correlation structure for within cluster individuals should be specified [13]. For normally distributed linear models interpretation of the regression coefficient for the intervention effect is the same in conditional and marginal models; however, for other members of the exponential family distributions (e.g. binary and count responses), the intervention effect from a marginal model is smaller than that from a conditional model and has a different interpretation [14]. In cohort and mixed designs, within individual correlation resulted from multiple measurements of the same participants over the course of the study should also be considered. This introduces an additional randomisation effect at the level of the individual in the model. A possible option for these designs are multilevel models [15]. Ignoring a level of clustering in the data and conducting simpler regression analysis would generally result in unbiased estimates of the fixed effects. However, the standard errors of all fixed effect coefficients are biased downwards [16, 17]. More specifically ignoring one level of data hierarchy in models for cohort or mixed designs SW-RCTs increase the risk of Type I statistical error. This underestimation bias is more problematic for non-linear models than for linear models. Results of a simulation study showed that with an average of only five observations per group, valid and reliable estimates of all parameters can be obtained when using two-level models with either a continuous or a discrete outcome [18]. Calendar time is a potential confounder in SW-RCTs and should be adjusted for in the analysis. Dealing with calendar time as a fixed effect and including a period effect parameter in the model to adjust for the calendar time has been proposed [1]. The estimated ICC and period effect from the model are not key issues in the interpretation of the intervention effect but reporting them are recommended [1] as they are helpful for designing future trials and evaluating underlying confounding effects of calendar time.

Consideration should be given to intention-to-treat analysis. ITT includes participants who did not get the intended treatment or who deviated from the trial protocol in the analysis [19]. This approach more closely reflects a real-world situation as it provides an unbiased estimate of the effect of intervention due to the loss of participants [20]. Strategy for intention to treat analysis in randomised trials with missing outcome data has been discussed and a four point framework for dealing with incomplete observations has been proposed [21]. This include: I. attempt to follow up all randomised participants, II) accounting for missing data in the main analysis, III) examining departures from the assumption about missingness made in the main analysis by performing sensitivity analyses, and IV) accounting for all randomised individuals, at least in the sensitivity analyses.

Analysis of the CHOICES Study

ITT analysis was performed in which all participants with missing follow-up data were included in the analysis. Primary analyses compared the following outcomes between the intervention and the control period: cognitive issues, access to additional funding (such as armed services subsidies), average years of receiving packaged care, accommodation type, and highest level of education.

All analysis accounted for hierarchical nature of the design. Two levels of hierarchy were considered: level one within participant repeated measures nested within the individuals, and level 2: participant-level data that were clustered within the three distinct participating communities. Hierarchical linear mixed models were used as the primary choice for main outcome. Linear mixed models using GEE technique with a compound summary working variance-covariance structure for within individual repeated measures that ignored the community aged care clustering effect were also

Table 1: Intervention Estimation Comparison for Sf-12 Domains Using Independent Sample t-Test, Two-way ANOVA, ANCOVA and GEE

SF12 domains	Control period		Intervention period		Model1[1]	Model2[2]	Model3[3]	Model4[4]
	Mean	SD	Mean	SD	Intervention (SE)	Intervention (SE)	Intervention (SE)	Intervention (SE)
Vitality	3.524194	0.094056	3.548387	0.098743	0.02 (0.09)	0.03 (0.13)	0.02 (0.08)	0.03 (0.09)
Physical functioning	2.806452	0.099617	2.806452	0.108441	0.00 (0.10)	0.00 (0.08)	0.00 (0.09)	0.00 (0.10)
Physical role functioning	5.926829	0.213864	5.512195	0.235004	-0.42 (0.24)	-0.37 (0.31)	-0.39 (0.22)	-0.39 (0.24)
General health perceptions	7.622951	0.199011	7.57377	0.229842	-0.05 (0.25)	-0.15 (0.30)	-0.04 (0.22)	-0.12 (0.25)
Bodily pain	3.08871	0.12387	3.08871	0.119011	0.00 (0.12)	-0.03 (0.16)	0.00 (0.10)	-0.01 (0.11)
Emotional role functioning	9.790323	0.158786	9.943548	0.160441	0.15 (0.19)	0.13 (0.21)	0.14 (0.17)	0.13 (0.19)
Social role functioning	3.386555	0.127486	3.470588	0.128993	0.08 (0.14)	0.10 (0.17)	0.09 (0.12)	0.09 (0.14)
Mental health	10.74797	0.149009	10.73984	0.183933	-0.01 (0.21)	-0.02 (0.22)	0.00 (0.18)	-0.02 (0.20)

[1]Independent sample t-test; [2]Two-way ANOVA, [3]ANCOVA, [4]GEE.

implemented and the results were compared with the hierarchical linear mixed models to evaluate necessity of an additional random effect in the models. Stata statistical software (Release 13) were used for data analysis. All primary analyses were adjusted for covariates comprising the calendar time in which the participants received the intervention (treated as a fixed categorical variable), mean age, cognitive issues, access to additional funding (such as armed services subsidies), average years of receiving packaged care, accommodation type, and highest level of education.

5. THE CHOICES CASE STUDY

Out of a total of approximately 2000 eligible clients, 198 agreed to participate and were randomised into five clusters (approx. 40 to each cluster). For all 198 participants, valid baseline data was collected. For each cluster, there were two assessment points: at Base Line (BL) approximately 1 month prior to the commencement of the intervention; and at the end of the trial (T2). The four instruments used to measure the impact of the intervention were the Adult Social Care Outcomes Toolkit, the Personal Wellbeing Index, a measure of satisfaction with the quality of care, and the 12-Item short Form health survey for measuring quality of life (SF-12) [22]. In order to compare different models properties in estimating intervention impact we focused on SF12 outcome for illustration purposes. The SF-12 measures 8 health domains as listed in Table **1**. Four models has been illustrated for comparison: independent sample t-test ignoring centres clustering effect and time trend (model 1); two-way ANOVE accounting for centres effects as a fix factor (model 2); ANCOVA accounting for centres effects as a fix factor and adjusting for relevant baseline SF-12 domain as a covariate (model 3) and GEE accounting for centres effects as a fix factor and adjusting for relevant baseline SF-12 domain by implementing within participant auto-correlation in a exchangeable variance-covariance structure. Over the course of the 10 month intervention, subjectively assessed health measures remained very stable, and all four models achieved similar point estimate and standard error (SE). Model 3 (ANCOVA) has been performed better in the term of SE estimation, followed by GEE and independent sample t-test.

6. MISSING DATA

In cluster trials there are two aspects to missing data: missing outcomes from individuals and missing clusters. Because data are usually positively correlated within a cluster, losing an individual in a cluster has less impact than if it were an individually randomized trial. However, losing a whole cluster has a large impact on the outcome. Missing data can reduce the power and efficiency of a study. They can also lead to biased results. If missingness of the outcome is unrelated to observed or unobserved characteristics of the individuals, the missing data are called Missing Completely At Random (MCAR). The complete case analysis for MCAR data, twill result in unbiased estimation of intervention effect but there will be loss of efficiency [23]. Missingness is considered to be Missing At Random (MAR) if missing data are related to an observed variables, but it is not related to the value of

the variable that has missing data. Multiple imputation and model-based approaches [24] are valid and unbiased methods for MAR data, as long as the models are specified correctly. The statistical literature is rich with methods for handling incomplete data [21, 25-29].

7. STRENGTHS AND WEAKNESSES OF CHOICES SW-CRT

When the stepped wedge design is compared with other designs, there are several advantages and disadvantages of choosing such a design. The aim of the present article was to determine the advantages and disadvantages of a stepped wedge cluster randomized design for a specific clinical application.

Disadvantages

Rigidity of Research Design

A SW-RCT is a relatively rigid design. Once participants are randomly assigned to clusters and once the participants of the first cluster have shifted to the intervention, it is difficult to make changes to the research design. However, in our case the research design had to be adapted to the fact that one organisation partially withdrew from the project considerably reducing the pool of potential participants. As a result, an additional organisation had to be incorporated in order to ensure that the study had the desired power. We did this by distributing participants recruited through this additional provider across all five clusters. However, this limited the intervention period experienced by these participants by two to three months (Figure **2b**). There is a gap in the research literature and there is currently no body of research that could assist researchers in dealing with such challenges.

Also, controlling the implementation of a SW-RCT involving frail older people can be challenging. Older people tend to face health problems and require hospital stays, go on holidays or respite breaks, and are generally very busy. This can lead to delays in the data collection process creating a condition where the data collection periods of clusters start to overlap. The rigidity of the SW-RCT makes it difficult to deal with such and other delays. Additional delays that we experienced were linked to the difficulty of

- obtaining ethics approval from a participating organisation,
- completing the recruitment process in time,
- completing the informed consent process, and
- obtaining data during or immediately after the Christmas holiday period.

Bearing this in mind, it would be advisable to plan for a longer data collection periods of six to eight weeks. However, this would increase the length of the trial. In our case this would have resulted in a trial approximately two years in length. This would not have been feasible.

Multiple Repeat Measures

In theory, the SW-RCT design requires less participants to arrive at statistically significant outcomes because the design incorporates repeat measures of the independent variables at the beginning of each new step throughout the intervention. In our case, this would have meant that participants would have had to be contacted every five weeks. It was clear from the outset that this was not a viable option. The data collection pack included four validated survey tools consisting of a total of 64 questions. In light of our experience collected during previous studies involving frail older people [30], we were certain that our participants would not be willing to complete the same survey tools six times (indeed, the research assistants found it difficult to convince a significant minority of participants to complete the survey tools twice, not to mention six times). Also, given the difficulty collecting data from a sample comprised of frail older people, it would have been logistically impossible to collect data at six time points. Bearing this in mind, we decided to forgo the increased power associated with multiple repeat measures in favour of a larger sample size, to make up for this loss.

Attrition

The stepped nature of the SW-RCT can result in a very long trial period. However, involving frail older people in long trials is problematic as a longer trial period will increase the attrition rate - particularly for the last clusters. In our case, clusters four and five ended up with considerably fewer participants. Whereas clusters one, two, and three, contained 30 and 31 participants, cluster four and five contained 23 and 21 participants respectively.

Integration of Qualitative and Quantitative Design Elements

If SW-RCTs are used to evaluate complex interventions, it is likely that they will comprise mixed

methods approaches involving qualitative and quantitative elements. However, the stepped design of a SW-RCT restricts the ways in which qualitative and quantitative strands can be integrated. In particular, an emergent, fully integrated design would be difficult to achieve as such an integration would affect each of the clusters in a different fashion, undermining the integrity of the research design. As a result of this constraint, a sequential integration where qualitative and quantitative methods are deployed independently in a given sequence or a parallel integration where qualitative and quantitative strands are conducted independently alongside each other are the most likely candidate for mixed methods approaches involving SW-RCT [31].

Advantages

The SW-CRT represents a viable alternative to conventional Cluster-Randomised Trials when the recruitment of a control group would be difficult or impossible. For example, the pending introduction of policy reforms create an operational context in which health and social care service providers would have been unwilling to provide a control group. Having to introduce organisational change may make it strategically undesirable to delay the implementation of the intervention, even by a two or three of months. A SW-CRT can potentially overcome this difficulty as its stepped design coincides with the gradual way many organisations implement change. The CHOICES study took advantage of this point. The logistic difficulties of implementing the intervention at seven sites at once was another considerations to choose a stepped wedge design for the CHOICES study.

The SW-CRT design is considered more ethical than parallel cluster RCT designs when the intervention is believed to do more good than harm [5]. In our case, there was enough evidence to support the intervention under study was not a burden to the participants associated with an increase of costs, there were also evidence to support that the intervention improve participants' satisfaction with community aged care services and quality of life.

Another general advantage of the stepped wedge design is that it always requires fewer number of clusters compared to an equivalent parallel design as the design is relatively insensitive to plausible variations of the ICC [10]. This advantage was not one of the considerations when choosing a stepped wedge design for the CHOICES trial.

8. FINAL RECOMMENDATIONS AND CONCLUSION

This paper has outlined a number of challenges faced by investigators implementing and analysing SW-CRTs. The general advice is that if you can use a parallel randomised design avoid using SW-CRTs [32]. However there are instances when a stepped wedge design is the only acceptable alternative [33]. For example, this was the case for the CHOICES study. While the cohort SW-CRT design generated a number of difficulties such as the difficulty to control the data collection phases and the associated switching point, a SW-CRT was the only viable option as service providers were not prepared to hold off implementing a model that was likely to translate into policy. Thus, we were unable to find a control group for the duration of the trial.

Reporting guidelines specific to SW-RCTs do not exist but the Consort 2010 extension to cluster randomised trials [34] and Consort extension to pragmatic trials [35] should be used as guidelines. Hemming *et al*. recommend some minor additions and modifications the Consort 2010 cluster extension for reporting of stepped wedge cluster randomised trials [1].

We believe the following recommendations should be addressed in designing, analysing, and reporting these studies.

Justify the practical aspects and study circumstances for using the SW design. The reasons for choosing a SW-RCT design should be explicitly stated. Although they are more problematic than parallel designs, it should be clear that SW designs are sometimes the only option available for addressing the research question. Ethical reasons have been mentioned as a motivation to choose the stepped wedge design rather than a parallel RCT. This statement is true at the cluster level but not necessary at the subject level. In cross-sectional SW-RCTs similar to parallel designs only a fraction of all subjects will receive the intervention. It is only in the cohort SW designs that all participants will switch to intervention. For this reason if there is evidence about effectiveness of the intervention or if there is a general believe that the intervention benefits overweight the risks, one should carefully think about whether an interventional study to obtain additional evidence is needed. Logistical, practical and/or financial issues have been mentioned as reasons to use a stepped wedge design. When an intervention cannot be implemented

simultaneously to all clusters the staggered implementation of intervention through SW-RCT could be a solution.

Randomise the order of receiving interventions in clusters to avoid bias. In addition when the outcome is measured at individual level lack of concealment of intervention implementation date could lead to differential selection of participants between per- and post-implementation periods.

Allow for clustering in the analysis. When a SW design is used, it is important that the analysis addresses the design appropriately. Systematic reviews suggest that there are still problems with analysing data from WE-CRTs [5, 6]. Performing an individual-level pre- and post-implementation does not address the design characteristics. The clustering as the unit of randomisation cannot be ignored also the confounding effect of the calendar time should be accounted for in the analysis. Report ICC and components of variance in publication. Sample size calculation for Stepped wedge designs is currently difficult by the lack of information about the magnitude of the ICC.

Adjust for confounding. Stepped wedge designs are not parallel randomised designs, so potential confounders will not adjust through random assignment of the intervention(s) to participants. In addition randomisation of the order on receiving intervention is not at individual level so the effect of confounding factors at the individual and cluster levels does not balance out by randomisation. If there are important confounding variables at individual and/ or cluster level appropriate regression models should be implemented. It is important to select appropriate methods of modelling to adjust for potential confounders as analysis at the cluster level tends to attribute group characteristics to individuals (ecological fallacy) [36, 37]. Multilevel models explicitly model the association of observations with clusters [38, 39], while generalized estimating equations treat it as a nuisance variable [13]. Both methods may require a fairly large number of clusters [40, 41]. It is a weakness of stepped wedge designs the intervention effect is partially confounded with calendar time. The calendar time should be adjusted as a confounder factor in the analysis.

REFERENCES

[1] Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ 2015; 350. https://doi.org/10.1136/bmj.h391

[2] 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): 181-96. https://doi.org/10.1002/sim.6325

[3] Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials 2007; 28(2): 182-91. https://doi.org/10.1016/j.cct.2006.05.007

[4] Woertman W, de Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL, Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. J Clin Epidemiol 2013; 66(7): 752-8. https://doi.org/10.1016/j.jclinepi.2013.01.009

[5] Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. BMC Med Res Methodol 2006; 6: 54. https://doi.org/10.1186/1471-2288-6-54

[6] Mdege ND, Man MS, 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): 936-48. https://doi.org/10.1016/j.jclinepi.2010.12.003

[7] Feldman HA, McKinlay SM. Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model. Stat Med 1994; 13(1): 61-78. https://doi.org/10.1002/sim.4780130108

[8] Zhan Z, van den Heuvel ER, Doornbos PM, Burger H, Verberne CJ, Wiggers T, de Bock GH. Strengths and weaknesses of a stepped wedge cluster randomized design: its application in a colorectal cancer follow-up study. J Clin Epidemiol 2014; 67(4): 454-61. https://doi.org/10.1016/j.jclinepi.2013.10.018

[9] Ottmann G, Millicer A, Bates A. CHOICES in Community Aged Care: Final Report, Uniting Care Life Assist/Deakin University Research Partnership, QPS, Glen Waverley 2015.

[10] de Hoop E, Woertman W, Teerenstra S. The stepped wedge cluster randomized trial always requires fewer clusters but not always fewer measurements, that is, participants than a parallel cluster randomized trial in a cross-sectional design. In reply. J Clin Epidemiol 2013; 66(12): 1428. https://doi.org/10.1016/j.jclinepi.2013.07.008

[11] Hemming K, Girling AJ. A menu-driven facility for power and detectable difference calculations in stepped-wedge randomized trials. The Stata Journal 2014; 14(2): 363-380.

[12] Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 1977; 72(3): 320-338. https://doi.org/10.1080/01621459.1977.10480998

[13] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986; 73: 13-22. https://doi.org/10.1093/biomet/73.1.13

[14] Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health 2004; 94(3): 423-32. https://doi.org/10.2105/AJPH.94.3.423

[15] Rice N, Leyland A. Multilevel models: applications to health data. J Health Serv Res Policy 1996; 1(3): 154-64.

[16] Merlo J, Lynch JW, Yang M, Lindstrom M, Ostergren PO, Rasmusen NK, Rastam L. Effect of neighborhood social participation on individual use of hormone replacement therapy and antihypertensive medication: a multilevel analysis. Am J Epidemiol 2003; 157(9): 774-83. https://doi.org/10.1093/aje/kwg053

[17] Merlo J. Multilevel analytical approaches in social epidemiology: measures of health variation compared with traditional measures of association. J Epidemiol Community Health 2003; 57(8): 550-2. https://doi.org/10.1136/jech.57.8.550

[18] Clarke P. When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. J Epidemiol Community Health 2008; 62(8): 752-8. https://doi.org/10.1136/jech.2007.060798

[19] Lachin JM. Statistical considerations in the intent-to-treat principle. Control Clin Trials 2000; 21(3): 167-89. https://doi.org/10.1016/S0197-2456(00)00046-5

[20] Abraha I, Cherubini A, Cozzolino F, De Florio R, Luchetta ML, Rimland JM, Folletti I, Marchesi M, Germani A, Orso M, Eusebi P, Montedori A. Deviation from intention to treat analysis in randomised trials and treatment effect estimates: meta-epidemiological study. BMJ 2015; 350. https://doi.org/10.1136/bmj.h2445

[21] Bell ML, Kenward MG, Fairclough DL, Horton NJ. Differential dropout and bias in randomised controlled trials: when it matters and when it may not. BMJ 2013; 346: e8668. https://doi.org/10.1136/bmj.e8668

[22] Ware JE Jr, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Medical Care 1996; 34(3): 220-233. https://doi.org/10.1097/00005650-199603000-00003

[23] Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient-reported outcomes. Stat Methods Med Res 2014; 23(5): 440-59. https://doi.org/10.1177/0962280213476378

[24] Little R, Rubin D. Statistical Analysis With Missing Data, Wiley, Hoboken, NJ 2002. https://doi.org/10.1002/9781119013563

[25] Taljaard M, Donner A, Klar N. Imputation strategies for missing continuous outcomes in cluster randomized trials. Biometrical Journal 2008; 50(3): 329-45. https://doi.org/10.1002/bimj.200710423

[26] Ma J, Akhtar-Danesh N, Dolovich L, Thabane L. Imputation strategies for missing binary outcomes in cluster randomized trials. BMC Med Res Methodol 2011; 11: 18. https://doi.org/10.1186/1471-2288-11-18

[27] DeSouza CM, Legedza AT, Sankoh AJ. An overview of practical approaches for handling missing data in clinical trials. J Biopharm Stat 2009; 19(6): 1055-73. https://doi.org/10.1080/10543400903242795

[28] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009; 338. https://doi.org/10.1136/bmj.b2393

[29] O'Neill RT, Temple R. The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it. Clin Pharmacol Ther 2012; 91(3): 550-4. https://doi.org/10.1038/clpt.2011.340

[30] Ottmann G, Mohebbi M. Self-directed community services for older Australians: a stepped capacity-building approach, Health Soc Care Community 2014; 22(6): 598-611. https://doi.org/10.1111/hsc.12111

[31] Teddlie C, Abbas T. Overview of contemporary issues in mixed methods research. Sage handbook of mixed methods in social and behavioral research 2010; pp. 1-44. https://doi.org/10.4135/9781506335193.n1

[32] Kotz D, Spigt M, Arts IC, 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(12): 1249-52. https://doi.org/10.1016/j.jclinepi.2012.06.004

[33] Mdege ND, Man MS, Taylor nee Brown CA, Torgerson DJ. 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(12): 1253-4. https://doi.org/10.1016/j.jclinepi.2012.06.003

[34] Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. BMJ 2012; 345: e5661. https://doi.org/10.1136/bmj.e5661

[35] Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, Oxman AD, Moher D. Improving the reporting of pragmatic trials: an extension of the CONSORT statement BMJ 2008; 337: a2390. https://doi.org/10.1136/bmj.a2390

[36] Piantadosi S, Byar DP, Green SB. The ecological fallacy. Am J Epidemiol 1988; 127(5): 893-904. https://doi.org/10.1093/oxfordjournals.aje.a114892

[37] Finney JW, Humphreys K, Kivlahan DR, Harris AH. Why health care process performance measures can have different relationships to outcomes for patients and hospitals: understanding the ecological fallacy. Am J Public Health 2011; 101(9): 1635-42. https://doi.org/10.2105/AJPH.2011.300153

[38] Rasbash J, Goldstein H. Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. Journal of Educational and Behavioral Statistics 1994; 19(4): 337-350. https://doi.org/10.3102/10769986019004337

[39] Goldstein H, McDonald RP. A general model for the analysis of multilevel data. Psychometrika 1988; 53(4): 455-467. https://doi.org/10.1007/BF02294400

[40] Snijders TAB. Power and Sample Size in Multilevel Linear Models, Encyclopedia of Statistics in Behavioral Science, John Wiley & Sons, Ltd. 2005.

[41] Maas CJ, Hox JJ. Sufficient sample sizes for multilevel modeling. Methodology 2005; 1(3): 86-92. https://doi.org/10.1027/1614-2241.1.3.86

Received on 17-04-2017 Accepted on 18-05-2017 Published on 04-08-2017

https://doi.org/10.6000/1929-6029.2017.06.03.4