36\textsuperscript{th} Annual Meeting
Preconference Workshop
P8
Handout
The stepped wedge cluster randomised trial (SW-CRT)

Society for Clinical Trials Conference
Arlington, USA
May 17th 2015

Workshop on research and reporting methods for the stepped wedge cluster randomised controlled trial

Faculty

• Professor Richard Lilford
  – University of Warwick, UK
• Dr Karla Hemming
  – University of Birmingham, UK
• Dr Monica Taljaard
  – University of Ottawa, Canada
• Dr Alan Girling
  – University of Birmingham, UK
• Dr Steven Teerenstra
  – Radboud University, The Netherlands
<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:00 – 1:10</td>
<td>Welcome and Introductions</td>
<td>Reconciling scientific rigour with the service imperative</td>
</tr>
<tr>
<td>1:10 – 1:40</td>
<td>Professor Richard Lifford</td>
<td>The stepped-wedge cluster randomised trial</td>
</tr>
<tr>
<td>1:40 – 2:00</td>
<td>Dr Karla Hemmings</td>
<td>The stepped-wedge cluster randomised trial including case studies: The Epoch Trial and Depression Management in Nursing homes</td>
</tr>
<tr>
<td>2:00 – 3:00</td>
<td>Dr Monica Talligrad</td>
<td>Sample size calculations. What sample size do I need? How do I calculate it? And what will be the most efficient design for me?</td>
</tr>
<tr>
<td>3:00 – 3:15</td>
<td>Break</td>
<td></td>
</tr>
<tr>
<td>3:15 – 3:45</td>
<td>Dr Alan Girling</td>
<td>Comparing different study designs. Efficiency and extensions to the conventional design</td>
</tr>
<tr>
<td>3:45 – 4:15</td>
<td>Dr Steven Teerenstra</td>
<td>Three treatment arms. Stepped wedge designs for comparing three treatment conditions</td>
</tr>
<tr>
<td>4:15 – 4:45</td>
<td>Dr Karla Hemmings</td>
<td>Analysis of the SW-CRT. Why we must adjust for temporal confounding.</td>
</tr>
<tr>
<td>4:45 – 5:00</td>
<td>Professor Richard Lifford</td>
<td>Preliminary recommendations for reporting and discussion</td>
</tr>
</tbody>
</table>
Reconciling Scientific Rigour with the Service Imperative

Professor Richard Lilford

Society for Clinical Trials Conference
Arlington, VA
May 17th 2015

Collaborations for Leadership in Applied Health Research Care

Hospital
Primary Care
Local Authority

£20m
Intervention
Development
Evaluation

£10m

National Institute for Health Research
Step Wedge Design

• What?

• Why?

Today’s talk: *Cluster* trials
What is a Step Wedge Design?

Many variations on a theme

What is the Counter-factual?

Two Types of Cluster Study:

Cross-sectional

Cohort
Why Consider a SWD, rather than Parallel Cluster Trial?

1. Ethics
2. Acceptability
3. Logistics
4. Scientific Reasons
   - Sample size
   - Interaction between intervention effect and time

Disadvantages of SWDs?

<table>
<thead>
<tr>
<th>Always</th>
<th>Sometimes</th>
<th>Spurious</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytically complex</td>
<td>Violation of sequence</td>
<td>All cluster studies</td>
</tr>
<tr>
<td></td>
<td>Increase trial duration</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Long gap between intervention &amp; outcome</td>
<td></td>
</tr>
</tbody>
</table>
Step Wedge Trials vs. Step Wedge Studies

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

But may be “instrumental variable”

Naturalistic SWD = Instrumental variable approach

The End
Introduction: what is the SW-CRT?

Defining features, common variations, and some salient examples

Karla Hemming

Why do we need another method of evaluation?
Evaluation of policy and service delivery interventions

- The Matching Michigan Study
  - Secular trend
  - Intervention effect

- The Oregon Experiment
  - Randomised
  - No primary outcome

- Mexican universal health insurance
  - 74 matched clusters
  - Staggered implementation

Evidence based policy interventions

**Working constraints**
- Stakeholder’s desires
- Pragmatic limitations
- A priori beliefs

**How the SW-CRT can help**
- All clusters ultimately get intervention
- Sequential roll out
- Robust evaluation
What is the SW-CRT?

The Stepped Wedge Cluster Randomised Trial
Example 1: A cross-sectional SW-CRT

The EPOCH study – a cross sectional SW-CRT

- Evidence based integrated care pathway (ICP)
- Emergency laparotomy
- Setting:
  - 90 hospitals
- Outcome:
  - 90 day mortality
- Sample size:
  - 27,500 patients
  - 90% power to detect a change from 25% to 22%
- Routinely collected outcome
  - No consent

www.nets.nihr.ac.uk/projects/hsdr/12500510
Example 2: A cohort SW-CRT

Depression management— a cohort SW-CRT

- Structural multidisciplinary approach to depression management
- Participants
  - Residents who provided informed consent
- Setting:
  - 33 units within nursing homes (Dementia / Somatic)
- Outcome:
  - Depression prevalence
- Sample size:
  - 793 patients
  - 80% power to detect a 30 to 40% reduction in prevalence (circa 20-30%)
- Outcome:
  - Questionnaire (Cornell scale for depression)

Example 3 – the Gambia hepatitis study

The Gambia hepatitis study

- Study started 1980s
  - 30 year follow-up
- Vaccine
  - efficacy against hep B
- Main outcome
  - liver disease
- Vaccine rolled out
  - national infant vaccination schedule

- Step lengths 10 to 12 weeks
- National coverage after 4 years
- Geographically defined areas
Is the SW-CRT the right design for my trial?

- Pragmatic considerations
  - Does it allow a randomised evaluation which otherwise would not be possible?

- Logistical considerations
  - Allows sequential role out

- Efficiency
  - Minimise number clusters / participants / observations

- Duration
  - Will it necessarily extend the trial?
  - Time between exposure and follow-up

Word of caution …

- Lack of concealment of allocation
  - Risk of selection bias

- Avoid individual patient recruitment
  - Routinely collected outcome data

Chalmers: “Although one of the reasons that the streptomycin trial has become iconic is … random number tables …. it was because successful concealment of allocation”
Example of a CRT with lack of concealment of allocation

- Results in baseline imbalance
- Due to recruitment of individuals after allocation known
- Will it be better or worse in SW-CRT?

<table>
<thead>
<tr>
<th></th>
<th>Control (n=511)</th>
<th>Intervention (n=729)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number recruited (%)</td>
<td>500 (98%)</td>
<td>700 (96%)</td>
</tr>
<tr>
<td>Resting blood pressure (mmHg)</td>
<td>110 (10)</td>
<td>115 (12)</td>
</tr>
<tr>
<td>Systolic blood pressure (mmHg)</td>
<td>73 (16)</td>
<td>72 (14)</td>
</tr>
</tbody>
</table>

Variations to the common design

- More than two treatment comparisons
- Hybrid designs – mixtures of parallel designs and stepped studies
- Transition periods
- Multiple levels (i.e. clustering within clustering)
- Repeated measures (cohort designs)
To follow….

- What sample size do I need?
- What is the most efficient design?
- How to extend the design to all for more than two treatments?
- How do I analyse a SW-CRT?

Summary

- SW-CRT a pragmatic study design which reconciles the need for robust evaluations with political or logistical constraints.
- Unbiased design when:
  - No individual patient recruitment (routinely collected outcome)
- Efficient design when:
  - Higher the ICC (process outcomes)
  - Limited number of clusters
- Design and analysis
  - Appropriate consideration of time effects in power and analysis
References 1 (methodology)

- Mdege ND, Man MS, Brown CATN, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. Journal of Clinical Epidemiology 2011; 64:936–948.


References 2 (motivating examples)


Research and reporting methods for the stepped wedge cluster randomized trial: Sample size calculations
Society for Clinical Trials workshop
Arlington, Virginia
May 17, 2015

Monica Taljaard
Senior Scientist, Clinical Epidemiology Program
Ottawa Hospital Research Institute

Outline

1. Review of cluster randomized trials (CRTs)
   - Concepts, definitions, notation
   - Sample size calculation for parallel arm designs

2. Stepped wedge (SW) cluster randomized trials
   - Concepts, definitions, notation
   - Sample size calculation for balanced complete designs
   - Extensions of the balanced complete design

3. Comparing parallel vs. stepped wedge designs

4. Conclusions
1. Review of CRTs

**Cluster randomization**

- CRT: A trial in which intact social units, or clusters of individuals — *rather than individuals themselves* — are randomized to different intervention groups.
- Key implication of cluster randomization:
  - Responses of multiple individuals in the same cluster usually positively correlated.
  - Due to the presence of positive intracluster correlation, standard statistical methods are invalid.
Implications of cluster randomization

- Failing to account for intracluster correlation leads to invalid inferences

<table>
<thead>
<tr>
<th>Implications of failing to account for correlation in:</th>
<th>Sample size calculation</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>↑ P(Type II error)</td>
<td>↑ P(Type I error)</td>
</tr>
<tr>
<td>n too small</td>
<td>Apparent n too large</td>
<td>(std errors too small)</td>
</tr>
</tbody>
</table>

Intracluster correlation

- Usually quantified by $\rho$, the “Intracluster Correlation Coefficient” (ICC)
- Suppose $k$ clusters of $m$ individuals per cluster allocated to each of an intervention and a control arm
- Let $Y$ be the response variable with $\text{Var}(Y) = \sigma^2$
- Then

$$\sigma^2 = \sigma_{b}^2 + \sigma_{w}^2$$

where

$\sigma_{b}^2$ = between-cluster variance
$\sigma_{w}^2$ = within-cluster variance
Intracluster correlation (ctd)

- We define
  \[ \rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}, \]
  the proportion of the total variance that is between clusters

- For \( Y \) either dichotomous or continuous, \( \rho \) may be estimated using standard one-way ANOVA, as:
  \[ \hat{\rho} = \frac{S_b^2}{S_b^2 + S_w^2} = \frac{MSB - MSW}{MSB + (m - 1) MSW} \]


The design effect

- Quantify the effect of clustering by the “Design Effect”
  - In an individually randomized trial with \( n \) individuals per arm, we have:
    \[ \text{Var}(\bar{Y}_2 - \bar{Y}_1) = \frac{2\sigma^2}{n} \]
  - In a CRT with \( n = km \) individuals per arm, we have:
    \[ \text{Var}(\bar{Y}_2 - \bar{Y}_1) = \frac{2\sigma^2}{km} \left[ 1 + (m - 1) \rho \right] \]

“Design Effect” (Deff)
Increasing $k$ vs. increasing $m$

- Note that

\[ \text{Var}(\bar{Y}_2 - \bar{Y}_1) = \frac{2\sigma^2}{km} \left[ 1 + (m - 1)\rho \right] \]

implies that:
- We can always improve precision by increasing $k$
  - As $k \to \infty$, then $\text{Var} \to 0$
- Increasing cluster sizes has limited effect
  - As $m \to \infty$, then $\text{Var} \to \frac{(2\sigma^2 \rho)}{k} / k = 2\sigma^2 / k$

Sample sizes for CRTs

- Sample size formulas for CRTs are readily available for a range of designs and variable types
- We focus here on:
  - Completely randomized designs (no stratification or matching)
  - Continuous or dichotomous outcomes
  - Two study arms
  - Equal allocation

---

### Required parameters

- Need to specify, *apriori*:
  - Desired statistical power \((1-\beta)\) and significance \((\alpha)\)
  - Minimally important effect size
  - Standard deviation \((\sigma)\) (continuous outcome)
  - Control arm proportion (dichotomous outcome)
- Additionally for a CRT:
  - Anticipated ICC \((\rho)\)
  - Cluster size \(m\)
  - \(CV_m\) (Coefficient of variation of cluster sizes, if variable)

### Comparing two means

- To test \(H_0: \mu_1 = \mu_2\) the required number of subjects *per arm* is given by

\[
km = \left( z_{1-\alpha/2} + z_{1-\beta} \right)^2 \frac{2\sigma^2}{(\mu_1 - \mu_2)^2} \left[ 1 + (m-1)\rho \right]
\]

- Divide by \(m\) to determine the required number of clusters:

\[k = \frac{n_{ind} \times Deff}{m}\]

- Recommend adding one cluster per arm if \(k < 15\)
Accounting for variable cluster sizes

- If cluster sizes vary, calculate \( k \) using \( \bar{m} \)
- Use coefficient of variation of cluster sizes (\( CV_m \)) to adjust for loss in efficiency:

\[
k_{adj} = \frac{k}{1 - CV_m^2 \zeta (1 - \zeta)}, \text{ where } \zeta = \frac{\bar{m}\rho}{\bar{m}\rho + (1 - \rho)}
\]

- Note: Usually \( CV_m < 0.7 \) and since \( \zeta (1 - \zeta) \) reaches a maximum at 0.25, a \( \sim 12\% \) inflation is usually adequate


Limited numbers of clusters

- If number of clusters \( k \) is limited, design may not be feasible
- Design is infeasible if:

\[
k < n_{\text{ind}} \times \rho
\]

- If design is feasible, determine \( m \) for a given number of clusters by rearranging the formula:

\[
m = \frac{n_{\text{ind}} (1 - \rho)}{k - \rho n_{\text{ind}}}
\]

Comparing two proportions

- To test $H_0: p_1 = p_2$ the required number of subjects per arm is given by

\[ km = \left( \frac{z_{1-\alpha/2} + z_{1-\beta}}{} \right)^2 \left[ \frac{\pi_1 (1 - \pi_1) + \pi_2 (1 - \pi_2)}{} \right] \left[ 1 + (m - 1) \rho \right] \]

\[ = n_{ind} \times Deff \]

- Required number of clusters per arm:

\[ k = \frac{n_{ind} \times Deff}{m} \]

Obtaining an estimate for ICC

- Other trial reports (similar population and similar endpoint)
- Calculate using routine data sources
- Pilot study (not recommended unless total sample size is >200)
- Databases or publications that report lists of ICCs
- Extrapolate based on studies of determinants of ICCs (e.g., Adams 2004)
  - Clinical outcomes: typically $\rho \leq 0.05$
  - Process measures: $\rho$ typically larger, up to 0.15

Disadvantages of small $k$

- Note that the sample size formula can provide an unrealistic answer ($k$ too small)
- Problems associated with small $k$:
  - Study may be severely underpowered (formulas derived using large-sample theory)
  - Chance imbalances across study arms likely
  - Limits perceived or actual generalizability of results
  - Limited options for analysis
  - Estimates for the ICC likely imprecise
  - Substantial loss in power if even one cluster drops out

2. Stepped wedge CRTs
Balanced complete SW design

- $k$ clusters randomly allocated to one of $t$ steps or "wedges"
- All clusters start in control condition; clusters within each wedge cross to intervention sequentially until all have received intervention
- Outcome measured in each of $T=t+1$ measurement periods
- Cross-sectional samples of $m$ individuals per cluster in each measurement period
  - We will not consider cohort designs here

Balanced complete SW design

- Balanced complete SW design with $t=5$ steps ($T=6$ measurement periods)

<table>
<thead>
<tr>
<th>Wedges</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Exposed to intervention
- Unexposed to intervention
General modeling framework

- Regression analysis using random effects model:

\[ Y_{ijl} = \mu + \beta_j X_{ij} \theta + u_i + \varepsilon_{ijl}; \]
\[ u_i \sim N\left(0, \sigma_b^2\right); \quad \varepsilon_{ijl} \sim N\left(0, \sigma_w^2\right) \]

Where

- \( \beta_j \) = fixed effect of time
- \( X_{ij} = 1 \) if intervention, 0 otherwise
- \( \theta \) = intervention effect
- \( u_i \) = random effect for cluster \( i \)
- \( \varepsilon_{ijl} \) = residual

Hussey & Hughes (2007) showed that the power to detect an intervention effect $\theta_A$ is:

$$power = \Phi \left( \frac{\theta_A}{\sqrt{Var(\hat{\theta})}} - z_{1-\alpha} \right)$$

where $\Phi$ is the cumulative standard normal distribution.

For the balanced complete design:

$$Var(\hat{\theta}) = \frac{6(T-1)(1-\rho)}{mT(T-2)} \sigma^2 \left[ 1 + \left( \frac{m(T+1)}{2} \right) \rho \right]$$


Sample size for balanced complete design

- Design effect to determine sample size:

$$Deff = \frac{1 + \rho \left( \frac{tm + m - 1}{2} \right)}{1 + \rho \left( \frac{tm + m - 1}{2} \right)} \times \frac{3(1-\rho)}{2 \left( \frac{t-1}{t} \right)}$$

- Total required sample size:

$$km(t + 1) = N_{ind} \times (t + 1) \times Deff$$

where $N_{ind}$ = TOTAL number of subjects required under individual randomization.

Sample size for balanced complete design

- Determine the required number of clusters by simple division:
  \[ k = \frac{N_{ini} \times \text{Deff}}{m} \]

- Some practical notes:
  - Need to specify number of steps \( t \) in advance
  - \( t \) is usually specified with due consideration of logistics, planned study duration, and available \( m \)
  - Measurement period must be of sufficient duration to allow the intervention effect to be realized
  - \( k \) not necessarily a multiple of the number of steps (round up to obtain a multiple of \( t \))

Example 1: Dementia Study

Zwijnen et al. BMC Health Services Research 2011, 14:41
http://www.biomedcentral.com/1472-6963/14/41

Grip on challenging behaviour: a multidisciplinary care programme for managing behavioural problems in nursing home residents with dementia. Study protocol

Sandra A Zwijnen1, Martin Smallbrugge1, Sytze Li Zuidema2, Raymond TCM Kompans3, Judith E Basman4, Maurits W van Tulder3, Jan A Eefsting1, Debby L Gemtsen3, Anne-Margriet Pott1AS

Abstract

Background: Behavioural problems are common in nursing home residents with dementia and they often are burdensome for both residents and nursing staff. In this study, the effectiveness and cost-effectiveness of a new care programme for managing behavioural problems will be evaluated.

Methods/Design: The care programme is based on Dutch national guidelines. It will consist of four steps: detection, analysis, treatment and evaluation. A stepped wedge design will be used. A total of 14 dementia special care units will implement the care programme. The primary outcome is behavioural problems. Secondary outcomes will include quality of life, prescription rate of antipsychotics, use of physical restraints and workload and job satisfaction of nursing staff. The effect of the care programme will be estimated using multilevel linear regression analysis. An economic evaluation from a societal perspective will also be carried out.
Example 1: Dementia Study

- Primary outcome: Behaviour problems measured with the Cohen-Mansfield Agitation Inventory (continuous)
- Design:

Table 1: Flow chart of the stepped wedge design

<table>
<thead>
<tr>
<th></th>
<th>T0</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Group 2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Group 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Group 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Group 5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

The groups will start with the care programme on six different points in time (T0 through T5).

Example 1: Dementia Study

- Assumptions:
  - 80% power, $\alpha = 0.05$
  - Standard deviation $\sigma = 16$
  - Minimally Important Effect Size = 4 points (0.25 SD difference)
  - ICC = 0.1
  - $m=20$ residents per home per measurement period
  - $t = 5$ steps ($T = 6$ measurement periods, each of 4 months duration for a total study duration of 24 months)

- NOTE: In this example we will ignore, for convenience, the fact that this is an open cohort design by assuming that different individuals (or a small fraction of a large cohort) are sampled in each home at each time
Example 1: Dementia Study

- Calculate required sample size for individual RCT: \( N_{ind} = 504 \)
- Calculate SW design effect:
  \[
  Deff = \frac{1 + \rho \left( \frac{mt + m - 1}{2} \right)}{1 + \rho \left( \frac{mt}{2} + m - 1 \right)} \times \frac{3(1 - \rho)}{2(1 - \frac{t}{T})} = 0.4593
  \]
- Calculate total required sample size:
  \[
  km(t + 1) = N_{ind} \times (t + 1) Deff = 504 \times 6 \times 0.4593 = 1389
  \]
- Divide to obtain required number of clusters:
  \[
  k = \frac{1389}{20(6)} = 11.6 \approx 12
  \]
- \( \Rightarrow \) A total of 12 nursing homes is required

\[
29
\]

Example 1: Dementia Study

- Power for the planned design with 14 homes, 20 residents per home, is 87.0%:

\[
\begin{align*}
Var(\hat{\theta}) &= 6(T - 1)(1 - \rho)\sigma^2 \left[ 1 + (mT - 1)\rho \right] \\
mkT(T - 2) \left( 1 + \left[ \frac{m(T + 1)}{2} - 1 \right] \rho \right) &= 1.6796 \\
\text{power} &= \Phi \left( \frac{\theta_A}{\sqrt{Var(\hat{\theta})}} - z_{1-\alpha} \right) = \Phi \left( \frac{4}{\sqrt{1.6796}} - 1.96 \right) = 0.870
\end{align*}
\]

\[
30
\]
Example 2: Active Villages Study

Primary outcome: proportion of adults reporting sufficient physical activity to meet internationally recognized guidelines

Design:

- The Devon Active Villages Evaluation (DAVE) trial of a community-level physical activity intervention in rural south-west England: a stepped wedge cluster randomised controlled trial

Abstract

Background: The majority of adults are not meeting the guidelines for physical activity despite activity being linked with numerous improvements to long-term health. In light of this, researchers have called for more community-level interventions. The main objective of the present study was to evaluate whether a community-level physical activity intervention increased the activity levels of rural communities.

Methods: 128 rural villages (clusters) were randomised to receive the intervention in one of four time periods between April 2011 and December 2012. The Devon Active Villages intervention provided villages with 12 weeks of physical activity opportunities for all age groups, including at least three different types of activities per village. Each village received an individually tailored intervention, incorporating a local needs-led approach. Support was provided for a further 12 months following the intervention. The evaluation study used a stepped wedge cluster randomised controlled trial design. All 128 villages were measured at each of five data collection periods using a postal survey.

The primary outcome of interest was the proportion of adults reporting sufficient physical activity to meet internationally recognised guidelines. Minutes spent in moderate- and vigorous activity per week was analysed as a secondary outcome. To compare between intervention and control modes, random effects linear regression and marginal logistic regression models were implemented for continuous and binary outcomes respectively.
Example 2: Active Villages Study

- Assumptions:
  - 80% power, \( \alpha = 0.05 \)
  - Control arm proportion = 25%
  - Minimally Important Effect Size = 5% (absolute increase)
  - ICC = 0.02
  - \( m = 10 \) residents per village per measurement period (to account for anticipated 20% response rate, 50 residents were actually surveyed)
  - \( t = 4 \) steps (\( T = 5 \) measurement periods, each of duration 3 months)

- Note: In this example we will ignore, for convenience, the imbalanced allocation of villages to steps

---

Example 2: Active Villages Study

- Required sample size for individual RCT: \( N_{\text{ind}} = 2496 \)
- SW design effect:

  \[
  \text{Deff} = \frac{1 + \rho \left( \frac{tm}{2} + m - 1 \right)}{1 + \rho \left( \frac{tm}{2} + m - 1 \right) \times \frac{3(1 - \rho)}{2 \left( \frac{t - 1}{t} \right)}} = 0.4912
  \]

- Total required sample size:

  \[
  km \left( t + 1 \right) = N_{\text{ind}} \times (t + 1) \text{Deff} = 2496 \times 5 \times 0.4912 = 6130
  \]

- Required number of clusters:

  \[
  k = \frac{6130}{10(5)} = 122.6 \approx 123
  \]

  \( \Rightarrow \) A total of 123 villages is required
Example 2: Active Villages Study

- Power for the planned design with 128 villages, 10 residents per village, is 83.8%

\[
Var(\hat{\theta}) = \frac{6(T-1)(1-\rho)(1-\pi)[1+(mT-1)\rho]}{mkT(T-2)} = 0.000288
\]

\[
\text{power} = \Phi\left(\frac{\theta_4}{\sqrt{Var(\hat{\theta})}} - z_{1-\alpha}\right) = \Phi\left(\frac{0.05}{\sqrt{0.000288}} - 1.96\right) = 0.838
\]

Extensions: Multiple measurements

- Multiple baseline measures \((b)\)
- Multiple measurements after each step \((q)\)

<table>
<thead>
<tr>
<th>Time</th>
<th>Wedges 1</th>
<th>Wedges 2</th>
<th>Wedges 3</th>
<th>Wedges 4</th>
<th>Wedges 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Unexposed</td>
<td>Exposed</td>
<td>Unexposed</td>
<td>Unexposed</td>
<td>Unexposed</td>
</tr>
<tr>
<td>2</td>
<td>Unexposed</td>
<td>Unexposed</td>
<td>Exposed</td>
<td>Unexposed</td>
<td>Unexposed</td>
</tr>
<tr>
<td>3</td>
<td>Unexposed</td>
<td>Unexposed</td>
<td>Unexposed</td>
<td>Exposed</td>
<td>Unexposed</td>
</tr>
<tr>
<td>4</td>
<td>Unexposed</td>
<td>Unexposed</td>
<td>Unexposed</td>
<td>Unexposed</td>
<td>Exposed</td>
</tr>
<tr>
<td>5</td>
<td>Unexposed</td>
<td>Unexposed</td>
<td>Unexposed</td>
<td>Unexposed</td>
<td>Unexposed</td>
</tr>
</tbody>
</table>

- Exposed to intervention
- Unexposed to intervention
Sample size accounting for multiple measurements

- **Design effect:**
  \[ Deff = \frac{1 + \rho \left( \frac{tqm + bm - 1}{2} \right)}{1 + \rho \left( \frac{tqm + bm - 1}{2} \right)} \times \frac{3(1 - \rho)}{2q(t-1/r)} \]

- **Total required sample size:**
  \[ km(qt + b) = N_{ind} \times (qt + b) Deff \]

  where \( N_{ind} = \) TOTAL number of subjects required for individual randomization

Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J Clin Epidemiol*, 66:752-758

More general designs

- Define \( X_{ij} \) for any design, where \( X_{ij} = 1 \) if cluster \( i \) is in the intervention condition at time \( j \) and 0 otherwise

- Can accommodate general designs, including parallel designs with pre- and post-measurements
Power formula for general designs

- Power to detect an intervention effect $\theta_A$ is:

$$power = \Phi \left( \frac{\theta_A}{\sqrt{\text{Var}(\hat{\theta})}} - z_{1-\alpha} \right)$$

where

$$\text{Var}(\hat{\theta}) = \frac{k(1-\rho)^2 \left[ \frac{(1-\rho)}{m} + T \rho \right]}{(kU - W)^\frac{(1-\rho)}{m} + \left[ U^2 + kTU - TW - kV \right] \rho}$$

$$V = \sum_{i=1}^{k} \left( \sum_{j=1}^{r} X_{ij} \right)^2 \quad W = \sum_{j=1}^{r} \left( \sum_{i=1}^{k} X_{ij} \right)^2 \quad U = \sum_{i=1}^{k} \sum_{j=1}^{r} X_{ij}$$


Extensions: Incomplete designs and multiple levels of clustering

- Hemming e.a. (2015) extended framework to accommodate
  - Incomplete designs
  - Multiple levels of clustering (e.g., patients nested within providers nested within medical practices)
- We will not cover these extensions here

<table>
<thead>
<tr>
<th>Time</th>
<th>Wedge</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>.</td>
<td>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>0</td>
<td>1</td>
<td>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>.</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Incomplete SW design with 1 before and 2 after measurements

<table>
<thead>
<tr>
<th>Time</th>
<th>Wedge</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>.</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Incomplete stepped wedge with implementation period

Software and Resources

- Menu-based facility in STATA (Hemming & Girling, 2014)
  - Power and detectable differences
  - General designs
  - Multiple levels of clustering
  - Outcomes: Dichotomous, continuous, rates
- R-package to be released — includes analytical and simulation-based approaches
- Special issue on stepped wedge trials soon to be published in Trials edited by James Hargreaves & Audrey Prost


Some general considerations

- Need to specify \( t \) with due consideration of practical and logistical constraints
- For a given \( k \), maximizing \( t \) (and thus, the number of measurement times) maximizes power
  - Optimal scenario is 1 cluster per step
  - But diminishing returns to increasing \( t \)
- SW design may be less sensitive to the ICC
  - Always good to do a sensitivity analysis for a range of values
- For cohort designs, calculations based on cross-sectional design are likely to overestimate the required sample size
3. Comparing parallel and stepped wedge CRT designs

Some debate in literature...

- Kotz e.a. (2012): Use of the SW design cannot be recommended: a critical appraisal and comparison with the classic cluster randomized controlled trial design. *J Clin Epi* 65(12): 1249-1252
- Woertman e.a. (2013): Stepped wedge designs could reduce the required sample size in CRTs. *J Clin Epi* 66: 752-758
- Hemming & Girling (2013): The efficiency of SW vs. CRTs: SW studies do not always require a smaller sample size. *J Clin Epi* 66:1427-1429
- De Hoop e.a. (2013): The stepped wedge CRT always requires fewer clusters but now always fewer measurements, that is, participants than a parallel CRT in a cross-sectional design. *J Clin Epi* 66: 1428
- Hemming e.a. (2013): SW CRTs are efficient and provide a method of evaluation without which some interventions would not be evaluated. *J Clin Epi* 66:1058-1060
- Hemming & Taljaard. Setting straight the sample size determination for stepped wedge and cluster randomised trials: design effects and illustrative examples. Under review.
Confusing the issues

- Conclusions differ depending on which parameters are fixed across designs
  - Assumed total cluster sizes (the same or different?)
  - Assumed study duration (the same or different?)
  - Assumed number of measurements (the same or different?)
  - Assumed number of clusters (the same or different?)

Practical considerations in deciding among alternative designs

- Rationale for considering SW design?
- Is the number of clusters limited by availability?
- What is the cost of recruiting additional clusters versus additional subjects per cluster?
- What is planned study duration?
- Impact on power of increasing cluster sizes vs. increasing number of clusters
  - Little benefit to increasing cluster sizes beyond $1/p$
Illustrative examples

- In some situations, there may be 4 practical choices:
  - “Quick” parallel CRT
  - Parallel CRT with extended recruitment (“The Fat CRT”?)
  - Parallel CRT with repeated measures
  - Balanced complete SW design

“The Quick CRT”

<table>
<thead>
<tr>
<th>Parallel CRT</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stepped Wedge CRT</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

- Cluster exposed to intervention
- Cluster unexposed to intervention
“The Fat CRT”

Parallel CRT

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

Stepped Wedge CRT

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>6</td>
</tr>
</tbody>
</table>

Cluster exposed to intervention
Cluster unexposed to intervention

“The multiple measures CRT”

Parallel CRT

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

Stepped Wedge CRT

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>6</td>
</tr>
</tbody>
</table>

Cluster exposed to intervention
Cluster unexposed to intervention
Example 1: Dementia Study

Recall:
- 80% power, \( \alpha = 0.05 \)
- Standard deviation \( \sigma = 16 \)
- Minimally Important Effect Size = 4 points (0.25 SD difference)
- ICC = 0.1
- \( m = 20 \) residents per home per measurement period
- \( T = 6 \) measurement times, each of 4 months duration for a total study duration of 24 months

Note: \( 1/\rho = 10 \) i.e., little benefit to increasing cluster sizes much further

Example 1: Dementia Study

Comparison of 4 design choices assuming ICC=0.1

<table>
<thead>
<tr>
<th>Design Choices</th>
<th>Required ( k ) (total)</th>
<th>Study duration</th>
<th>Total # observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multicenter trial using individual randomization</td>
<td>25</td>
<td>4 months</td>
<td>504</td>
</tr>
<tr>
<td>Quick CRT</td>
<td>74</td>
<td>4 months</td>
<td>1480</td>
</tr>
<tr>
<td>Fat CRT</td>
<td>56</td>
<td>24 months</td>
<td>6720</td>
</tr>
<tr>
<td>Multiple measures CRT (1 pre, 5 post)</td>
<td>21</td>
<td>24 months</td>
<td>2520</td>
</tr>
<tr>
<td>Balanced complete SW design with ( T = 6 )</td>
<td>12</td>
<td>24 months</td>
<td>1440</td>
</tr>
</tbody>
</table>
Example 1: Dementia Study

- Comparison of 4 design choices assuming smaller ICC=0.001

<table>
<thead>
<tr>
<th>Design Choice</th>
<th>Required k (total)</th>
<th>Study duration</th>
<th>Total # observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multicenter trial using individual randomization</td>
<td>25</td>
<td>4 months</td>
<td>504</td>
</tr>
<tr>
<td>Quick CRT</td>
<td>26</td>
<td>4 months</td>
<td>520</td>
</tr>
<tr>
<td>Fat CRT</td>
<td>6</td>
<td>24 months</td>
<td>720</td>
</tr>
<tr>
<td>Multiple measures CRT (1 pre, 5 post)</td>
<td>6</td>
<td>24 months</td>
<td>720</td>
</tr>
<tr>
<td>Balanced complete SW design with $T=6$</td>
<td>9</td>
<td>24 months</td>
<td>1080</td>
</tr>
</tbody>
</table>

Example 2: Active Villages Study

- Recall:
  - 80% power, $\alpha = 0.05$
  - Control arm proportion = 25%
  - Minimally Important Effect Size = 5% (absolute increase)
  - ICC = 0.02
  - $m=10$ residents per village per measurement period
  - $t = 4$ steps ($T = 5$ measurement periods, each of duration 3 months)
  - Note: $1/\rho = 50$ i.e., some benefit to increasing cluster sizes
Example 2: Active Villages Study

- Comparison of 4 design choices assuming ICC=0.02

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Required k (total)</th>
<th>Study duration</th>
<th>Total # observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multicenter trial using individual randomization</td>
<td>250</td>
<td>3 months</td>
<td>2496</td>
</tr>
<tr>
<td>Quick CRT</td>
<td>296</td>
<td>3 months</td>
<td>2960</td>
</tr>
<tr>
<td>Fat CRT</td>
<td>100</td>
<td>15 months</td>
<td>5000</td>
</tr>
<tr>
<td>Multiple measures CRT (1 pre, 4 post)</td>
<td>97</td>
<td>15 months</td>
<td>4850</td>
</tr>
<tr>
<td>Balanced complete SW design with T=5</td>
<td>123</td>
<td>15 months</td>
<td>6150</td>
</tr>
</tbody>
</table>

Example 2: Active Villages Study

- Comparison of 4 design choices assuming larger ICC =0.1

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Required k (total)</th>
<th>Study duration</th>
<th>Total # observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multicenter trial using individual randomization</td>
<td>250</td>
<td>3 months</td>
<td>2496</td>
</tr>
<tr>
<td>Quick CRT</td>
<td>476</td>
<td>3 months</td>
<td>4760</td>
</tr>
<tr>
<td>Fat CRT</td>
<td>296</td>
<td>15 months</td>
<td>14800</td>
</tr>
<tr>
<td>Multiple measures CRT (1 pre, 4 post)</td>
<td>165</td>
<td>15 months</td>
<td>8250</td>
</tr>
<tr>
<td>Balanced complete SW design with T=5</td>
<td>136</td>
<td>15 months</td>
<td>6800</td>
</tr>
</tbody>
</table>
Summary

- Direct comparisons between the designs is complex – best to evaluate power of alternative designs within logistical constraints of the planned study on a case-by-case basis.
- Quick CRT may not be feasible (required $k$ may exceed that available); but may be preferable if inexpensive to recruit more clusters as it can have shorter duration.
- Parallel designs (fat / repeated measures) are preferable when ICC is small; although repeated measures may offer little benefit.
- SW designs are preferable when ICC is large.

4. Conclusions
General conclusions

- The Stepped Wedge Design Effect — a function of $\rho$, cluster size, and number of steps — can be used to determine sample sizes for cross-sectional SW trials
- If $\rho$ unknown, recommend sensitivity analyses for a range of alternative values — although SW designs less sensitive to $\rho$
- Parallel CRTs designs may not be feasible with limited numbers of clusters — SW designs offer reasonable choice
- SW designs a good choice when $\rho$ is large; parallel CRTs a good choice when $\rho$ is small
- Regardless of sample size formula, small numbers of clusters may not be desirable for many reasons!

Areas for future work

- Methodology is still in active development
  - Sample size methodology for cohort designs
  - Sample size for other types of outcomes
  - Accounting for variation in cluster sizes
  - Accounting for attrition
  - Need to account for inter-period correlation?
  - Minimum recommended number of clusters?
  - Minimum number of observations per cluster per time interval?
Comparative Efficiency of Stepped-Wedge Designs

Alan Girling
University of Birmingham, UK
A.J.Girling@bham.ac.uk

A Statistical Design Question

- How does the Stepped Wedge Design perform as a tool for estimating a treatment effect?
- Ethical and Logistical concerns are (temporarily) suspended
- Question addressed under the basic Hussey & Hughes model
  - Exact results are possible
  - Insight into more general scenarios
Assumptions

- Study of fixed duration (8 months)
- Constant recruitment rate in each cluster
- Continuous Outcome
  - Additive treatment effect
  - Cross-sectional observations with constant ICC = ρ
- Cross-over in one direction only (i.e. Treated to Control prohibited)
- The analysis allows for a secular trend (“time effect”)
  - This has been questioned; but if time effects are ignored, the ‘best’ statistical design involves simple before-and-after studies in each cluster (i.e. not good at all!)

Goal: To compare statistical performance of different designs under these assumptions, especially the Stepped-Wedge

1. Comparing Two Designs
Two Simple Candidate Designs

1. Parallel Study with (multiple) baseline controls
   “Controlled Before-and-After Design” (CBA)

   Clusters
   0 0 0 0 1 1 1 1
   0 0 0 0 1 1 1 1
   0 0 0 0 0 0 0 0
   0 0 0 0 0 0 0 0

   Treatment Effect Estimate =
   (Mean difference between two groups in months 5 – 8)
   minus
   r x (Mean difference between two groups in months 1 – 4)
   (r is a correlation coefficient “derived from the ICC”)

Two groups of clusters. Treatment implemented in one group only, in month 5.

2. Simple Parallel Design (PD)

   Clusters
   1 1 1 1 1 1 1 1
   1 1 1 1 1 1 1 1
   0 0 0 0 0 0 0 0
   0 0 0 0 0 0 0 0

   Treatment Effect Estimate =
   (Mean difference between two groups over all months)
Performance measured by Precision of the effect estimate:

\[ \text{Precision} = \frac{1}{\text{(Sampling Variance)}} = \frac{1}{\text{(Standard Error)}}^2 \]

Where \( R \) is the “Cluster-Mean Correlation”

\[ R = \frac{m \rho}{1 + (m-1) \rho} \]

... and is the same for both designs

\[(m = \text{number of observations per cluster, } \rho = \text{ICC})\]

Relative Efficiency of designs – by comparing straight lines on a Precision-Factor plot

CBA design is better (“more efficient”) if \( R > 2/3 \). Otherwise the Parallel design is better.
2. The Cluster-Mean Correlation (CMC)

- Relative efficiency of different designs depends on cluster-size ($m$) and ICC ($\rho$), but only through the CMC ($R$)
- The CMC ($R$) = proportion of the variance of the average observation in a cluster that is attributable to differences between clusters
  \[ R = \frac{m\rho}{1+(m-1)\rho} \approx \frac{m\rho}{1+m\rho} \]
- (The ICC ($\rho$) = proportion of variance of a single observation attributable to differences between clusters)
- CMC can be large (close to 1) even if ICC is small
• CMC can be large (close to 1) even if ICC is small
• So CBA can be more efficient than a parallel design for reasonable values of the ICC, if the clusters are large enough

\[ R = \frac{m \cdot \rho}{1 - \rho} \]

3. Some More Designs with 4 Groups of Clusters
### Controlled Before & After

<table>
<thead>
<tr>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Before & After + Parallel

<table>
<thead>
<tr>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Precision Factors**:

\[
\frac{1}{2} - \frac{1}{4}R < \frac{3}{4} - \frac{1}{2}R < \frac{9}{16} - \frac{9}{32}R
\]

### + Baseline & Full Implementation

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th></th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Stepped Wedge

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### How to find Precision Factors: \( A - B \times R \)

<table>
<thead>
<tr>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>½</td>
<td>½</td>
<td>½</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>½</td>
<td>½</td>
<td>½</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>½</td>
<td>½</td>
<td>½</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>½</td>
<td>½</td>
<td>½</td>
</tr>
</tbody>
</table>

**Row means**

\[\frac{1}{2} \times \frac{1}{4} \times R\]

\[\frac{3}{4} \times \frac{1}{2} \times R\]

\[\frac{9}{16} \times \frac{9}{32} \times R\]

**Column variances**

\[\frac{1}{4}\]

\[\frac{1}{16}\]

\[\frac{1}{8}\]

\[1/2\]

\[1/4\]

\[1/16\]

\[1/8\]

\[B = 4 \times 1/16 = 1/4\]

\[A = 4 \times \text{Average of these} = 4 \times 1/8 = 1/2\]

A is \(4 \times \) the “variance within-columns” of (0/1) treatment indicator

B is \(4 \times \) the “variance between-rows” of (0/1) treatment indicator
- Stepped-Wedge is best when the CMC is close to 1
- Parallel Design is best when the CMC is close to 0 – but risky if ICC is uncertain
- Before & After/Parallel mixture is a possible compromise
4. Is there a ‘Best’ Design?

- If Cross-over from Treatment to Control is permitted, the Bi-directional Cross-Over (BCO) design has Precision Factor = 1 at every R, better than any other design.

\[
\text{Precision Factor} = 1 - 0 \times R
\]

- Usually Treatment cannot be withdrawn within the study duration. Then the Precision Factor cannot exceed

\[
1 - R + \frac{1}{3} R^2
\]

t for any feasible design.
• 'Best' designs are those whose precision lines are tangent to the boundary curve
• Parallel Design is tangent at R = 0
• Stepped-Wedge design close to the boundary at R = 1
• 3rd design can be improved

Some (nearly) ‘Best’ designs with 4 groups of clusters

• Available design performance limited by choosing only 4 groups of clusters
Some (nearly) ‘Best’ designs with 8 groups of clusters

‘Best’ Design for large studies: This is a mixture of Parallel and Stepped-Wedge clusters.

- This Design achieves tangency to the boundary curve at R
- Includes Parallel and Stepped-Wedge designs as special cases
- When R is close to 1 the Stepped-Wedge is the best possible design
5. Conclusions

• Efficiency of different cluster designs depends on the ICC (\( \rho \)) and the Cluster-size (m) but only through the CMC (R).

• The Stepped-Wedge Design is most advantageous when R is close to 1.
  – In studies with large clusters this can arise even if the ICC is relatively small.

• Precision Factor plots for comparing designs
  – Comparison of designs is a linear comparison in R
  – The theoretically ‘best’ design choice is sensitive to R, and combines Parallel with Stepped-Wedge clusters

Limitations

• Continuous data, simple mixed model
  – Natural starting point – exact results are possible
  – Applies to Binary observations through large-sample approximations

• Cross-sectional designs only
  – Extension to cohort designs through nested subject effects is (relatively) straightforward
  – More complex (realistic) time-dependency
Three treatment arms
Stepped wedge like and other designs

Steven Teerenstra
biostatistics, Radboud Instituted for Health Sciences

joint work with
Hilly Calsbeek, Hub Wollersheim from IQ Healthcare

Why comparing three treatments?
- Two interventions in context to placebo
  - Implementation strategy T, enhanced T+
  - What if no active strategy was applied: C
  - T+ superior to T+
    • (descriptive) comparison T vs C, T+ vs C
- Golden standard non-inferiority trial
  - placebo P, active control AC, test T
  - T non-inferior to AC,
  - AC (and/or T) superior to P (assay sens.)
- Equal interest in 3 interventions
  - A vs B, A vs C, B vs C
Setting

- **Cross-sectional**
  - equal cluster size, equal number of clusters per group

- **constant effects over time;** $T_+ > T > > C$
  - two intervention effects: $\theta_{T_+, T} > 0$, $\theta_{T, C} > 0$
  - power on $T_+$ vs $T$
  - descriptive $T$ vs $C$, $T_+$ vs $C$

- **Model**
  - $Y_{ijk} = \mu + \beta_i + \theta_{T_+, T} + \theta_{T, C} + u_j + e_{ijk}$ if $T_+$
  - $Y_{ijk} = \mu + \beta_i + \theta_{T, C} + u_j + e_{ijk}$ if $T$
  - $Y_{ijk} = \mu + \beta_i + u_j + e_{ijk}$ if $C$

What designs make more / less sense?

*building blocks*
Some possible designs built from these
Different goals / constraints

- Small number of clusters
  - e.g. limited number of memory clinics

- Small number of measurements
  - e.g. trial has to finish in 2 years due to grant obligation

- Small number of total subjects
  - #clusters * cluster size * #measurements
    (because design is cross-sectional)

- Small costs
  - Optimization of power given cost function
  - Given cost per cluster, per subject, per measurement
Different goals / constraints

- Small number of clusters
  - e.g. limited number of memory clinics

- Small number of measurements
  - e.g. trial has to finish in 2 years due to grant obligation

- Small number of total subjects
  - #clusters * cluster size * #measurements
    (because design is cross-sectional)

- Small costs
  - Optimization of power given cost function
  - Given cost per cluster, per subject, per measurement
Observations

- SW designs are less influenced by ICC than (pre)post test designs

- Impact of ICC increases with cluster size
• Smallest number of clusters
  • double SW (parallel or sequential)
    – If n=5: parallel double SW fewest clusters
    – If n=50: sequential double SW fewest clusters
  • seq. pre-post-post, seq.double pre-post, seq.double hybrid SW

• Small number of measurements
  – 2 measurements
    • Small ICC: hybrid pre-posttest
    • Larger ICC: parallel hybrid double preposttest
  – 3 measurements
    • Parallel double SW

• ‘SW’ favorably for comparing 3 trt arms

Sample size
1. calculate sample size per group as usual
   • for a individual randomized posttest design
2. multiply by (1+(n-1)*ICC)
   • to account for clustering of subjects in clusters
   • n=cluster size
3. multiply by appropriate design effect
   • gives the sample size per group
4. multiply by #groups for sample size per time
   • So divide by cluster size to get total number of clusters
Total number of clusters ..

<table>
<thead>
<tr>
<th>Effect = 0.2 standard deviation</th>
<th>design</th>
<th>posttest</th>
<th>sequential</th>
<th>sequential double</th>
<th>sequential double stepped wedge</th>
<th>parallel double stepped wedge</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>5</td>
<td>249</td>
<td>168</td>
<td>166</td>
<td>166</td>
<td>166</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>63</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>39</td>
<td>24</td>
<td>22</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>0.05</td>
<td>5</td>
<td>288</td>
<td>164</td>
<td>178</td>
<td>178</td>
<td>178</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>108</td>
<td>50</td>
<td>44</td>
<td>44</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>87</td>
<td>38</td>
<td>24</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>0.1</td>
<td>5</td>
<td>333</td>
<td>194</td>
<td>182</td>
<td>182</td>
<td>174</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>165</td>
<td>52</td>
<td>44</td>
<td>44</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>144</td>
<td>30</td>
<td>24</td>
<td>24</td>
<td>22</td>
</tr>
</tbody>
</table>

Discussion

- Not exhaustive list of options…other?
- Practicality of these designs?
  - Few measurements:
    - more implementation effort? (more clusters per group)
    - Learning effect / fading out effect cannot be investigated?
    - Effect has to come to full potential before measurement
- Order of ‘best’ designs may be different when minimizing cost / total sample size
- …
Some illustrative examples on the analysis of the SW-CRT

Karla Hemming

Calendar time is a confounder
Why calendar time is a confounder...

Time effects

- When designing a SW-CRT time needs to be allowed for in the sample size calculation.

- Time also needs to be allowed for in the analysis
Analysis

• Summarise key characteristics by exposure / unexposed status
  – Identify selection biases

• Analysis either GEE or mixed models
  – Clustering
  – Time effects

• Imbalance of calendar time between exposed / unexposed:
  – The majority of the control observations will be before the
    majority of the intervention observations
  – Time is a confounder!

• Unadjusted effect meaningless
Treatment effect

• After accounting for any secular changes, what is the effect of the intervention, averaged across steps?

• The intervention effect is modelled as a change in level, constant across steps (or time)

Interpretation of intervention effect in SW design: continuous time

Estimated after averaging across all sites / time points in each condition
Example one
Example 1: Maternity sweeping

- Objective: evaluate a training scheme to improve the rate of membrane sweeping in post term pregnancies
  - Primary outcome:
    - Proportion of women having a membrane sweep
    - Baseline rate 40%
    - Hope to increase to 50% during 12 weeks post intervention
  - Cluster design:
    - 10 teams (clusters); 12 births per cluster per week
    - Pragmatic design – rolled out when possible
    - Transition period to allow training

Example 1: Maternity sweeping (transition period)
Example 1: Underlying trend

Example 1: results

<table>
<thead>
<tr>
<th></th>
<th>Unexposed to intervention n=1417</th>
<th>Exposed to intervention n=1356</th>
<th>Relative Risk</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of women offered and accepting membrane sweeping</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number (%)</td>
<td>629 (44.4%)</td>
<td>634 (46.8%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster adjusted</td>
<td></td>
<td>1.06 (0.97, 1.16)</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>Time and cluster adjusted</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed effects time</td>
<td>0.88 (0.69, 1.05)</td>
<td>0.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear time effect</td>
<td>0.90 (0.73, 1.11)</td>
<td>0.34</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example 1: Impact of secular trend

Unadjusted RR = 1.06 95% CI (0.97, 1.16)

Adjusted (for time) RR = 0.88 95% CI (0.69, 1.05)

Explanations

- Rising tide
  - General move towards improving care – perhaps due to very initiative that prompted study investigators to do this study
- Contamination
  - Unexposed clusters became exposed before their randomisation date
- Lack of precision
  - Intervention wasn’t ruled out as being effective
Example 2: Critical care outreach

- Intervention: Critical care outreach
- Setting: Hospital in Iran
- Clusters: Wards
- Outcome: Mortality
Example 2: Design

Example 2: Underlying trend
Example 2: results

<table>
<thead>
<tr>
<th></th>
<th>Unexposed to intervention</th>
<th>Exposed to intervention</th>
<th>Treatment effect</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Patients</td>
<td>7,802</td>
<td>10,880</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mortality</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number (%)</td>
<td>370 (4.74)</td>
<td>384 (3.53)</td>
<td>OR (95% CI)</td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.73 (0.64, 0.85)</td>
<td></td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Cluster adjusted</td>
<td>1.02 (0.83, 1.26)</td>
<td></td>
<td>0.817</td>
<td></td>
</tr>
<tr>
<td>Fixed effects for time</td>
<td>0.82 (0.56, 1.19)</td>
<td></td>
<td>0.297</td>
<td></td>
</tr>
<tr>
<td>Linear effect for time</td>
<td>0.96 (0.76, 1.22)</td>
<td></td>
<td>0.750</td>
<td></td>
</tr>
<tr>
<td>Covariate adjusted</td>
<td>0.97 (0.64, 1.47)</td>
<td></td>
<td>0.489</td>
<td></td>
</tr>
</tbody>
</table>

Example 2: Underlying trend

Unadjusted OR = 1.02 95% CI (0.83, 1.26)
Adjusted OR = 0.82 95% CI (0.56, 1.19)

Contamination?
Lack of power?
Explanations

- Rising tide
  - General move towards improving care – perhaps due to very initiative that prompted study investigators to do this study

- Contamination
  - Unexposed clusters became exposed before their randomisation date

- Lack of precision
  - Intervention wasn’t ruled out as being effective

Models fitted

- Mixed effects model
- Random effect for cluster
- Fixed effect for time and intervention

\[ Y_{ijl} = \mu + \beta_j + \theta X_{ij} + u_i + e_{ijl} \]

\[ u_i \sim N(0, \sigma_u); \; e_{ijl} \sim N(0, \sigma_w) \]

(time \( j \), cluster \( i \), individual \( l \))
Parameter estimation

- Models fitted used Stata 13
  - Used meglm function
  - Uses mean-variance adaptive Gauss-Hermite quadrature
  - Default number of integration points (7)
  - Experienced convergence difficulty (LOS example) then used the Laplace approximation – but clear instability

- Used random effects
  - GEE alternative
  - GEE possibly more robust to model miss-specification
  - GEE possibly problematic when small number of clusters (there exist adjustments)

Model assumptions and extensions
Model assumptions 1

- Underlying secular trend
  - The underlying secular trend is same across all clusters

Variation in underlying secular trends

\[ y_{i,k} = p + \beta S + \delta X_{i} + u_{i} + e_{i,k} \]

S is strata
Model assumptions 2

- Time invariant treatment effect
  - There is no delayed intervention effect
  - No change in intensity of the effect over the course of time
  - No time by treatment interaction
  - Time (since introduction) isn't an effect modifier

Time variant treatment effect

\[ Y_{ij} = \mu + \beta_i + (i \times X_{ij}) + \eta_i + \epsilon_{ij} \]
Model assumptions 3

- Intra cluster correlation
  - The correlation between two individuals is independent of time
  - Two observations in the same cluster / time period have the same degree of correlation as two observations in the same cluster but different time periods

\[ Y_{ijt} = \mu + \beta_j + \theta_i X_{ij} + u_i + v_{ij} + e_{ijt} \]

\[ u_i \sim N(0, \sigma_u), \quad v_{ij} \sim N(0, \sigma_v), \quad e_{ijt} \sim N(0, \sigma_e) \]

Model assumptions 4

- Treatment effect heterogeneity
  - The effect of the intervention is the same across all clusters
  - Typical assumption in CRTs
  - In a meta-analysis common to assume between cluster heterogeneity in treatment effect

\[ Y_{ijt} = \mu + \beta_j + \theta_i X_{ij} + u_i + e_{ijt} \]

\[ \theta_i \sim N(0, \sigma_\theta) \]
Summary

- Time is a potential partial confounder
- Designs which are completely confounded with time shouldn't be used
- Time must be allowed for as a covariate in primary analysis
- Model extensions require sufficient power and pre-specification

References


Reporting Guidelines

CONSORT for Stepped Wedge Trials

- No tailored reporting guidelines at present
- Suggest following CONSORT 2010 extension to CRTs
- CONSORT extension to stepped wedge trials currently in development
- For preliminary recommendations, see:
## Preliminary reporting recommendations

### Table 2: Suggested modifications to the Consort 2010 cluster extension for reporting of stepped wedge cluster randomised trials

<table>
<thead>
<tr>
<th>Section and topic</th>
<th>Modified checklist item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title and abstract</td>
<td>Identification as a stepped wedge cluster randomised trial</td>
</tr>
<tr>
<td>Introduction</td>
<td>Rationale: Stakeholders not amenable to parallel randomisation; Need for sequential rollout; Desire for all clusters to receive the intervention; Evidence of preliminary effectiveness; Likely to be an efficient design for anticipated intra-cluster correlation and cluster size.</td>
</tr>
<tr>
<td>Methods</td>
<td>Trial design: Definition of the cluster; Cluster size distinguished from cluster size per observation period; Length of the steps (observation periods); Number of clusters randomised at each step; Cohort (repeated measures on individuals), cross-sectional design (different individuals), or mixture (open cohort); Schematic representation of the trial design.</td>
</tr>
<tr>
<td>Sample size justification</td>
<td>Allowance for clustering; Allowance for the number of steps; Allowance for any repeated measures on individuals; Clear reference to the methods used.</td>
</tr>
<tr>
<td>Analysis</td>
<td>Allowance for clustering (that is, random effect models); Allowance for the number of steps (that is, fixed effect for step); Allowance for repeated measures on individuals, if appropriate.</td>
</tr>
<tr>
<td>Results</td>
<td>Characteristics of sample reported by exposed and unexposed observation periods, or by randomisation group; Adjusted (for times) treatment effect and 95% CI should be interpreted as unbiased estimate of effect size; Schematic representation of actual trial study design; Inference to these analyses should follow the randomised design and might be different to that what is stated here.</td>
</tr>
</tbody>
</table>

### Preliminary reporting recommendations

- **Trial design**
  - Definition of the cluster
  - Cluster size distinguished from cluster size per observation period
  - Length of the steps (observation periods)
  - Number of clusters randomised at each step
  - Cohort (repeated measures on individuals), cross-sectional design (different individuals), or mixture (open cohort)
  - Schematic representation of the trial design
Preliminary reporting recommendations

- Sample size justification
  - Allowance for clustering
  - Allowance for the number of steps
  - Allowance for any repeated measures on individuals
  - Clear reference to the methods used

Preliminary reporting recommendations

• Analysis
  – Allowance for clustering
  – Allowance for time effects
  – Allowance for repeated measurements

• Results
  – Characteristics of study population reported by those exposed and unexposed to the intervention
  – Time adjusted treatment effect should be reported and interpreted as the unbiased estimate of the treatment effect
  – Schematic representation of the study design