A methodological review of current practice with extremely small stepped-wedge cluster randomized trials

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

1. Department of Internal Medicine, Yale School of Medicine. 2. Department of Biostatistics, Yale School of Public Health. 3. Clinical Epidemiology Program, Ottawa Hospital Research Institute. 4. Department of Epidemiology and Biostatistics, Western University. 5. Department of Population Health Sciences and Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison. 6. Child Health Evaluative Sciences, The Hospital for Sick Children. 7. Université de Tours, Université de Nantes. 8.School of Epidemiology and Public Health, University of Ottawa.

Yale

What is a stepped-wedge cluster randomized trial?

The stepped-wedge cluster randomized trial (SW-CRT) randomizes clusters to transition from the control to the intervention condition in a staggered fashion. Compared to the parallel-arm cluster randomized design, the SW-CRT has the advantage that all clusters eventually receive the intervention during the study. The design also offers some logistical advantage in that implementation of the intervention is naturally staggered over time. The design can further improve statistical efficiency by combining both between-cluster and within-cluster comparisons for treatment effect estimation. Despite these advantages, potential challenges in the proper design and conduct of SW-CRTs, as well as statistical complexities for estimating the treatment effect, have been increasingly recognized in the literature.

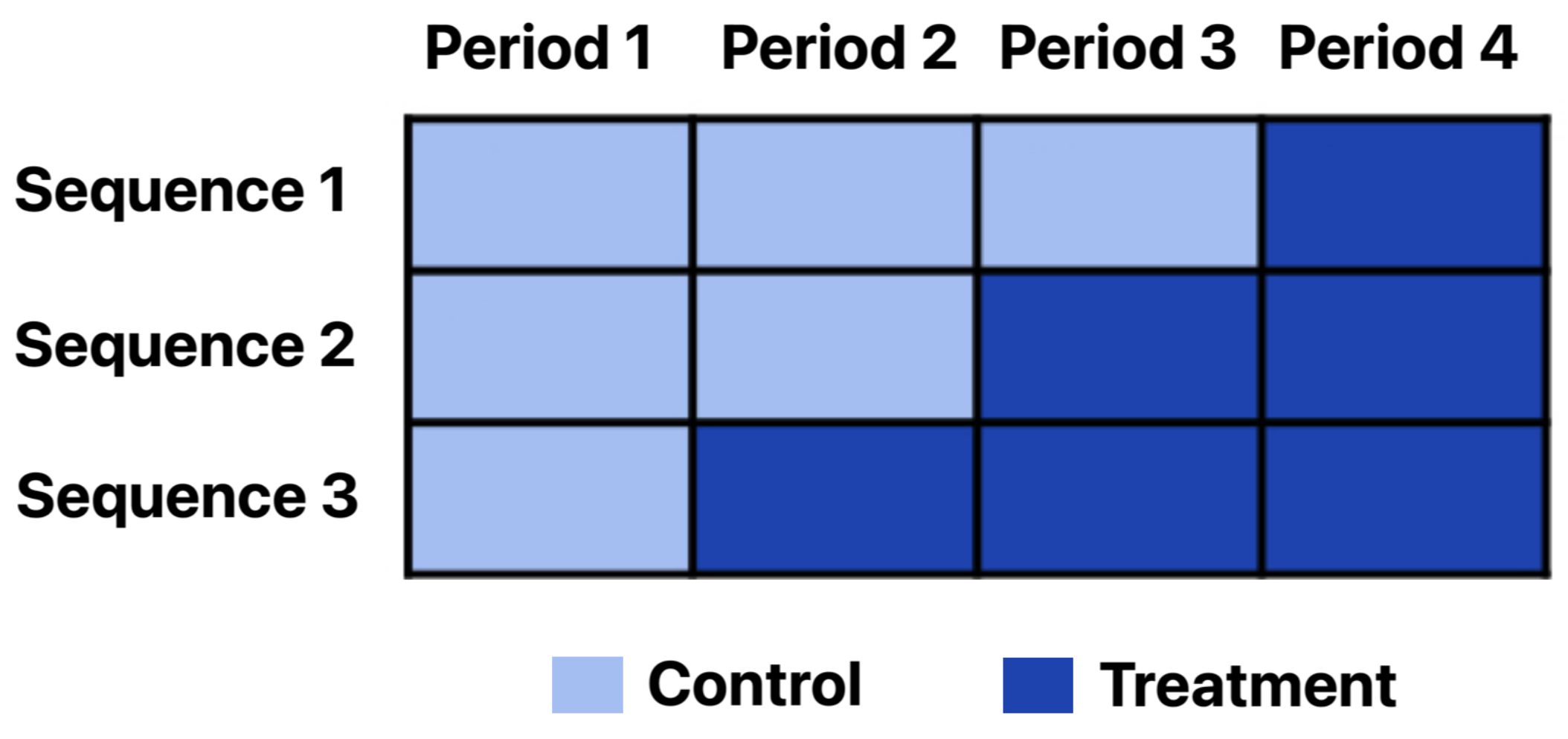

Figure 1. A sample diagram of a stepped-wedge with three sequences and four periods

Motivation for reviewing small SW-CRTs

Although there is no consensus on the minimum required number of clusters for a SW-CRT, published trials typically include a fairly small number of clusters, for example, a median (Q1-Q3) of 11 (8-18) clusters in a recent review.[2] The design and analysis of a small SW-CRT can be problematic due to

- **No guideline for small-sample correction at design.** No explicit guidelines for implementing small sample adjustments at the design stage.
- **Mistakenly used power procedure based on large-sample approximations.** Commonly used power calculation procedures assume that the number of clusters is relatively large.[1] Calculations based on large-sample approximations can be misleading and lull investigators into a false sense about the power that can be achieved with a limited number of clusters
- **Failure to distinguish a within-period and different between-period intracluster correlation coefficient (ICC)** When assuming that within- and between-period ICCs are equal, the estimated number of clusters can mathematically approach one as the cluster size increases.[3]
- **Challenges to account for the complex longitudinal correlation structure.** Few simulation studies have been conducted to examine the performance of available statistical methods for SW-CRTs with very small numbers of clusters. Small-sample adjustments to standard errors and appropriate degrees of freedom have been recommended to preserve the validity of inference under a small number of clusters with generalized estimating equations (GEEs) or generalized linear mixed models (GLMMs). Alternative methods include cluster-level methods and permutation tests which do not rely on large-sample approximations or Bayesian inference which is computationally more flexible with a small number of clusters.

Data collection for the review

Electronic searches were used to identify primary reports of full-scale SW-CRTs published 2016-2022; the subset of **61** SW-CRTs that randomized **2-6 clusters** was identified. Key data extraction elements include

1. **Trial size.** Numbers of clusters randomized and analyzed, numbers of sequences, sample sizes, sample size and power calculations.
2. **Study design characteristics.** Types of clusters and interventions, types of SW-CRT designs, whether labelled an implementation trial.
3. **Design justifications.** Reasons for the use of a SW-CRT as required by the CONSORT extension for SW-CRTs and for the small number of clusters.
4. **Analytical approaches used for the primary outcome.** Reporting of methods of analysis appropriate for small numbers of clusters, use of covariates in the analysis, presence of baseline imbalances, and whether the primary result was statistically significant.

Trial size

The median sample size was 1,426 (Q1-Q3: 420-7,553). Ten (16.4%) SW-CRTs presented no sample size or power calculations for the primary outcome.The majority (33, 84.6%) accounted for at least the within-period ICC; only 1 trial distinguished a different within and between-period ICC in the sample size calculation.

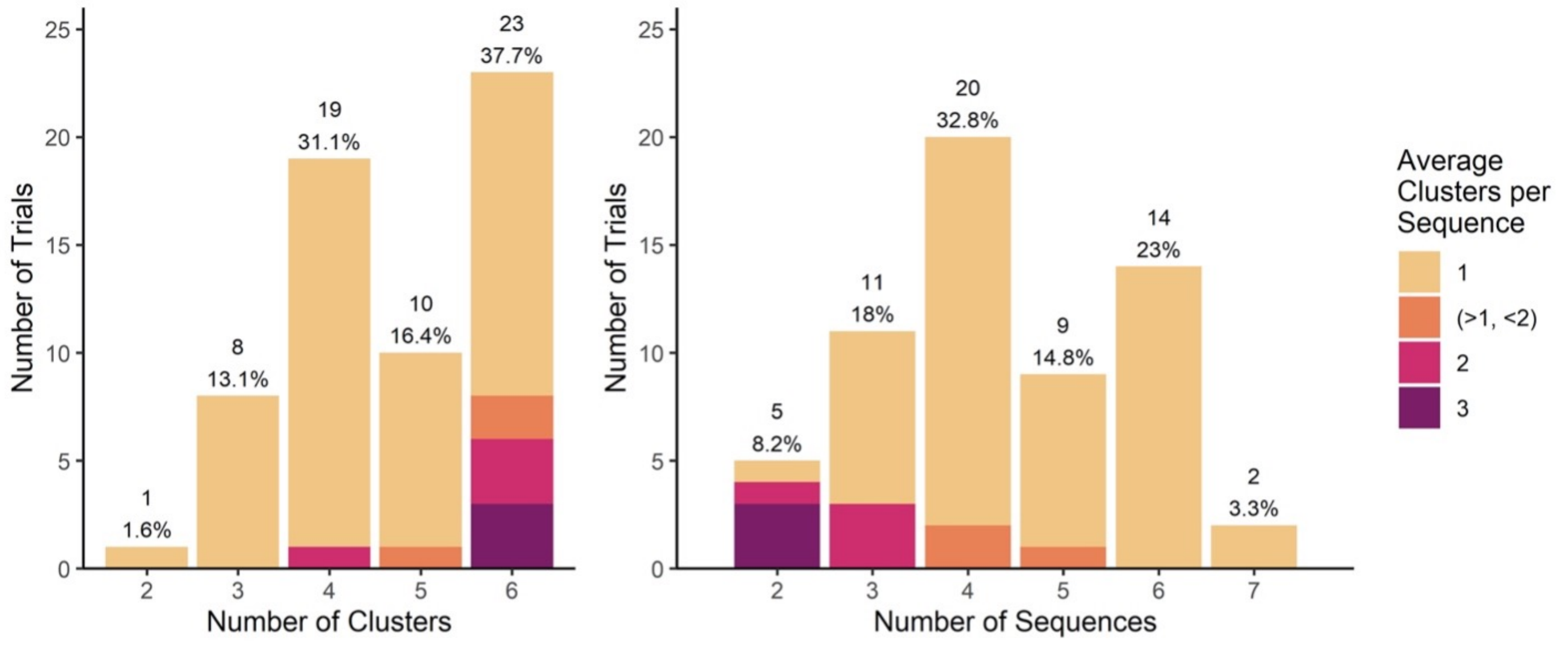

Figure 2. Histograms of the number of clusters and number of sequences for small SW-CRT (N=61) stratified by average clusters per sequence. 3 trials each included a single additional, non-randomized cluster in the analysis (allowing for up to 7 sequences)

Study design characteristics

The majority of these trials were published after 2018 (49, 80.4%); were conducted in North America (16, 26.2%) or Europe (16, 26.2%); carried out in hospitals, hospital wards/units/teams (23, 37.7%) or communities or geographical areas (17, 27.9%); and targeted healthcare professionals (27, 44.3%). Most trials employed a cross-sectional design (49, 80.2%). The majority of trials were complete designs (45, 73.8%). About a quarter (16, 26.2%) were clearly labelled as implementation or hybrid implementation-effectiveness trials.

Design justifications

Most trials (45, 73.8%) did not acknowledge the small number of clusters as a limitation. About one-fifth of trials (12, 19.7%) used language suggesting the trial may have been considered a "preliminary" evaluation (vs. "full-scale").

Most trials justified the use of the stepped-wedge roll-out (44, 72.1%)

Stated reasons include logistical or administrative convenience (24, 54.5%), to improve methodological rigor (21, 47.7%), to facilitate recruitment (14, 31.8%), to improve power or efficiency (14, 31.8%), to reduce bias from between-cluster differences (9, 20.5%), desire to implement likely beneficial intervention (9, 20.5%), or unethical to withhold intervention from clusters (5, 11.4%).

Only 16 (26.2%) trials explained the small number of clusters

Among them, 5 (31.3%) mentioned limited availability of eligible clusters, 5 (31.3%) were limited by available resources, 4 (25.0%) were limited by other feasibility considerations (i.e., timelines, logistics, difficulties in implementation), and 6 (37.5%) indicated that clusters were defined based on the need to minimize contamination which led to a small number of available clusters.

Analytical approaches used for the primary outcome

The majority of trials (44, 72.1%) used GLMMs for the primary analysis. Only 5 (8.2%) accounted for a distinct between-period correlation in the analysis. Methods of analysis specifically considering correction for a small number of clusters were found in only 4 (6.6%). Another 8 (13.1%) used a fixed effects model.

Methods for small-sample correction	Count(%)
GLMM with degree of freedom correction	2 (3.2)
GEE with small-sample correction	1 (1.6)
Bayesian analysis	1 (1.6)
Cluster-period-level analysis	0 (0)
Permutation tests	0 (0)
Fixed-effects model	8 (13.1)
Unclear	2 (3.2)
No	47 (77.0)

Table 1. Reported method of analysis considering correction for the small number of clusters

Over half of the trials (34, 55.7%) reported baseline imbalances. The primary results were statistically significant in favor of the intervention in 33 (54.1%) trials.

References

[1] Fan Li, James P Hughes, Karla Hemming, Monica Taljaard, Edward R Melnick, and Patrick J Heagerty. Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: an overview. *Statistical Methods in Medical Research*, 30(2):612–639, 2021.

[2] Pascale Nevins, Kendra Davis-Plourde, Jules Antoine Pereira Macedo, Yongdong Ouyang, Mary Ryan, Guangyu Tong, Xueqi Wang, Can Meng, Luis Ortiz-Reyes, Fan Li, et al. A scoping review described diversity in methods of randomization and reporting of baseline balance in stepped-wedge cluster randomized trials. *Journal of clinical epidemiology*, 157:134–145, 2023.

[3] Monica Taljaard, Steven Teerenstra, Noah M Ivers, and Dean A Fergusson. Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. *Clinical Trials*, 13(4):459–463, 2016.

Funding: NIA IMPACT Collaboratory U54AG063546; Canadian Institutes of Health Research PJT-153045; Patient-Centered Outcomes Research Institute Award ME-2022C2-27676

Correspondence email: guangyu.tong@yale.edu