Q&A

Mind the Gap Webinar: When is the Stepped Wedge Study a Good Study Design Choice?

Karla Hemming, Ph.D.

January 21, 2022

Q: Are there additional steps researchers should follow when using a group-based randomized stepped-wedge design to rule out biases? For example, one may expect secular trends to differ across intervention groups/clusters due to the group-based intervention where participants have contact with one another.

A: The main biases in cluster trials likely come from selectively recruiting or identify systematic different participants in treatment and control conditions. Mechanisms to avoid these biases include recruiting before randomization, or where that is not possible recruiting blind to treatment condition. Temporal biases are a potential source of bias in SW trials with a small number of clusters and where the assumptions made at the analysis stage (such as a common secular trend) are not met. These might be mitigated in large samples (trends will balance across sequences) or in small settings by recruiting clusters that are likely to be more homogenous in their trends.

Q: is there a program or online calculator for performing power calculations for a stepped wedge design? Something for preparatory work that is accessible to operational folks at health systems.

A: There are lots of online calculators to determine SS for SW and cluster trials. This app has a link to many if not all of them:

https://douyang.shinyapps.io/swcrtcalculator/

Q: Can you say more about strategies for adequately adjusting standard errors for within-cluster correlation?

A: Standard errors need to allow for both the correlation within clusters and over time. There are a few nice papers on this, here is one:

Kasza J, Hemming K, Hooper R, Matthews J, Forbes AB. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. Stat Methods Med Res. 2019 Mar;28(3):703-716. doi: 10.1177/0962280217734981. Epub 2017 Oct 13. PMID: 29027505.

And this paper contains a summary of these issues:

Hemming K, Kasza J, Hooper R, Forbes A, Taljaard M. A tutorial on sample size calculation for multiple-period cluster randomized parallel, cross-over and stepped-wedge trials using the Shiny CRT Calculator. Int J Epidemiol. 2020 Jun 1;49(3):979-995. doi: 10.1093/ije/dyz237. PMID: 32087011; PMCID: PMC7394950.

They also need to allow for small samples, in settings where there are less than about 40 clusters. In GEEs these small sample corrections have been evaluated for SW trials here:

Thompson JA, Hemming K, Forbes A, Fielding K, Hayes R. 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 Feb;30(2):425-439. doi: 10.1177/0962280220958735. Epub 2020 Sep 24. PMID: 32970526; PMCID: PMC8008420.

Q: I am curious about the secular effects. Can you describe the kinds of secular effects that might have differential impacts across the clusters when testing a clinical intervention.

A: Temporal trends are probably a lot more common than anticipated. Generally health is improving and over a long study (say 3 to 5 years) some of these improvements would be expected to have an impact even in absence of the intervention. Other trends can be kick started by what some people have called the rising tide: the notion that a particular issue is a problem in health prompts investigators to explore interventions, but at a time when many others are also exploring and implementing interventions. This paper includes an example where there was probably a secular trend:

Hemming K, Taljaard M, Forbes A. Analysis of cluster randomised stepped wedge trials with repeated cross-sectional samples. Trials. 2017 Mar 4;18(1):101. doi: 10.1186/s13063-017-1833-7. PMID: 28259174; PMCID: PMC5336660.

Q: You mentioned that we must use small sample corrections when low number of clusters. Could you speak into how to interpret sample size calculations when showing sufficient power but small numbers of clusters? Are there any power calculation techniques that take into account the analytical burden of small sample size corrections?

A: In parallel cluster trials sample size can allow for small sample corrections, approximately, by adding 2 to 4 extra clusters per arm. In SW trials exactly how sample size can allow for small numbers of clusters is still unknown.

Q: What can be done to test for bias in an SW-CRT when you have a small number of clusters?

A: Testing for bias is very hard. Information about the trial processes is usually helpful when trying to critically appraise / identify if a CRT might be at risk over identification and recruitment bias. Bias in the estimation of the treatment effect due to a poorly specified model are harder to detect. Results that are robust to a range of sensitivity analyses (i.e. modeling time effects in different ways) are likely to be more reliable.

Q: Might Dr Hemming share her email or contact information? This is an excellent presentation

A: k.hemming@bham.ac.uk

Q: Can you elaborate on the difference of "needed" clusters in parallel cluster design over stepped wedge? I don't quite understand why there is a difference...

A: In the setting where the cluster size, power and target difference are specified power calculations allow you to determine the number of clusters "needed" under a parallel design and the number "needed" under a SW design. These can be different and whether the SW will require more or less than a parallel design is dependent on the specific circumstances.

Q: Is the SW-CRT design "more justifiable" for certain types of interventions, for example, some interventions may require a high degree of behavioral compliance than other interventions.

A: Cluster randomization is certainly more justifiable for certain types of interventions. For example, where the intervention required clinicians to change behaviors (they could not subsequently treat individual patients differentially). I don't believe this justification extends further to justifying rolling out the intervention to all clusters.

Q: can you share your R Shiny app?

A: https://github.com/karlahemming/Cluster-RCT-Sample-Size-Calculator

https://clusterrcts.shinyapps.io/rshinyapp/

Q: About that cartoon.... how do you reconcile this with implementation trials?

A: The cartoon suggests that it is a myth that the SW design is a good choice where "it is expected the intervention will work". See Binik A. Delaying and withholding interventions: ethics and the stepped wedge trial. J Med Ethics 2019;45:662–67. In implementation trials, an "implementation strategy" is tested to see if it can improve / nudge health care providers to adopt interventions/ treatments that are already known to be effective. The "implementation strategy" is what is being evaluated, not the interventions/ treatments that are already known to be effective. Whether or not the implementation strategy works or does not work is unknown, hence why it is being tested.

Q: Is a purely randomization-based analysis for a cluster-randomized trial practical, or must we always rely on model-based results?

A: Randomisation based analysis for CRTs do not necessarily reply on model based results. These are strong contenders for approaches that make minimal assumptions. They are however less powerful and rarely used.

Q: Re the power argument. I hear that a priori the SW-CRT may be able to achieve 80% power where the simple, parallel CRT cannot, but at the end of the day only the evidence that is actually collected matters. So, in terms of post-study evidence, precision, and confidence interval size, does the SW-CRT ever beat the simple CRT?

A: Power and precision have a one to one relationship. If the SW-CRT for a particular scenario is more powerful it will also be more precise. Whether this holds at the analysis stage will depend on whether assumptions made about key parameters such as outcome prevalence and ICC were correct.

Q: One slide mentioned that we could learn from early sites in the SW-RCT. If we learn from early sites and use those lessons for later sites, does that not introduce bias?

A*: If the intervention changes over the course of the roll-out then an estimated effect will typically be some sort of average over time periods. This is not a bias, but it might be quite different to the effect at any given point.*

Q: It seems this largely focus on efficacy trials, does this also apply to effectiveness implementation trials where you also want to include a cost-effectiveness component

A: Yes, everything discussed applies to implementation trials too. See one question above.

Q: In a staggered parallel design would adjustment for time be necessary?

A: *Adjustment for time would probably improve precision; but unadjustment for time would not lead to a biased treatment effect (time is balanced in a staggered parallel design).*