A look at planning and designing stepped wedge trials

Faculty Disclosure

In compliance with ACCME Guidelines, I hereby declare:

I do not have financial or other relationships with the manufacturer(s) of any commercial services(s) discussed in this educational activity.

Name: Kevin Dobbin

Title: Associate Professor

MOREHOUSE SCHOOL OF MEDICINE

Background

- Stepped wedge trials are for interventions randomized at the group/cluster level.
- Examples: Clinics, hospitals, or medical practices can be randomized to a control vs. intervention
- Traditionally, cluster randomized designs (CluRD) have been the standard.
- Stepped wedge (SWD) design newer alternatives.

Stepped wedge vs. cluster randomized

Parallel

		Time: 1
Cluster	1	1
	2	1
	3	0
	4	0

Crossover

		Time: 1	2
Cluster	1	1	0
	2	1	0
	3	0	1
	4	0	1

Stepped Wedge

		Time: 1	2	3	4	5
Cluster	1	0	1	1	1	1
	2	0	0	1	1	1
	3	0	0	0	1	1
	4	0	0	0	0	1

Download high-res image (42KB) Download full-size image

Fig. 1. Treatment schedules for parallel, crossover, and stepped wedge designs. "0" represents con
existing treatment; "1" represents an intervention.

[From Hussey and Hughes, Cont. Clin. Trials, 2007]

Background: SWD advantages

- All clusters/groups receive the intervention eventually.
  - This can make recruiting sites easier.
- Statistical power can be increased
  - This can be a big deal when total number of clusters is limited.
- "Rolling out an intervention" can be advantageous.
  - Facilitates "learn as you go."
  - Trainers can travel from site to site.

Background: controversy

Kotz et al. (J. CLin. Epid., 2012) criticized the stepped wedge design.

- For every plus they found a minus
  - + SWD means everyone gets the intervention
  - - One can do this with a CluRD
  - + SWD sequential implementation advantageous
  - - Sequential implementation can be used in CluRD
  - + SWD requires fewer clusters than CluRD
  - - More measurements are required to model time

CluRD

Response can be binary, continuous, survival, etc.
With a binary response:

$$Logit(Pr(Y_{ij} = 1)) = \mu + \alpha_i + \theta\, Trt_i$$

- α_i represents the group effect, $i = 1, ..., g$ (fixed or random)
- Trt_i is the indicator of whether or not group i is treated
- θ is the parameter of interest.

If fixed effect: $\sum_i \alpha_i = 0$.

Sources of variation

- Different observations in the same group/cluster are usually positively correlated.
- Individual observations in same group are not independent.
- Observations in different groups are assumed independent.

The intracluster correlation coefficient is

$$ICC = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_e^2}$$

The larger the ICC, the bigger the cluster effect relative to the individual effect.

Degrees of freedom

$$n_{cluster} = number\ of\ clusters\ (total\ both\ groups)$$
$$n_p = number\ of\ participants\ per\ cluster$$

To fix ideas, we'll assume 2 groups.

- Inferences can be made at the cluster or individual level.
- Cluster level
  - Analyses tend to be cleaner and simpler, with fewer assumptions.
  - Error DF: $n_{cluster} - 2$
- Individual level
  - Analyses tend to require additional modeling assumptions.
  - Error DF: Depend on the ICC.

CluRD inflation factors

- The inflation factor for a cluster randomized design is

$$IF_{CluRD} = (1 + (n_p - 1)ICC)$$

(e.g., Donner and Klar, 2000)

- This formula can be applied to a continuous response or a binary response.

The basic approach is to calculate the sample size required using a completely randomized design (CoRD), and then multiply by the IF.

$$n_{CluRD} = n_{CoRD} \cdot IF_{CluRD}$$

Example

Objective: Detect an increase in proportion from 0.30 to 0.40 with 90% power at a two-sided $\alpha = 0.05$.

$$\begin{aligned} n_{CoRD} &= \frac{(z_{\alpha/2} + z_{\beta})^2(0.3 * 0.7 + 0.4 * 0.6)}{(0.4 - 0.3)^2} \\ &= 476.1 \end{aligned}$$

So with $n_p = 50$ per cluster, the number of clusters per group required for CoRD is

$$476.10/50 = 9.5 \textit{ round up to } 10$$

Example: CluRD adjustment

$$
\begin{aligned}
ICC &= 0.05 \\
n_p &= 50 \\
n_{CluRD} &= 476 * (1 + (50 - 1) * 0.05) \\
&= 1642.2
\end{aligned}
$$

$1650/50 = 33$ clusters per group needed.

SWD

$$Logit(Pr(Y_{it} = 1)) = \mu + \alpha_i + \beta_t + \theta X_{it}$$

- α_i is group factor, $i = 1, ..., g$.
- β_t is time factor, $t = 1, ..., t$
- X is a matrix of treatment indicators representing design

Diagram for SW

Statistical model must disentangle intervention and time effects.

Cluster 1		I	I	I	I	I	I
Cluster 2			I	I	I	I	I
Cluster 3				I	I	I	I
Cluster 4					I	I	I
Cluster 5						I	I
Cluster 6							I
Times	1	2	3	4	5	6	7

Interaction

$$Logit(Pr(Y_{it} = 1)) = \mu + \alpha_i + \beta_t + \alpha\beta_{it} + \theta X_{it}$$

- An assumption of the SWD generally made is that there is not an interaction between time and group.
- In some applications, this may be dubious.
- The assumption could be checked partially by assuming $\alpha\beta_{it} = \alpha_i\beta_t$, and using a modification of the Tukey test for interaction/additivity.
  - This approach compares models each of which contains θ.
- See Scheffe (1959).

Random or fixed effects

$$Logit(Pr(Y_{it} = 1)) = \mu + \alpha_i + \beta_t + \theta X_{it}$$

- Hussey and Hughes (2006) used fixed effects for time and random effects for group.
- For a particular setting, this may or may not be the best choice, depending on the number of groups and times:
  - Mixed effects models can underperform in some settings.
  - Refs: Gelman (2005), Pinhero and Bates (2000)

Degrees of Freedom

- Assumptions
  1. Cluster level analysis
  2. Time and cluster are fixed effects
  3. $n_{times} = n_{cluster} + 1$
- The degrees of freedom for error are:

$$\begin{aligned} DF_{SWD} &= n_{cluster}(n_{cluster}+1) - n_{cluster} - (n_{cluster}-1) - 1 - 1 \\ &= n_{cluster}^2 - n_{cluster} \end{aligned}$$

Error DF table

Cluster level analyses with cluster as fixed effects and random effects

	Fixed		Random	
$n_{cluster}$	DF_{SWD}	DF_{CluRD}	DF_{SWD}	DF_{CluRD}
1	0	0	0	0
2	2	0	3	1
3	6	0	8	2
4	12	2	15	5
5	20	3	23	7
6	30	4	35	9
7	42	5	48	11
8	56	6	63	13
9	72	7	80	15

Cell variances

- The cell variances in the cluster level analyses are functions of the ICC.
- The total sample size for each cell in the CluRD is n_p.
- The total sample size for each cell in the SWD is $n_p/(n_{cluster}+1)$.

$$Var(cell\ mean) = \sigma^2 \left\{ \frac{1+(n_{cell}-1)ICC}{n_{cell}} \right\}$$

Cell std. dev. relations

Assumes fixed replicates per cluster, and 6 clusters

Table of τ_{SWD}/τ_{CluRD}

n_p	ICC=0.05	ICC=0.50
10	1.51	1.04
20	1.31	1.02
30	1.22	1.01
40	1.18	1.01
50	1.14	1.01
60	1.12	1.01
70	1.11	1.00

Normality

- By the Central Limit Theorem (CLT), means become more normal as sample size increases.
- The CluRD design will have larger sample (group) sizes.
- The SWD design will have smaller sample (group by time) sizes.

This could impact some statisical inferences.

Inflation factor

- Woertman et al. (2013) and Hemming and Taljaard (2016) develop inflation factors for the stepped wedge designs.

In our notation, the inflation factor is

$$IF_{SWD} = \frac{(n_c+2)*(1+ICC[n_p+(n_p/(n_c+1))-1])*3*(1-ICC))}{(1+ICC((n_p/2)+(n_p/(n_c+1)-1)))*2*[n_c+1-1/(n_c+1)]}$$

A bit opaque.

SW IF w.r.t. CoRD

Inflation factors for stepped wedge compared to CoRD

n_c	n_p	ICC	IF_{SWD}
6	10	0.05	1.99
6	50	0.05	2.48
12	10	0.05	1.86
12	50	0.05	2.35
6	10	0.50	1.46
6	50	0.50	1.53
12	10	0.50	1.41
12	50	0.50	1.49

IF SWD vs. CluRD

"Inflation" factors for stepped wedge compared to CluRD

n_c	n_p	n_{total}	ICC	IF_{SWD}/IF_{CluRD}
6	10	60	0.05	1.37
6	50	300	0.05	0.72
12	10	120	0.05	1.28
12	50	600	0.05	0.68
6	10	60	0.50	0.27
6	50	300	0.50	0.06
12	10	120	0.50	0.26
12	50	600	0.50	0.06

Discussion: SW

No time by cluster interaction assumption

- Is this assumption valid?
- If treatments are rolled out over time, it seems intuitively less likely to be true.
- If the assumption is violated, can one come up with a good model that leaves df for error?

SW model modification

Should there be a treatment by time interaction?

- This seems intuitively plausible – as time goes by, implementation may get better or worse.
- This would impact comparisons.

Negative (or worse) result

Equipoise is not being sure whether an intervention will help or not.

- It has been pointed out that SW may be putting the cart before the horse, since all clusters wind up with the intervention.
  - It seems that one has already decided it will work.
- So if the conclusion is that it doesn't work, or makes things worse?

So is SWD only appropriate when we don't have true equipoise?

Something for nothing?

- Some of the improvement in degrees of freedom can look "too good to be true."
- This is partly due to the change in units: from "group" in CluRD to "group-time" in SWD.
- This seems analogous to comparing a full plot to a split plot experiment.
  - Maybe an adjustment to account for the unit change is needed.

Obervations

- CluRD seems likely to remain the standard for controlled clinical trials.
- When number of clusters is very limited, the SWD seems appealing relative to CluRD.
- When n_p is large-ish and the ICC is large, the SWD also seems appealing.

Modifications

- Split clusters into two groups, one with a stepped wedge and the other with stepped wedge and sham intervention.
- Append the stepped wedge with a set of "always control" groups.

References

- Donner A and Klar N (2000) Design and analysis of Cluster randomization trials in health research. Wiley.
- Hussey MA and Hughes JP (2007) Design and analysis of stepped wedge cluster randomized trials. Contemp. Clin. Trials.
- Gelman A (2005) Analysis of variance – why it is more important than ever. The Annals of Statistics.
- Pinhero JC and Bates DM (2002) Mixed-effects models in S and S-plus.
- 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. Jnl of Clinical Epid.
- Hemming K and Taljaard M (2010) Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. J Clin. Epid.
- Kotz D, Spigt M, Arts ICW, Crutzen R, Viechtbauer W (2012) Use of the stepped wedge design cannot be recommended .... J Clin. Epid.