NIH Collaboratory Rethinking Clinical Trials®

Health Care Systems Research Collaboratory

Design & Analysis of Embedded Pragmatic Clinical Trials Workshop Summary

May 2, 2019

Lister Hill Auditorium, NIH Campus

Table of Contents and Contributors

Introduction

In 2019, NIH Health Care Systems Research Collaboratory held a comprehensive workshop to explore and discuss statistical issues encountered with embedded pragmatic clinical trials (ePCTs). We define ePCTs as pragmatic clinical trials that are embedded within health care delivery systems and that leverage the context to both deliver interventions and record outcomes. The workshop entailed panel discussions on topics including measurement and data, cluster designs, and choosing between a parallel and stepped wedge design. During the panel discussions, the PI and statistician of the ePCTs presented and discussed the biostatistical challenges encountered during the design and analysis of their studies. This document summarizes the lessons learned and recommends tools to help others design and analyze future ePCTs.

Panel 1. Measurement and Data: Outcomes, Exposures, and Subgroups Based on EHR Data

During this panel, participants discussed issues regarding the measurement of outcome and exposure variables, including associated errors and potential bias, which is particularly important when the source of information regarding outcomes is from extant data systems, like the electronic health record (EHR) or claims data.

Moderator, Rui Wang

Vince Mor (PI) and Roee Gutman (Statistician) for PROVEN

- Pragmatic Trial of Video Education in Nursing Homes (PROVEN)

Nancy Latham and Dave Ganz (PIs) and Peter Peduzzi (Statistician) for STRIDE

- Strategies to Reduce Injuries and Develop Confidence in Elders (STRIDE)

Susan Huang (PI) and Ken Kleinman (Statistician) for ABATE

- Active Bathing to Eliminate (ABATE) Infection

There are many methodological challenges associated with measuring and analyzing outcomes in ePCTs, including addressing secular trends, missing data, compliance, and contamination due to knowledge of the trial or intervention. First, secular trends may affect data availability, outcome prevalence, and compliance to the intervention. To

counter this, PIs should monitor factors that may be subject to secular trends and develop plans to respond appropriately. For example, to monitor compliance to the intervention, investigators may consider measuring this at multiple levels. In the PROVEN trial, the intervention is exposure to a video about advance care planning, and compliance is measured at two levels: by video status reports documenting compliance at the facility-level (video offered or not) and at the individual-level (video shown or not) (Mor et al. 2017). The team considered planned "as treated" secondary analysis based on intervention compliance strata.

Second, a defining feature of ePCTs is the use of routinely collected data from the EHR or claims; however, these data may not capture the entire patient population and may miss events. For example, in the PROVEN trial, hospital transfers were measured via Medicare/Medicaid mandated Minimum Data Set (MDS) resident assessment instrument; however the MDS does not capture all hospital transfers like Emergency Department visits and Observation Stays. On the other hand, reliance on Medicare health care utilization claims data would miss the 35% of the Medicare population that are members of Medicare Advantage plans.. In response, PIs could consider augmenting these data with other alternative sources.

Third, the intervention or knowledge of the participating in the intervention may affect outcome ascertainment. For these case studies, the control arm is usual care, and in the experimental arm, knowledge of being in an intervention arm of a trial may influence the recording of diagnoses or symptoms such that they are recorded more assiduously and completely, resulting in more symptoms or problems among the experimental patients relative to the controls. For example, in the ABATE trial, which tested whether routine bathing and decolonization with chlorhexidine soap would reduce hospital-based infections (Huang et al. 2019), investigators considered the possibility that physician decisions regarding sending clinical cultures may be affected by knowledge of the intervention and/or the trial. As another example, in STRIDE, seeking medical attention after a fall was part of the original primary outcome definition because it was considered a measure of injury severity. However, as part of the trial, participants were

assigned fall care managers, and these care managers could therefore be asked by participants about seeking medical attention. To address this problem, for specific types of injuries where seeking medical attention was thought to be more discretionary, the definition of the primary outcome was changed from seeking medical attention to requiring an overnight stay in the hospital. The revised definition is expected to lead to fewer events, and future work may include a quantitative assessment of the bias and variance tradeoff.

Panel 2. To Cluster or Not to Cluster?

Pragmatic trials embedded in health care delivery systems should be designed with the organizational structure in mind, as individual patients are typically nested within providers, clinics, and higher-level organizational units. During this panel, PIs and statisticians discussed the trade-offs associated with clustering and elements of intervention delivery and analytical approaches that address this multi-level structure.

Moderator, Keith Goldfeld

Miguel Vazquez (PI) and Chul Ahn (Statistician) for ICD-Pieces

- Improving Chronic Disease Management with Pieces (ICD-Pieces)

Lynn DeBar (PI) and William Vollmer (Statistician) for PPACT

- Collaborative Care for Chronic Pain in Primary Care (PPACT)

Greg Simon (PI) and Susan Shortreed (Statistician) for SPOT

- Suicide Prevention Outreach Trial (SPOT)

The decision whether to cluster randomize or not depends on at least two key factors: (1) the nature of the intervention (who is providing it, to whom, and in what context) and (2) the structure of the health care system and patterns of treatment. Even if cluster *randomization* is avoided, *analysis* including cluster or group-level effects may still be required based on the nesting of patients within treatment arm (e.g., if the intervention is group based or if the intervention is delivered through a set of providers). If the cluster, or group, is nested within the treatment condition, investigators may be learning less than expected from individuals, and clustering needs to be addressed in the design and analysis. If the group is *not* nested in the treatment

condition, then stratification, as is typically done to account for clinic variation within individually randomized trials, may be an option to help gain precision regarding the intervention effect. In other words, it is important to consider both the randomization and whether there is clustering in the allocation of assignment, and the delivery of care to weigh potential clustering of either intervention or control outcomes even when using individual randomization.

When deciding the unit for randomization and analysis, many challenges need to be considered. (1) Data may not be readily available that detail the structure of clustering (e.g., may not have provider information for privacy protection), or the clustering structure may be too complex to apply current analytic methods, such as with multiple levels of clustering (practices within facilities within systems) or crossed clustering (patients may see multiple providers, or providers may practice at multiple sites). If the cluster structure is too complex, then it may not be possible to conduct a cluster-randomized trial. (2) Standard analytic methods for certain outcomes or clustering structures may not be available for planning the study (e.g., when cluster sizes are highly variable [see Unequal Cluster Sizes in Cluster Randomized Clinical Trials]). (3) Outcomes based on long follow-up periods may expose a study to greater risk of contamination (in addition to other challenges that any long-term study faces, like increased probability of loss to follow-up). This may be a reason why cluster randomization should not be selected. (4) Extreme heterogeneity of clusters may preclude adequate balance when only a small to moderate number of clusters can be randomized. (5) Unanticipated implementation issues (e.g., slow or incomplete intervention implementation or changes in systems that conflict/compete with intervention) may generate additional complex analytic challenges, for which data may not be readily available.

The panel provided three case examples to illustrate different rationales regarding the decision to cluster randomize or not.

Practice-level randomization

ICD-Pieces involved a care delivery intervention to improve outcomes for patients with co-morbid conditions (kidney disease, diabetes, and hypertension). While the intervention was at the physician level, the study included diverse health care systems (including safety net, private insurance, and VA); thus, the investigators decided it was necessary to stratify randomization by health care system to balance the types of clinics in each arm. A key question in the design of the trial was the ultimate unit of randomization: clinic, practice within clinic, or individual physicians within practice or clinic. Because (1) two of the systems had a small number of clinics and (2) there was concern about contamination across individual providers within a practice, the investigators opted for randomization at the practice level to gain more power and reduce contamination. An important challenge remained: cluster sizes at the practice level were quite variable, and because of this they needed to develop tailored analytic methodology to determine power/sample size. Specifically, the statisticians developed sample size formulas that incorporate variable cluster sizes for binary and continuous outcomes in stratified cluster randomization trials (Wang et al. 2017; Xu et al. 2019).

Physician-level randomization

PPACT provided a team-based care delivery aimed at patients on long-term opioid therapy for chronic pain, and the study included three diverse Kaiser systems (Northwest, Georgia, and Hawaii) (DeBar et al. 2018). The intervention was delivered by an interdisciplinary team (nurse, physical therapist, behavioral health, pharmacist) directly to a group of patients, so in theory they could have considered individual randomization, but did not. However, even if randomization had been at the patient level, there still would have been structural clustering at the treatment delivery level, which would need to be addressed in the analysis. The PPACT research team initially considered clustering at the clinic level, but determined there was too much heterogeneity across the relatively small number of clinics and were concerned about balancing clinic characteristics across treatment arms. Ultimately, investigators decided to randomize at the primary care provider (PCP) level. Approximately 10 patients per PCP were enrolled. For PCPs with small panels of patients, patients from

two PCPs were combined to form a randomization cluster. The planned analysis was going to include a PCP-level random effect, though not a cluster random effect (which in most cases would have been the same). In this study, the realities on the ground—primarily heterogeneity of clinics, fluidity of intervention, and inclusion of very small clinics—drove much of the design, providing challenges for the ultimate analysis of the outcome.

Individual randomization

SPOT was a three-arm, individually randomized trial targeting patients identified as at-risk for suicide (Simon et al. 2016). The study was conducted within four large, integrated health care systems. The two active interventions (one based on care management and the other on developing skills) were largely delivered directly to the patient through text and phone. The care management arm directly tried to engage patients with their health care providers (psychiatrists, PCPs, therapists, etc.) when assessed risk was elevated. Randomization was carried out at the patient-level for two main reasons. First, the intervention was carried out at the patient level. Second, patients receive mental health care from multiple providers which given the fluid nature of mental health care would likely lead to contaimination given the long intervention and follow-up period. The intervention period was 12 months and the outcome was suicide attempt in the 18 months following randomization. If randomization was conducted a cluster level (either provider or clinic), a provider would still likely see patients from different intervention arms defeating the purpose of cluster-level randomization. Although it may be desirable to adjust for this structural clustering analytically, the complexity of the patterns of care, which involve multiple visits and multiple providers over an extended time frame, may limit the feasibility of complex tailored analyses.

Panel 3. Choosing a Parallel Group or Stepped-Wedge Design

Cluster-randomized trials are often limited in the number of clusters available for study, and therefore a variety of design alternatives are considered. One contemporary design is

the stepped-wedge trial where all clusters start out as controls, and the intervention is then "activated" over time, cluster by cluster, as determined by randomization. The design leverages repeated measurements of clusters and allows each cluster to be observed in both intervention and control states. However, there are pros and cons to this design, and a parallel-group design is an alternative option where clusters are randomized to intervention or control conditions and remain under the same condition throughout the trial. Parallel-group cluster randomization may be desirable especially when studying complex conditions that are subject to temporal trends.

Moderator, Fan Li

Jerry Jarvik (PI) and Patrick Heagerty (Statistician) for LIRE

- Lumbar Imaging with Reporting of Epidemiology (LIRE)

Ted Melnick (PI) and Jim Dziura (Statistician) for EMBED

- Pragmatic Trial of User-Centered Clinical Decision Support to Implement Emergency Department-Initiated Buprenorphine for Opioid Use Disorder (EMBED)

Doug Zatzick (PI) Patrick Heagerty (Statistician) for TSOS

- A Policy-Relevant U.S. Trauma Care System Pragmatic Trial for PTSD and Comorbidity (Trauma Survivors Outcomes and Support [TSOS])

A key advantage of the stepped-wedge design over the parallel design is that the former allows a staggered rollout of intervention according to a pre-specified schedule. The successive rollout of intervention ensures that all health care units receive intervention before the end of the study, and is particularly attractive for interventions that are thought to be effective (e.g., disseminating an intervention to a different population) with minimum harm to providers and patients. Both the LIRE and TSOS studies adopted the stepped-wedge design due to such perceived benefit across health care systems. In fact, the TSOS study originally proposed a parallel design, but switched to a stepped-wedge design due to foreseeable logistical convenience and perceived ethical benefit (Zatzick et al. 2016). The logistical resources and efforts may be less demanding if the intervention is rolled out according to a staggered schedule. By comparison, the parallel design may only allow half the health care units to receive the intervention before the end of study, which may not be viewed positively when health care systems are considering participation (Murray 1998; Cook et al. 2016). However, although a parallel design requires more initial resources to implement the

intervention concurrently in half of the population, it does not involve multiple time periods for data collection and generally results in a shorter trial duration, which permits an earlier time schedule for post-study dissemination and implementation activities. This last point is among the reasons why the EMBED study switched from a stepped-wedge to a parallel design. EMBED is designed to help address the opioid crisis by assessing the effectiveness of the user-centered clinical decision support system on increasing adoption of initiation of buprenorphine (BUP) into the routine emergency care of individuals with opioid use disorder (Melnick et al. 2019). The study proposed a stepped-wedge design during the planning phase but switched back to a parallel design due to the high likelihood of confounding by temporal trends from ongoing efforts to mitigate the opioid epidemic.

The potentially longer duration of a study with stepped-wedge design makes it more susceptible to uncontrolled treatment factors changing outside of the study, which make it more difficult to describe the actual intervention received, potentially affecting implementation in the future. This is a major consideration for adopting such a design. An additional implementation challenge for the stepped-wedge design is trial protocol delay and pauses, which may inevitably delay the rollout of interventions and data collection activities. Both the LIRE and TSOS studies experienced some protocol delay but eventually resolved these issues by communicating with the site management teams to troubleshoot the social and political problems.

Finally, given the site heterogeneity across multiple domains, the stepped-wedge design may offer the advantage of having each cluster contribute patients to both the intervention and control condition. Therefore, while the parallel design only allows for between-cluster comparisons, the stepped wedge design allows for both within-cluster (horizontal) comparisons and between-cluster (vertical) comparisons (Hussey and Hughes 2007; Matthews and Forbes 2017). In fact, in the absence of time effect, such as in a short trial, it has been demonstrated that the stepped-wedge design always improves the power over the parallel design (Zhou et al. 2017). However, with non-negligible time effect, the relative efficiency between these two designs depends on the

magnitude of correlation parameters and other design resources (Hemming and Girling 2013; Woertman et al. 2013; Hemming and Taljaard 2016). The potential efficiency advantage supports the LIRE and TSOS studies in achieving the desired level of statistical power with a limited number of health care units, and is also among the reasons why the EMBED study proposed the stepped-wedge design in the first place. Besides efficiency, an equally important consideration in choosing the stepped-wedge design is the ability to control for the unobserved time trends, defined as the outcome trajectory in the absence of intervention. Current practice in analyzing stepped-wedge designs assumes a categorical time indicator in the fixed-effects component and only occasionally accounts for the between-cluster heterogeneity in the time trends (Nickless et al. 2018). Failure to account for between-cluster heterogeneity in the time trends could result in bias of the intervention effect estimate and incorrect type I error rate. In fact, the concern of current practice in dealing with time effects prompted the EMBED study team to switch back to a parallel design, especially because there is a high likelihood of confounding by temporal trends from ongoing efforts to mitigate the opioid epidemic. Further, in trials with a longer duration (such as the UK sweeping study in (Hemming et al. 2017), the categorical time effect specification may cost too many parameters and degrees of freedom, leading to reduced precision of the intervention effect. From this perspective, it would be helpful to borrow ideas from the classical longitudinal data analysis and model the temporal trend as parametric or semi-parametric curves with splines or polynomial structures (Hemming et al. 2017; Nickless et al. 2018).

Panel 4. Unique Complications

Embedded pragmatic clinical trials often encounter challenges that are associated with research embedded in a dynamic delivery system environment. Issues include questions about appropriate consent, strategies for monitoring trials regarding conduct quality and patient safety, and plans for handling unplanned changes in the research environment.

Moderator, Andrea Cook

Myles Wolf (PI) and Hrishikesh Chakraborty (Statistician) for HiLO

- Pragmatic Trial of Higher vs. Lower Serum Phosphate Targets in Patients Undergoing Hemodialysis (HiLo)

Beverly Green (PI) and William Vollmer (Statistician) for STOP CRC

- Strategies and Opportunities to Stop Colorectal Cancer (STOP CRC)

Laura Dember (PI) J. Richard Landis and Jesse Hsu (Statisticians) for TIME

- Time to Reduce Mortality in End-Stage Renal Disease (TiME)

Because ePCTs are embedded in healthcare delivery systems, unique challenges occur, such as the need to potentially consent individuals in a cluster-randomized trial, issues involved in ascertaining outcomes, dynamic patient populations, unanticipated delays, and issues with implementing the intervention. The HiLo trial provided an example of two of these issues; the trial tested which phosphate management strategy (high or low) would confer lower rates of all-cause mortality and hospitalization in patients with end-stage renal disease. Because the intervention was greater than minimal risk, individual consent was needed, and yet the trial was randomized by dialysis facility. The team solved this problem by providing tablets in dialysis units with video and paper consent. Another challenge was that the primary outcome of HiLo was all-cause hospitalization. In patients with end-stage renal disease, hospitalizations are associated with greater severity of disease; however, some patients fear hospitalization more than death, and there are patients who die without being hospitalized. To counter this problem, the investigators switched to a hierarchical endpoint of all-cause mortality followed by all-cause hospitalizations.

The STOP CRC trial, which was conducted in 26 Federally Qualified Health Centers, experienced delays that affected their patient population. Clinic membership was defined as having a visit within the past 12 months, and eligible patients were accrued after clinic randomization, but fell off the registry list if 12 months passed without additional visits. At one site, a system-wide EHR upgrade delayed intervention start-up by 4 months, and clinic training delays led to even longer delays, resulting in patients being removed from the list of patients expected to receive the interventions but who remained in the STOP CRC trial denominator. To account for this, investigators

performed a secondary lagged analysis evaluating patients who were accrued after the delays. This lagged analyses resulted in a net increase in uptake of the intervention (Coronado et al. 2018).

The TiME trial was a cluster randomized parallel group trial conducted in 266 free-standing outpatient dialysis facilities operated by two national dialysis providers (Dember et al. 2019). The trial enrolled 7,035 patients and the implementation of the trial was highly centralized and used no on-site research staff. The trial was designed to determine if longer hemodialysis sessions improved survival and reduced hospitalizations for patients with end-stage renal disease. Facilities were randomized to either the intervention group (hemodialysis session duration of at least 4.25 hours) or the usual care group (no trial driven session duration). Implementation of the 4.25 hour sessions at the intervention facilities was challenging and lower than anticipated. Potential contributors to this difficulty were 1) patient and nephrologists factors, 2) facility factors, or 3) dialysis provider organization factors.

This was an extremely rich, multi-level cluster design, with 252 facilities, 7,035 patients, and 1,129,867 hemodialysis sessions. The TiME investigators and statisticians were able to obtain granular data regarding both prescribed and delivered session duration for all participants. There was substantial heterogeneity in fidelity to the intervention. An analysis of the relative contributions of session, patient, facility, and dialysis provider organization to the variance of session durations within 5 different session duration categories (<210, 201-<225, 225-<240, 240-<255, and ≥255 minutes) in the intervention and usual care groups indicates that, in the intervention group, the facility was an important contributor to variability in implementing the ≥4.25 hour (≥255 minutes) session duration (the target for the intervention facilities) but not to variability in session durations of <4.25 hours.

These findings reinforce the need to understand patient-level, nephrologist -level, and facility-level factors that would allow a more responsive uptake of the intervention. For

future work, it would be interesting to know which facility factors that contributed to the high adherence in some of the facilities but not in others.

Resources

These documents focus on detailed aspects of statistical design for conducting pragmatic clinical trials:

- Key Issues in Extracting Usable Data from Electronic Health Records for Pragmatic Clinical Trials
- The Intraclass Correlation Coefficient
- Unequal Cluster Sizes in Cluster-Randomized Clinical Trials
- Pair-Matching vs Stratification in Cluster-Randomized Trials
- Frailty Models in Cluster-Randomized Trials
- Small-Sample Robust Variance Correction for Generalized Estimating Equations for Use in Cluster-Randomized Trials
- Assessing Data Quality for Healthcare Systems Data Used in Clinical Research (Version 1.0)
- Analyses of Randomized Controlled Trials in the Presence of Noncompliance and Study Dropout

NIH Collaboratory Living Textbook Sections:

- Cluster Randomized Trials
- Choosing Between Cluster and Individual Randomization

References

Cook AJ, Delong E, Murray DM, Vollmer WM, Heagerty PJ. 2016. Statistical lessons learned for designing cluster randomized pragmatic clinical trials from the NIH Health Care Systems Collaboratory Biostatistics and Design Core. Clin Trials Lond Engl. 13(5):504–512. doi:10.1177/1740774516646578.

Coronado GD, Petrik AF, Vollmer WM, Taplin SH, Keast EM, Fields S, Green BB. 2018. Effectiveness of a Mailed Colorectal Cancer Screening Outreach Program in Community Health Clinics: The STOP CRC Cluster Randomized Clinical Trial. JAMA Intern Med. 178(9):1174. doi:10.1001/jamainternmed.2018.3629. [accessed 2019 Aug 23]. http://archinte.jamanetwork.com/article.aspx?doi=10.1001/jamainternmed.2018.3629.

DeBar L, Benes L, Bonifay A, Deyo RA, Elder CR, Keefe FJ, Leo MC, McMullen C, Mayhew M, Owen-Smith A, et al. 2018. Interdisciplinary team-based care for patients with chronic pain on long-term opioid treatment in primary care (PPACT) - Protocol for a pragmatic cluster randomized trial. Contemp Clin Trials. 67:91–99. doi:10.1016/j.cct.2018.02.015.

Dember LM, Lacson E, Brunelli SM, Hsu JY, Cheung AK, Daugirdas JT, Greene T, Kovesdy CP, Miskulin DC, Thadhani RI, et al. 2019. The TiME Trial: A Fully Embedded, Cluster-Randomized, Pragmatic Trial of Hemodialysis Session Duration. J Am Soc Nephrol JASN. 30(5):890–903. doi:10.1681/ASN.2018090945.

Hemming K, Girling A. 2013. The efficiency of stepped wedge vs. cluster randomized trials: stepped wedge studies do not always require a smaller sample size. J Clin Epidemiol. 66(12):1427–1428. doi:10.1016/j.jclinepi.2013.07.007.

Hemming K, Taljaard M. 2016. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. J Clin Epidemiol. 69:137–146. doi:10.1016/j.jclinepi.2015.08.015.

Hemming K, Taljaard M, Forbes A. 2017. Analysis of cluster randomised stepped wedge trials with repeated cross-sectional samples. Trials. 18(1):101. doi:10.1186/s13063-017-1833-7.

Huang SS, Septimus E, Kleinman K, Moody J, Hickok J, Heim L, Gombosev A, Avery TR, Haffenreffer K, Shimelman L, et al. 2019. Chlorhexidine versus routine bathing to prevent multidrug-resistant organisms and all-cause bloodstream infections in general medical and surgical units (ABATE Infection trial): a cluster-randomised trial. Lancet Lond Engl. 393(10177):1205–1215. doi:10.1016/S0140-6736(18)32593-5.

Hussey MA, Hughes JP. 2007. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials. 28(2):182–191. doi:10.1016/j.cct.2006.05.007.

Matthews JNS, Forbes AB. 2017. Stepped wedge designs: insights from a design of experiments perspective. Stat Med. 36(24):3772–3790. doi:10.1002/sim.7403.

Melnick ER, Jeffery MM, Dziura JD, Mao JA, Hess EP, Platts-Mills TF, Solad Y, Paek H, Martel S, Patel MD, et al. 2019. User-centred clinical decision support to implement emergency department-initiated buprenorphine for opioid use disorder: protocol for the pragmatic group randomised EMBED trial. BMJ Open. 9(5):e028488. doi:10.1136/bmjopen-2018-028488.

Mor V, Volandes AE, Gutman R, Gatsonis C, Mitchell SL. 2017. PRagmatic trial Of Video Education in Nursing homes: The design and rationale for a pragmatic cluster randomized trial in the nursing home setting. Clin Trials Lond Engl. 14(2):140–151. doi:10.1177/1740774516685298.

Murray DM. 1998. Design and analysis of group-randomized trials. New York: Oxford University Press (Monographs in epidemiology and biostatistics).

Nickless A, Voysey M, Geddes J, Yu L-M, Fanshawe TR. 2018. Mixed effects approach to the analysis of the stepped wedge cluster randomised trial-Investigating the confounding effect of time through simulation. PloS One. 13(12):e0208876. doi:10.1371/journal.pone.0208876.

Simon GE, Beck A, Rossom R, Richards J, Kirlin B, King D, Shulman L, Ludman EJ, Penfold R, Shortreed SM, et al. 2016. Population-based outreach versus care as usual to prevent suicide attempt: study protocol for a randomized controlled trial. Trials. 17(1):452. doi:10.1186/s13063-016-1566-z.

Wang J, Zhang S, Ahn C. 2017. Power analysis for stratified cluster randomisation trials with cluster size being the stratifying factor. Stat Theory Relat Fields. 1(1):121–127. doi:10.1080/24754269.2017.1347309. [accessed 2020 Mar 19]. https://www.tandfonline.com/doi/full/10.1080/24754269.2017.1347309.

Woertman W, de Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL, Teerenstra S. 2013. Stepped wedge designs could reduce the required sample size in cluster randomized trials. J Clin Epidemiol. 66(7):752–758. doi:10.1016/j.jclinepi.2013.01.009.

Xu X, Zhu H, Ahn C. 2019. Sample size considerations for stratified cluster randomization design with binary outcomes and varying cluster size. Stat Med. 38(18):3395–3404. doi:10.1002/sim.8175.

Zatzick DF, Russo J, Darnell D, Chambers DA, Palinkas L, Van Eaton E, Wang J, Ingraham LM, Guiney R, Heagerty P, et al. 2016. An effectiveness-implementation hybrid trial study protocol targeting posttraumatic stress disorder and comorbidity. Implement Sci IS. 11(1):58. doi:10.1186/s13012-016-0424-4.

Zhou X, Liao X, Spiegelman D. 2017. “Cross-sectional” stepped wedge designs always reduce the required sample size when there is no time effect. J Clin Epidemiol. 83:108–109. doi:10.1016/j.jclinepi.2016.12.011.