Published in final edited form as:

Clin Trials. 2024 April ; 21(2): 199–210. doi:10.1177/17407745231208397.

Adherence to key recommendations for design and analysis of Stepped-Wedge Cluster Randomized Trials: A Review of trials published 2016–2022

Pascale Nevins[1], **Mary Ryan**[2], **Kendra Davis-Plourde**[2,3], **Yongdong Ouyang**[1,4], **Jules Antoine Pereira Macedo**[5], **Can Meng**[3], **Guangyu Tong**[2,6], **Xueqi Wang**[2,7], **Luis Ortiz-Reyes**[1], **Agnès Caille**[5,8], **Fan Li**[2,6], **Monica Taljaard**[1,4]

[1]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada

[2]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA

[3]Yale Center for Analytical Sciences, Yale School of Public Health, New Haven, CT, USA

[4]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada

[5]Université de Tours, Université de Nantes, INSERM, SPHERE U1246, Tours, France

[6]Center for Methods in Implementation and Prevention Science, Yale University, New Haven, CT, USA

[7]Section of Geriatrics, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA

[8]INSERM CIC 1415, CHRU de Tours, Tours, France

Abstract

Background/Aims: The stepped-wedge cluster randomized trial (SW-CRT), in which clusters are randomized to a time at which they will transition to the intervention condition — rather than a trial arm — is a relatively new design. SW-CRTs have additional design and analytical considerations compared to conventional parallel arm trials. To inform future methodological development, including guidance for trialists and the selection of parameters for statistical simulation studies, we conducted a review of recently published SW-CRTs. Specific objectives were to describe (1) the types of designs used in practice, (2) adherence to key requirements for statistical analysis, and (3) practices around covariate adjustment. We also examined changes in adherence over time and by journal impact factor.

Methods: We used electronic searches to identify primary reports of SW-CRTs published 2016–2022. Two reviewers extracted information from each trial report and its protocol, if available, and resolved disagreements through discussion.

Corresponding Author: Monica Taljaard, Clinical Epidemiology Program, Ottawa Hospital, Civic Campus, 1053 Carling Avenue, Civic Box 693, Admin Services Building, ASB 2-004, Ottawa, ON K1Y 4E9, 613-798-5555 x18618 phone, mtaljaard@ohri.ca.

Declaration of conflicting interests
The Authors declare that there are no conflicts of interest.

Results: We identified 160 eligible trials, randomizing a median (Q1-Q3) of 11 (8–18) clusters to 5 (4–7) sequences. The majority (122, 76%) were cross-sectional (almost all with continuous recruitment), 23 (14%) were closed cohorts and 15 (9%) open cohorts. Many trials had complex design features such as multiple or multivariate primary outcomes (50, 31%) or time-dependent repeated measures (27, 22%). The most common type of primary outcome was binary (51%); continuous outcomes were less common (26%). The most frequently used method of analysis was a generalized linear mixed model (112, 70%); generalized estimating equations were used less frequently (12, 8%). Among 142 trials with fewer than 40 clusters, only 9 (6%) reported using methods appropriate for a small number of clusters. Statistical analyses clearly adjusted for time effects in 119 (74%), for within-cluster correlations in 132 (83%), and for distinct between-period correlations in 13 (8%). Covariates were included in the primary analysis of the primary outcome in 82 (51%) and were most often individual-level covariates, however, clear and complete pre-specification of covariates was uncommon. Adherence to some key methodological requirements (adjusting for time effects, accounting for within-period correlation) was higher among trials published in higher versus lower impact factor journals. Substantial improvements over time were not observed although a slight improvement was observed in the proportion accounting for a distinct between-period correlation.

Conclusions: Future methods development should prioritize methods for SW-CRTs with binary or time-to-event outcomes, small numbers of clusters, continuous recruitment designs, multivariate outcomes, or time-dependent repeated measures. Trialists, journal editors, and peer reviewers should be aware that SW-CRTs have additional methodological requirements over parallel arm designs including the need to account for period effects as well as complex intracluster correlations.

Keywords

Methodological review; covariate adjustment; small sample correction; mixed-effects regression; intracluster correlation

Introduction

The stepped-wedge cluster randomized trial (SW-CRT) is a relatively new but increasingly popular trial design that is often used for evaluating complex interventions such as health service delivery. Unlike parallel arm CRTs, in which clusters are allocated to either control or intervention arms,[1] SW-CRTs are characterized by the fact that all clusters typically start in the control condition and gradually cross to the intervention with the timing of the cross-over being determined by random allocation. This design offers several potential advantages over a traditional parallel arm CRT design such as increased power and statistical efficiency and the ability to implement the intervention in all clusters over the course of the trial.[2,3] Terminology and key design terms related to SW-CRTs are presented in Table 1.

Due to its inherent features, the design and analysis of a SW-CRT is more complex than for a parallel arm CRT. A key requirement for appropriate statistical analysis of SW-CRTs is accounting for the confounding effects of time, by, for example, including a fixed period effect in a multivariable model.[4,5] Another key requirement is to account for the similarity among participants in the same cluster (within-cluster correlation), typically by specifying a suitable correlation structure. Because outcomes are collected from multiple clusters over time, the correlation structure should ideally allow for a within-period intracluster correlation (i.e., correlation among multiple participants in the same cluster and period), and at least a distinct between-period intracluster correlation (i.e., the correlation among multiple participants in the same cluster but in different periods). For example, the nested exchangeable correlation model[6] and the exponential decay model[7] represent two alternative methods that separately define the within-period and between-period intracluster correlations in a cross-sectional design. Extensions of the decaying correlation structure to accommodate continuous recruitment (called continuous-time decay) are also available, in which case the intracluster correlation between observations is a function of the distance between measurement times;[8] in what follows, we also consider this model as allowing for distinct within-period and between-period intracluster correlations. In the case of cohort designs, an intra-individual correlation (i.e., the correlation among repeated measurements from the same participant in different periods) should also be modelled.[6] Failure to accurately model the correlation structure may lead to an increased risk of type I error.[9,10] Because SW-CRTs often randomize a small number of clusters, methods of analysis that preserve the type I error rate, such as cluster-level analyses, non-parametric methods, generalized linear mixed models (GLMM) with degrees-of-freedom corrections, or generalized estimating equations (GEE) with small sample corrections, are essential.[11,12,13] As in other trial designs, adjusting for prespecified baseline prognostic factors in the analysis can help control for potential confounding, improve power and efficiency, and mitigate potential bias due to attrition, although ability to conduct covariate-adjusted analyses may be limited when the number of clusters is small.

The increasing use of the SW-CRT design across a range of research contexts has motivated the rapid advancement of methodology for these trials, but several gaps remain.[5] Perhaps unsurprisingly, the focus of initial methodological development has been on SW-CRTs with a (univariate) continuous outcome using large-sample methods. We previously published a descriptive analysis of SW-CRTs with a focus on randomization procedures and reporting of baseline covariate balance.[14] In the present manuscript, we report on a descriptive analysis of the same set of SW-CRTs with the primary objectives of describing: (1) design features commonly used in practice; (2) analytical approaches and adherence to key requirements for statistical analysis including accounting for period effects, complex correlations, and methods appropriate for small number of clusters; and (3) current practices around covariate adjustment in the analysis. We also examined adherence to key methodological requirements over time and by journal impact factor. The ultimate goal of this review is to inform future methodological development and shape more detailed guidance on design, analysis and reporting for SW-CRTs.

Methods

Our search strategy, eligibility criteria, screening and data sources have been described in detail elsewhere[14] and are briefly summarized here.

Search strategy and eligibility criteria

According to a prespecified protocol,[15] we aimed to identify primary reports of SW-CRTs published 1 January 2016 though 4 March 2022 (the date of the search). We used three sources to identify eligible trials: first, we included all trials included in a previously published review of implementation challenges in SW-CRTs by Caille et al.[16] (spanning January 2019 - September 2020); second, we updated the PubMed search used by Caille et al. to cover the period October 2020 - March 2022; and third, we searched an established database of primary reports of pragmatic trials (covering January 2014 - April 2019) to identify SW-CRTs.[17] Trials were considered eligible if they were SW-CRTs, conducted in humans, randomized at least five clusters, had a minimum of two sequences and three periods, and were published in English. To reflect recent practice and ensure a roughly equal number of years before and after the publication date (November 2018) of the CONSORT extension for SW-CRTs,[18] we included only primary reports published since 1 January 2016, and excluded protocols, feasibility studies, or those reporting only secondary analyses.

Screening and identification of source material

After title and abstract screening in Covidence,[19] full texts of potentially eligible reports were screened by two independent reviewers. Disagreements were resolved by discussion with a senior team member. We attempted to locate a protocol for each included trial by searching the full text and supplementary material for any mention of a protocol. When a protocol could not be located, an email was sent to the corresponding author of the publication in question, requesting a copy of the study protocol if available.

Data elements

An extraction form (see Appendix in the supplemental material) was developed to standardize the capture of data elements of interest. To describe the design characteristics of these trials (objective 1), we extracted information on the type of SW-CRT design (cross-sectional with continuous or single time-point recruitment, closed cohort, or open cohort); whether the trial was planned as complete or incomplete; and the number of clusters randomized and analyzed, number of sequences and the sample size. Sample size was defined as the number of participants (or patient-visits) in a cross-sectional design, number of participants in an open or closed cohort design, or the offset or person-time in a design with a rate or time-to-event outcome. We determined if the authors clearly identified one or more primary outcomes, noted if the primary outcome was multivariate (e.g., a questionnaire-based scale consisting of multiple subscales that are reported separately), and classified the measurement scale of the primary outcome. For cross-sectional SW-CRTs, we extracted whether there were time-dependent repeated measures (i.e., multiple outcome assessments on individuals at timepoints not defined by the step length) for the primary and any secondary outcomes. The journal impact factor in the year of publication was obtained from Journal Citation Reports;[20] or, when unavailable, from the SCImago Journal and Country Rank.

Pertaining to our second objective of describing analytical approaches used in SW-CRTs, we focused on the primary analysis of the primary outcome. To identify a single primary outcome for extraction, we chose the primary outcome defined by the trial authors; if more than one primary outcome was defined or if the authors did not clearly identify a primary outcome, we selected the outcome driving the sample size or, if no sample size calculation was presented, the first outcome listed in the section describing the outcomes of interest or the outcome presented more prominently. If the primary analysis was not clearly identified, reviewers were instructed to choose the analysis corresponding to the main result reported in the abstract, or otherwise the first analysis presented for the primary outcome. We extracted information on the statistical method used in the primary analysis, and whether the primary analysis accounted for a time effect, the within-period correlation, and a distinct between-period correlation structure. Within-period correlation was considered accounted for if authors used at least a random effect for the cluster or subcluster in the model, used GEE with robust standard errors, or conducted a cluster-period level analysis. Fixed effects regression does not yield an estimate for the intracluster correlation and was classified for our purposes as not accounting for clustering. A distinct between-period correlation was considered accounted for if the analysis included at least a cluster or subcluster by period random effect in a mixed-effects model or used GEE with either a block-exchangeable or nested exchangeable working correlation structure (but not simple exchangeable, which assumes within- and between-period correlations are equal). We also extracted whether methods of analysis appropriate for small number of clusters were used in trials with fewer than 40 clusters (simulation studies have shown that small sample corrections are generally needed to preserve the type I error rate with fewer than 40 clusters).[12,13] Applicable methods included a cluster-level analysis, GLMM with a specified degrees-of-freedom correction, GEE with a bias-corrected variances, or a randomization/permutation-based test. For non-continuous outcomes, we extracted whether both relative and absolute effects were reported. Finally, we extracted whether the primary results were positive (statistically significant in favour of the intervention) or negative.

To describe the reporting of and methods for covariate adjustment in the analysis of SW-CRTs (objective 3), we extracted whether covariates were included in the primary (as defined above) or secondary analyses. We extracted whether both adjusted and unadjusted results were presented and if so, whether results differed in statistical significance; the number of cluster- and individual-level covariates adjusted in the primary analysis; whether there was any adjustment for the baseline measure of the primary outcome; how continuous covariates were handled; and whether covariates adjusted for in the primary analysis were clearly prespecified. Covariates were considered prespecified when (1) they were specified in an available protocol, (2) they were used in restricting the randomization, or (3) the report stated that covariates were chosen *a priori*. We also extracted whether a rationale for covariate adjustment was provided and whether there were missing data on covariates, and if so, whether this was noted as a barrier to covariate adjustment in the analysis. The method used for handling missing covariates was extracted whenever missing data on covariates were noted.

Data extraction

All 11 statistician-reviewers involved in the extractions participated in pilot testing the form on eight SW-CRTs chosen to represent a variety of scenarios. After training and calibration was complete, four trials were randomly assigned to pairs of reviewers each week until all trials had been allocated. Pairs alternated each week to avoid diverging extractions. Reviewers completed extractions independently and met weekly to resolve any discrepancies; when consensus could not be reached within pairs, supervisory statisticians on the reviewing team were consulted. All data were captured in Airtable.[21]

Analysis

Counts and frequencies were used to describe categorical variables. The range, mean and standard deviation, and/or median and interquartile range were used to describe continuous variables. We calculated the absolute difference between the number of clusters randomized and analyzed to describe prevalence of including non-randomized clusters in the final analysis and cluster-level attrition. We compared the proportion of trials with positive results between trials which accounted for both time effects and within-cluster correlation (the minimum features required for the appropriate analysis of SW-CRTs) and those that did not. To examine changes over time or variation with journal impact factor, adherence to key methodological requirements (accounting for time, within- and between-period intracluster correlation) was tabulated by Journal Impact Factor (above or below the median) and publication date (before or after 2019: the CONSORT extension for SW-CRTs was published in November 2018[18]), and described using differences in proportions with 95% confidence intervals All analyses were conducted using R (v. 4.2.3).[22]

Results

Screening and inclusion

A flow diagram representing the identification and screening of SW-CRTs included in this review is presented in Supplementary Figure 1. The review from Caille et al. provided 55 trial reports.[16] The search in the pragmatic trials database initially identified 92 reports; after full-text screening and the removal of trials published before 2016, 46 reports were eligible. The updated search to 2022 yielded 117 reports after title/abstract screening, of which 65 passed full-text screening and were included in our review. Across the three sources, 166 trials were allocated to reviewers, however, during extraction a further six were discovered to not meet all inclusion criteria. Our review thus contained 160 SW-CRTs.

Descriptive characteristics

Descriptive characteristics of the 160 SW-CRT publications are presented in Table 2. Most trials were cross-sectional (122, 76.3%), 23 (14.4%) were closed cohorts and 15 (9.4%) were open cohorts. Cross-sectional trials mainly used continuous recruitment (116, 95.1%). The majority of trials were complete designs (115, 71.9%). The median number of clusters randomized per trial was 11 (Q1-Q3: 8–18) and median number of sequences was 5 (Q1-Q3: 4–7). The median sample size was 2,724 (Q1-Q3: 643–14,734). Relative to the number of clusters randomized, 7 trials included additional clusters in the analysis (a median of 1 additional cluster per trial), and 11 trials included fewer clusters in the primary analysis (a median of 2 fewer clusters per trial). Only 5 of 11 trials with cluster-level attrition provided an explanation for the reduced number of clusters (e.g., due to data collection burden, lack of resources, or failure to recruit participants), while 5 of 7 trials with additional non-randomized clusters provided explanations (e.g., due to the inclusion of pilot study clusters, or to accommodate research timelines and sub-studies). Most trials identified a single primary outcome (103, 64.4%), but 45 (28.1%) had two or more co-primary outcomes and 5 (3.1%) had multivariate outcome(s). For 7 (4.4%), the authors did not clearly identify a primary outcome. The single primary outcome identified for extraction was most often binary (81, 50.6%) or continuous (42, 26.3%). Among trials with cross-sectional designs, time-dependent repeated measures were present for the primary outcome in 12 (9.8%) and for at least one secondary outcome in 26 (21.3%); overall, 27 (22.1%) had repeated measures on at least one outcome. We located a protocol for most trials (125, 78.1%). The median journal impact factor was 7.0 (Q1-Q3: 3.4–13.4).

Methods of analysis

Summaries of methods of analysis used are presented in Table 3. The majority of trials (112, 70.0%) used GLMM for the primary analysis, while 12 (7.5%) used GEE, and 11 (6.9%) used fixed-effects generalized linear models. Time effects were accounted for in the primary analysis of 119 (74.4%) trials; 132 (82.5%) accounted for within-period correlation; only 13 (8.1%) accounted for a distinct between-period correlation. Among trials with fewer than 40 clusters, methods of analysis appropriate for small numbers of clusters were used in 9 (6.3%): five used GLMM with degrees-of-freedom correction, two used GEE with bootstrap resampling, one used Wild Bootstrap based inference, and one used a permutation-based test. Trials with appropriate methods had a median of 9 (range: 6–18) clusters. Among 118 trials with a non-continuous primary outcome, both absolute and relative treatment effects were presented in 24 (20.3%), with most presenting only relative treatment effects (77, 65.3%). The primary results were statistically significant in favour of the intervention in 76 (47.5%) trials. Of 106 (66.3%) trials accounting for both time and within-cluster correlation in the primary analysis, 44 (42%) had positive results while of the 54 (33.8%) trials lacking at least one of these elements in the analysis, 32 (59%) had positive results.

Covariate adjustment in the analysis

Details regarding covariate adjustment in the analyses are presented in Table 4. Covariate-adjusted analyses were presented in 113 (70.6%) trials: 82 (51.3%) adjusted for at least one covariate in the primary analysis while 31 (19.4%) adjusted for covariates in secondary analyses only. Overall, 55 (34.4%) trials presented both covariate-adjusted and unadjusted analyses, with results typically not differing in statistical significance. Of the 82 with covariate adjustment in the primary analysis, 36 (43.9%) included one or more cluster-level covariates and 67 (81.7%) included one or more individual-level covariates. Most trials adjusting for cluster-level covariates included a single cluster-level covariate, whereas those adjusting for individual-level covariates included a median of 3 (Q1-Q3: 1–6) individual-level covariates. In terms of how continuous covariates were handled, 17 (15.0%) used simple linear terms, 2 (1.88%) used splines, 24 (21.2%) categorized the variable, 58 (51.3%) did not specify what method was used, and 20 (17.7%) had no continuous covariates. Included covariates were clearly prespecified in 14 (17.1%) trials, clearly chosen post hoc in 20 (24.4%), and a mixture of prespecified and post hoc in 19 (23.2%). In a further 4 (4.9%) trials, covariates were clearly prespecified but some were omitted from the analysis and in 25 (30.5%) trials it was unclear whether covariates were prespecified or chosen post hoc. A rationale for covariate adjustment was provided in 68 of 113 trials (60.2%) with the most common rationale being to account for chance imbalances or confounding (50/68, 73.5%).

The presence of missing data on covariates was explicitly reported in 42 (37.2%) trials. Complete case analysis (or no missing data method) was used in 32 (28.3%) trials; the missing indicator method was used in 4 (3.5%), single imputation in 2 (1.8%), multiple imputation in 2 (1.8%) and a mixture of methods or unclear method in 2 (1.8%).

Variation in adherence to key methodological requirements

Variation in adherence to key analysis requirements is presented in Table 5. Trials published in higher (≥7.0) impact factor journals more often adjusted for time effects (absolute difference 16.3%, 95% CI for difference in proportions: 0.03 to 0.30, p=0.02), and accounted more often for within-period intracluster correlation (17.5%, 95% CI: 0.06 to 0.29, p=0.004), but we observed only a small difference in accounting for a distinct between-period intracluster correlation (3.8%, 95% CI: −0.05 to 0.12, p=0.39). Comparing trials published after versus before the CONSORT extension for SW-CRTs, we observed only small differences in the prevalence of adjusting for time effects (−4.9%, 95% CI: −0.18 to 0.09, p=0.48) and for within-period intracluster correlation (2.6%, 95% CI: −0.09 to 0.14, p=0.66); however, 11.6% trials published after 2019 accounted for a distinct between-period intracluster correlation compared to only 4.1% trials published in or before 2019 (absolute difference 7.5%, 95% CI: −0.01 to 0.16, p=0.08).

Discussion

Summary of main findings

In this review of 160 recently published SW-CRTs, we found that the majority had cross-sectional designs with continuous recruitment, and half had a binary primary endpoint. More complex designs such as cross-sectional designs with time-dependent repeated measures were common. GLMMs were the most used analysis method. Despite numerous publications emphasizing the need to account for time and clustering, approximately one-quarter in our sample did not account for a time effect and one-fifth did not account for intracluster correlation; distinct within- and between-period intracluster correlations were accounted for in less than one in 10 trials. Trials published in higher impact factor journals more often reported these key features. The use of methods of analysis suitable for small numbers of clusters was exceedingly rare: only one in fifteen trials with fewer than 40 clusters used appropriate methods. Covariate adjustment in the primary analysis was used in half of the SW-CRTs, but the covariates were often not prespecified.

Comparison with previous reviews

Cross-sectional designs have been cited as making up between 33–55% of SW-CRTs in previous reviews,[11,23,24,25] most recently up to 2017. Our finding that more than three-quarters of SW-CRTs published recently use cross-sectional designs is thus somewhat surprising and may represent a shift in the use of SW-CRTs over time or differences in how review teams classify these designs. Our result that most cross-sectional SW-CRTs use continuous recruitment is consistent with previous reviews;[23,25] however our result that

32% of studies contain multiple primary outcomes or multivariate outcomes is much higher than the 7% found previously.[26] Our review identified binary primary outcomes as the most common for SW-CRTs, consistent with previous reviews.[26,27]

The observed prevalence of methods of analysis for SW-CRTs in our review (70% GLMMs and 8% GEE) is consistent with previous reviews of SW-CRTs, such as the review by Barker et al.[11] which found that 59% of 102 SW-CRTs published up to 2015 used GLMMs and 17% GEE, and the review by Kristunas et al.[26] which found that 56% used GLMMs and 13% GEE. Time effects were accounted for in 60% of SW-CRTs in the review by Barker et al.,[11] compared to 74% in our study, which may indicate an improvement over time, although our stratified comparison of before versus after 2019 saw no substantial improvement. Although previous reviews of SW-CRTs did not report on covariate adjustment, a review of 300 (mostly parallel arm) CRTs published 2000–2008 by Wright et al., found that 73% of CRTs reported at least one covariate-adjusted analysis[28] which is comparable to our finding of 71% in SW-CRTs. In the review by Wright et al.,[28] 17% of CRTs reporting adjusted analyses clearly chose covariates post-hoc; adherence to this principle of pre-specification may be substantially more difficult in SW-CRTs as only 17% of trials in our review clearly prespecified and included all covariates in the analysis.

Strengths and limitations

Important strengths of our study include the large sample size—the largest to-date on SW-CRTs—and rigorous double extraction of all variables by trained statistician-reviewers. We identified available protocols for 78% of included trials, which provided additional details for several analysis-related extractions. A limitation of our study is that we used SW-CRTs from an existing pragmatic trials database to supplement a stepped-wedge specific search implemented in PubMed; our search may therefore not have captured all SW-CRTs published in this date range although the pragmatic trials database used SW-CRT related terms in its search strategy. Finally, our stratified analysis of trials published before versus after the CONSORT extension for SWCRTs found no real improvement, but this may merely reflect the inevitable lag from the time of trial design and protocol development to its final publication: it may take several years for a CONSORT statement to have a measurable impact on the methodological quality of published trials.

Implications for research and practice

We have identified several design features of SW-CRTs that suggest areas for further methodological development. First, although existing methods do not differentiate between continuous and fixed time-point recruitment, continuous recruitment designs are very common and require more attention in methods development.[8,29] Second, despite adherence to the implementation schedule being a noted challenge in the implementation of SW-CRTs,[16,30] incomplete designs made up less than one-third of our sample. Additional guidance on the importance of incorporating transition periods, as well as dissemination of recent methods for batched[31] and staircase designs[32,33] may be useful. Third, approximately one third of trials identified multiple or multivariate primary outcomes: trialists may benefit from more applied papers incorporating recently published methods for SW-CRTs with co-primary outcomes.[34] Fourth, cross-sectional designs with time-dependent repeated measures on individuals are not uncommon, but we are unaware of any statistical papers addressing methods of analysis for such designs. Finally, while continuous outcomes are often the focus for initial statistical methods development, our review found that non-continuous outcomes are more common in SW-CRTs and may need to be prioritized in future methods development. Time-to-event outcomes were relatively rarely reported, although it is possible that investigators treated such outcomes as binary due to the relative absence of methods for time-to-event analyses in SW-CRTs.

Our finding that many SW-CRTs do not account for time effects or clustering in the analysis is concerning. Peer reviewers and trialists should be aware that estimated intervention effects from models with and without accounting for time can be in opposite directions and should insist on treatment effects obtained from models that account for time, even if the time effect is not statistically significant. Whereas the need to account for at least one distinct between-period correlation has been well-established in the methodological literature since 2016,[6] we found few trials accounting for more complex correlation structures beyond simple exchangeable. This may reflect the inevitable delay before new methodology makes its way into practice; it may also reflect computational challenges, perhaps due to trials being too small to fit more complex correlation structures.[35] It is also concerning that almost no trialists reported on the use of methods appropriate for small number of clusters, despite the fact that the median number of clusters randomized was only 11. Further work is required regarding small sample corrections for SW-CRTs[10,12,36,37,38,39,40] Finally, our review found sub-optimal practices around covariate adjustment in SW-CRTs: although nearly three quarters of trials presented a covariate-adjusted analysis of the primary outcome, covariates were often not prespecified, which raises concerns about model selection. A possible explanation is that pre-specification is more complex for a SW-CRT due to the requirements to adjust for period effects (often modelled categorically) and account for distinct within- and between-period intracluster correlation structures; thus, confidence about the ability to adjust for covariates, especially cluster-level covariates, at the design stage may be limited and such decisions may then be postponed to the analysis stage. To this end, guidance and best practices for specifying covariates to adjust in the design and analysis stages remain to be developed for SW-CRTs.

Conclusions

The use of SW-CRTs has rapidly increased over the past two decades and has outpaced its methodological development. More guidance, including tutorial-style manuscripts and other tools should be developed to guide trialists, statisticians, peer reviewers and editors in the use of robust designs and methods for SW-CRTs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

This work was supported by the Canadian Institutes of Health Research through the Project Grant competition (competitive, peer-reviewed), award number PJT-153045; and by the National Institute of Aging (NIA) of the National Institutes of Health under Award Number U54AG063546, which funds NIA Imbedded Pragmatic

Alzheimer's Disease and AD-Related Dementias Clinical Trials Collaboratory (NIA IMPACT Collaboratory). This work was also supported by Patient-Centered Outcomes Research Institute Awards (ME-2020C3-21072 and ME-2022C2-27676). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, PCORI, or its Board of Governors or Methodology Committee. The funder played no role in the study. KDP and FL were also funded by the Yale Clinical and Translational Science Award (UL1 TR001863).

References

1. Donner A and Klar N. Design and analysis of cluster randomization trials in health research. London: Arnold Publishers Limited, 2000.
2. Hemming K, Haines TP, Chilton PJ, et al. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ 2015 Feb 6; 350: h391. [PubMed: 25662947]
3. Prost A, Binik A, Abubakar I, et al. Logistic, ethical, and political dimensions of stepped wedge trials: critical review and case studies. Trials 2015 Aug 17; 16: 351. [PubMed: 26278521]
4. Hussey MA and Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials 2007 Feb; 28(2): 182–191. [PubMed: 16829207]
5. Li F, Hughes JP, Hemming K, et al. Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: An overview. Stat Methods Med Res 2021; 30(2): 612–639. [PubMed: 32631142]
6. Hooper R, Teerenstra S, de Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Stat Med 2016 Nov 20;35(26):4718–4728. [PubMed: 27350420]
7. Kasza J, Hemming K, Hooper R, et al. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. Stat Methods Med Res 2019;28(3):703–716. [PubMed: 29027505]
8. Grantham KL, Kasza J, Heritier S, et al. Accounting for a decaying correlation structure in cluster randomized trials with continuous recruitment. Stat Med 2019 May 20;38(11):1918–1934. [PubMed: 30663132]
9. Kasza J and Forbes AB. Inference for the treatment effect in multiple-period cluster randomised trials when random effect correlation structure is misspecified. Stat Methods Med Res 2019; 28(10–11): 3112–3122. [PubMed: 30189794]
10. Ford WP and Westgate PM. Maintaining the validity of inference in small-sample stepped wedge cluster randomized trials with binary outcomes when using generalized estimating equations. Stat Med 2020; 39(21): 2779–2792. [PubMed: 32578264]
11. Barker D, McElduff P, D'Este C, Campbell MJ. Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. BMC Med Res Methodol 2016; 16: 69. [PubMed: 27267471]
12. Thompson JA, Hemming K, Forbes A, et al. Comparison of small-sample standard-error corrections for generalised estimating equations in stepped wedge cluster randomised trials with a binary outcome: A simulation study. Stat Methods Med Res 2021; 30(2): 425–439. [PubMed: 32970526]
13. Leyrat C, Morgan KE, Leurent B, Kahan BC. Cluster randomized trials with a small number of clusters: which analyses should be used? Int J Epidemiol 2018; 47(1): 321–331. [PubMed: 29025158]
14. Nevins P, Davis-Plourde K, Macedo JP, et al. A Scoping Review described diversity in methods of randomization and reporting of baseline balance in Stepped-Wedge Cluster Randomized Trials. J Clin Epidemiol 2023; 157: 134–145. [PubMed: 36931478]
15. Nevins P, Davis-Plourde K, Ouyang Y, et al. Handling of covariates in stepped-wedge cluster randomized trials: Protocol for a methodological review. uO Research 2022 Aug, http://hdl.handle.net/10393/43901.
16. Caille A, Taljaard M, Le Vilain-Abraham F, et al. Recruitment and implementation challenges were common in stepped-wedge cluster randomized trials: results from a methodological review. J Clin Epidemiol 2022; 148: 93–103. [PubMed: 35483552]

17. Nicholls SG, Carroll K, Hey SP, et al. A review of pragmatic trials found a high degree of diversity in design and scope, deficiencies in reporting and trial registry data, and poor indexing. J Clin Epidemiol 2021; 137: 45–57. [PubMed: 33789151]

18. Hemming K, Taljaard M, McKenzie JE, et al. Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. BMJ 2018; 363: k1614. [PubMed: 30413417]

19. Veritas Health Innovation Ltd. Covidence, https://www.covidence.org/ (2022, accessed March 2022).

20. Clarivate Analytics. Journal Citation Reports, https://jcr.clarivate.com (2022, accessed July 2022).

21. Formagrid, Inc. Airtable, https://airtable.com/ (2022, accessed May 2023).

22. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2023, https://www.R-project.org/

23. Grayling MJ, Wason JM and Mander AP. Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. Trials 2017; 18(1): 33. [PubMed: 28109321]

24. Eichner FA, Groenwold RHH, Grobbee DE, Oude Rengerink K. Systematic review showed that stepped-wedge cluster randomized trials often did not reach their planned sample size. J Clin Epidemiol 2019; 107: 89–100. [PubMed: 30458261]

25. Beard E, Lewis JJ, Copas A, et al. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. Trials 2015; 16: 353. [PubMed: 26278881]

26. Kristunas C, Morris T and Gray L. Unequal cluster sizes in stepped-wedge cluster randomised trials: a systematic review. BMJ Open 2017 Nov 15; 7(11): e017151.

27. Martin J, Taljaard M, Girling A, Hemming K. Systematic review finds major deficiencies in sample size methodology and reporting for stepped-wedge cluster randomised trials. BMJ Open 2016; 6(2): e010166.

28. Wright N, Ivers N, Eldridge S, et al. A review of the use of covariates in cluster randomized trials uncovers marked discrepancies between guidance and practice. J Clin Epidemiol 2015; 68(6): 603–609. [PubMed: 25648791]

29. Hooper R and Copas A. Stepped wedge trials with continuous recruitment require new ways of thinking. J Clin Epidemiol 2019; 116: 161–166. [PubMed: 31272885]

30. Unni RR, Lee SF, Thabane L, et al. Variations in stepped-wedge cluster randomized trial design: Insights from the Patient-Centered Care Transitions in Heart Failure trial. Am Heart J 2020; 220: 116–126. [PubMed: 31805422]

31. Kasza J, Bowden R, Hooper R, Forbes AB. The batched stepped wedge design: A design robust to delays in cluster recruitment. Stat Med 2022; 41(18): 3627–3641. [PubMed: 35596691]

32. Kasza J, Hooper R and Forbes A. Staircase cluster randomised trial designs. Presentation, Monash University, https://www.sctweb.org/presentations/2019/Kasza_SCT2019.pdf (2019, accessed May 2023).

33. Hooper R, Kasza J and Forbes A. The hunt for efficient, incomplete designs for stepped wedge trials with continuous recruitment and continuous outcome measures. BMC Med Res Methodol 2020; 20(1): 279. [PubMed: 33203361]

34. Davis-Plourde K, Taljaard M and Li F. Power analyses for stepped wedge designs with multivariate continuous outcomes. Stat Med 2023; 42(4): 559–578. [PubMed: 36565050]

35. Rezaei-Darzi E, Kasza J, Forbes A, et al. Use of information criteria for selecting a correlation structure for longitudinal cluster randomised trials. Clin Trials 2022; 19(3): 316–325. [PubMed: 35706343]

36. Scott JM, deCamp A, Juraska M, et al. Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. Stat Methods Med Res 2017; 26(2): 583–597. [PubMed: 25267551]

37. Li F, Turner EL and Preisser JS. Sample size determination for GEE analyses of stepped wedge cluster randomized trials. Biometrics 2018;74(4):1450–1458. [PubMed: 29921006]

38. Design Li F. and analysis considerations for cohort stepped wedge cluster randomized trials with a decay correlation structure. Stat Med 2020;39(4):438–455. [PubMed: 31797438]

39. Li F, Yu H, Rathouz PJ, et al. Marginal modeling of cluster-period means and intraclass correlations in stepped wedge designs with binary outcomes. Biostatistics 2022;23(3):772–788. [PubMed: 33527999]

40. Zhang Y, Preisser JS, Turner EL, et al. A general method for calculating power for GEE analysis of complete and incomplete stepped wedge cluster randomized trials. Stat Methods Med Res 2023;32(1):71–87. [PubMed: 36253078]

Table 1.

Terminology and key design features of SW-CRTs

Design terms	**Explanation**
Sequence	Group of one or more clusters which is defined by the time at which its cluster(s) will transition from control to intervention condition.
Step	The transition timepoint; usually equidistantly spaced in time across the duration of the trial. The steps define “periods”, which are the unit blocks of time.
Cluster-period	The intersection of a period and a sequence, a single cell; the unit on which observations are taken.
Complete design	Design in which observations are collected from each cluster-period.
Incomplete design	Some cluster-periods are excluded from data collection, for example, to allow for implementation of the intervention during a transition period or to reduce the data collection burden.
Closed cohort design	All participants are identified at the beginning of the trial and the same participants are repeatedly measured in every cluster-period.
Open cohort design	Participants are repeatedly measured in multiple cluster-periods, though not all participants contribute an equal number of measurements: by design, participants may join or leave the cohort while the trial is underway.
Cross-sectional design	A design in which different participants are identified and measured in each cluster-period.
Continuous recruitment	Participants are recruited in continuous time throughout each cluster-period, e.g., as they arrive at a clinic.
Fixed time recruitment	Participants are recruited at one time-point per cluster-period e.g., through the administration of a cross-sectional survey.

Table 2.

Characteristics of included stepped-wedge cluster randomized trials (N = 160)

Characteristic	Frequency (%)
Type of stepped-wedge trial	
Cross-sectional	122 (76.3)
Continuous recruitment	116 (95.1)
Fixed time-point recruitment	6 (4.9)
Open cohort	15 (9.4)
Closed cohort	23 (14.4)
Complete or incomplete design	
Complete	115 (71.9)
Incomplete	45 (28.1)
Number of clusters randomized	
Median (Q1, Q3)	11 (8, 18)
Min, Max	5, 291
Not reported	1
Number of sequences	
Median (Q1, Q3)	5 (4, 7)
Min, Max	2, 81
Not reported	2
Sample size[a]	
Median (Q1, Q3)	2724 (643, 14733.5)
Range	44, 4801573
Not reported	5
Analysis included additional non-randomized clusters[b]	
Yes	7 (4.4)
No	152 (95.0)
Additional number included: median (range)	1 (1–8)
Analysis excluded randomized clusters[b]	
Yes	11 (6.9)
No	148 (92.5)
Number clusters excluded: median (range)	2 (1–34)
Primary outcome(s) clearly identified	
Yes: One primary outcome	103 (64.4)
Yes: Two or more co-primary outcomes	45 (28.1)
Yes: Multivariate outcome(s) [c]	5 (3.1)
No outcome(s) clearly defined as primary	7 (4.4)

Characteristic	Frequency (%)
Type of primary outcome[d]	
Continuous	42 (26.3)
Binary	81 (50.6)
Ordinal	1 (0.6)
Time-to-event	9 (5.6)
Count or Rate	27 (16.9)
If trial has a cross-sectional design, are there time-dependent repeated measures for the primary outcome (N = 122)	
Yes	12 (9.8)
No	110 (89.4)
If trial has a cross-sectional design, are there time-dependent repeated measures for any secondary outcomes (N = 122)	
Yes	26 (21.3)
No	91 (74.6)
Not applicable	5 (4.1)
Is a protocol available	
Yes	125 (78.1)
No	35 (21.9)
Journal Impact Factor	
Median (Q1, Q3)	7.0 (3.4, 13.4)

[a] Defined as number of participants or visits in a cross-sectional design, number of participants in an open or closed cohort design, or the off-set or person-time in a design with a rate or time-to-event outcome.

[b] The number of clusters included in the analysis (relative to the number in the randomization) could not be determined for one trial.

[c] Multivariate outcomes are, for example, a questionnaire-based scale consisting of multiple subscales that are reported separately.

[d] Based on the unit of analysis of the single primary outcome defined by the trial authors. If more than one or no clear primary outcomes were defined, extractors selected the outcome driving the sample size or, if no sample size calculation was presented, selected the first outcome listed under “outcomes”.

Table 3.

Characteristics of the primary analysis of the primary outcome [a] (N = 160)

Characteristic	Frequency (%)
Statistical method used	
Generalized Estimating Equations (GEE)	12 (7.5)
Generalized Linear Mixed Model (GLMM)	112 (70.0)
Fixed-effects General Linear Model (GLM)	11 (6.9)
Cox or Accelerated failure time model	9 (5.6)
Simple/Naïve analysis	9 (5.6)
Other	5 (3.1)
Unclear	2 (1.3)
Adjusted for time or period effects	
Yes	119 (74.4)
No	39 (24.4)
Unclear	2 (1.3)
Accounted for within-period intracluster correlation	
Yes	132 (82.5)
No	24 (15.0)
Unclear	4 (2.5)
Allowed for a distinct between-period correlation	
Yes	13 (8.1)
No	146 (91.3)
Unclear	1 (0.6)
Reported method of analysis appropriate for small numbers of clusters[b](N = 142 with <40 clusters)	
Yes	9 (6.3)
No	133 (93.7)
Reported absolute and/or relative treatment effects (N = 118 with non-continuous outcome)	
Only absolute	17 (14.4)
Only relative	76 (65.0)
Both absolute and relative	24 (20.3)
Primary results	
Positive (i.e., statistically significant in favour of intervention)	76 (47.5)
Negative	84 (52.5)

[a] The single primary outcome defined by the trial authors. If more than one or no clear primary outcomes were defined, extractors selected the outcome driving the sample size or, if no sample size calculation was presented, selected the first outcome listed under “outcomes”.

[b] Applicable methods for a small sample correction included a cluster-level analysis, GLMM with a specified degrees-of-freedom correction, GEE with a bias corrected variances, or a non-parametric approach.

Table 4.

Covariate adjustment in analyses of the primary outcome (N = 160)

Characteristics	Frequency (%)
Covariates included in the analysis	
Yes: at least in the primary analysis	82 (51.3)
Yes: in secondary analyses	31 (l9.4)
No covariates in any analyses of primary outcome	41 (25.6)
Unclear	6 (3.8)
Both adjusted and unadjusted analyses presented	
Yes	55 (34.4)
Yes, and they differ in significance	8 (14.5)
Yes, but they do not differ in significance	44 (80.0)
Yes, but insufficient information to determine significance	3 (5.5)
No	105 (65.6)
Number of cluster-level covariates in primary analysis (N = 82)	
0	44 (52.4)
1	25 (30.5)
≥2	11 (13.4)
Range	0 – 5
Unclear	2 (2.4)
Number of individual-level covariates in primary analysis (N = 82)	
0	13 (15.9)
1	9 (11.0)
≥2	58 (70.7)
Median (Q1, Q3)	3 (1, 6)
Range	0 – 16
Unclear	2 (2.4)
Adjustment for baseline measure of primary outcome? (N = 113)	
Yes	11 (9.7)
No or not applicable	102 (90.3)
Handling of continuous covariates* (N = 113)	
Simple linear terms	17 (15.0)
Splines	2 (1.8)
Categorization	24 (21.2)
Not specified	58 (51.3)
No continuous covariates	20 (17.7)
Covariates adjusted in the primary analysis prespecified? (N = 82)	
Clearly prespecified and all specified covariates included	14 (17.1)
Clearly prespecified, but some covariates omitted	4 (4.9)

Characteristics	Frequency (%)
Clearly chosen post hoc	20 (24.4)
Mixture (some prespecified, some post hoc)	19 (23.2)
Unclear	25 (30.5)
Rationale for covariate adjustment?[a](N = 113)	
Yes	68 (60.2)
Chance imbalance/confounding	50 (73.5)
Correlation with outcome	17 (25.0)
Improve precision of treatment effect	5 (7.3)
Non-compliance or non-participation	2 (2.9)
Missing data bias	1 (1.5)
No	45 (39.8)
Presence of missing data on covariates noted (N = 113)	
Yes	42 (37.2)
No	70 (61.9)
Unclear	1 (0.9)
Missing data on covariates mentioned as a barrier to adjustment (N = 113)	
Yes	4 (3.6)
No	109 (96.5)
Method for handling missing covariates in any analysis of the primary outcome (N = 42)	
Complete case analysis or no method specified	32 (28.3)
Missing indicator method	4 (3.5)
Single imputation	2 (1.8)
Multiple imputation	2 (1.8)
Mixture or unclear	2 (1.8)

[a] Multiple selections possible.

Table 5.

Changes in adherence to key requirements for analysis over time and difference between higher versus lower impact factor journals. Entries are Frequency (%)

	Publication Year			**Journal Impact Factor**		
	≤2019 (N = 74)	**>2019 (N = 86)**	**Difference in proportions (95% CI)**	**≤7.0 (N = 80)**	**>7.0 (N = 80)**	**Difference in proportions (95% CI)**
Adjusted for time or period effects			−0.05 (−0.18, 0.09)			0.16 (0.03, 0.30)
Yes	57 (77.0)	62 (72.1)		53 (66.3)	66 (82.5)	
No	16 (21.6)	23 (26.7)		27 (33.7)	12 (15.0)	
Unclear	1 (1.4)	1 (1.2)		0	2 (2.5)	
Accounted for within-period intracluster correlation			0.03 (−0.09, 0.14)			0.18 (0.06, 0.29)
Yes	60 (81.1)	72 (83.7)		59 (73.8)	73 (91.3)	
No	13 (17.6)	11 (12.8)		18 (22.5)	6 (7.5)	
Unclear	1 (1.4)	3 (3.5)		3 (3.7)	1 (1.3)	
Allowed for a distinct between-period correlation			0.08 (−0.006, 0.16)			0.04 (−0.04, 0.12)
Yes	3 (4.1)	10 (11.6)		5 (6.3)	8 (10.0)	
No	71 (95.9)	75 (87.2)		75 (93.7)	71 (88.7)	
Unclear	0	1 (12)		0	1 (13)	