Introducing a Stata program for cluster-level analysis of CRTs

Stephen Nash

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Outline

- Why cluster-level analysis?
- An example of unadjusted analysis
- The Hayes & Moulton method
- `clan`: a nascent Stata program
- A live demonstration!
- Current capabilities
- What happens next?
- Thoughts and suggestions

Purpose of this talk

- Describe cluster-level summary analysis
- Introduce new Stata program
- Seek feedback / beta testers

* AKA Anything could happen in the next ten minutes

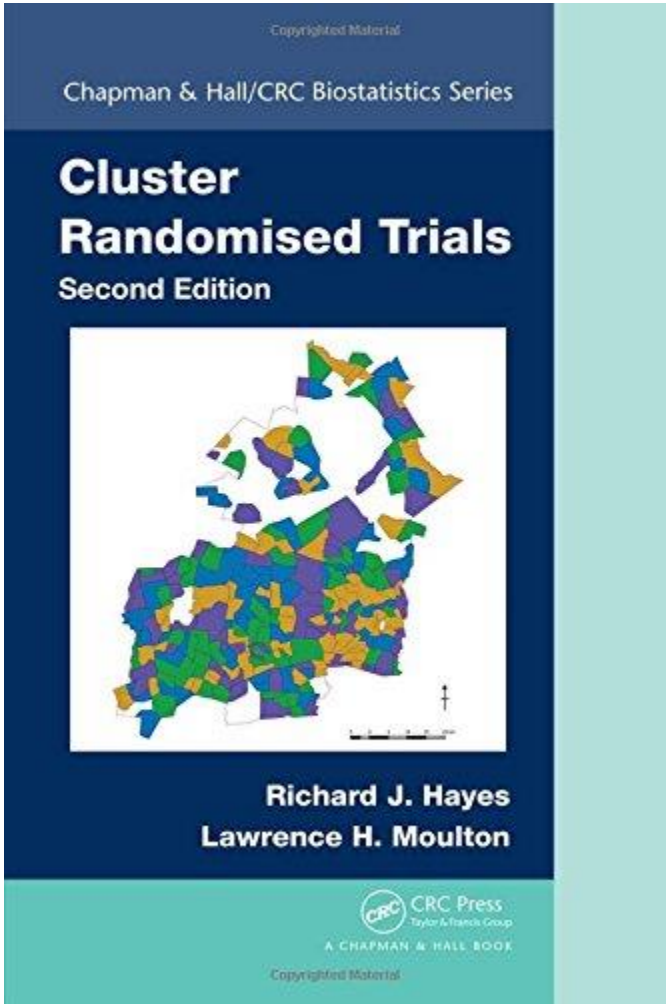
Why cluster-level analysis?

- It is possible to analyse at the individual level
- Adjust for clusters with random effects or GEE
- But may not be robust for 'small' number of clusters
- Cluster-level analysis is
  - intuitive[1]
  - simple[2]
  - robust to small number of clusters[3]

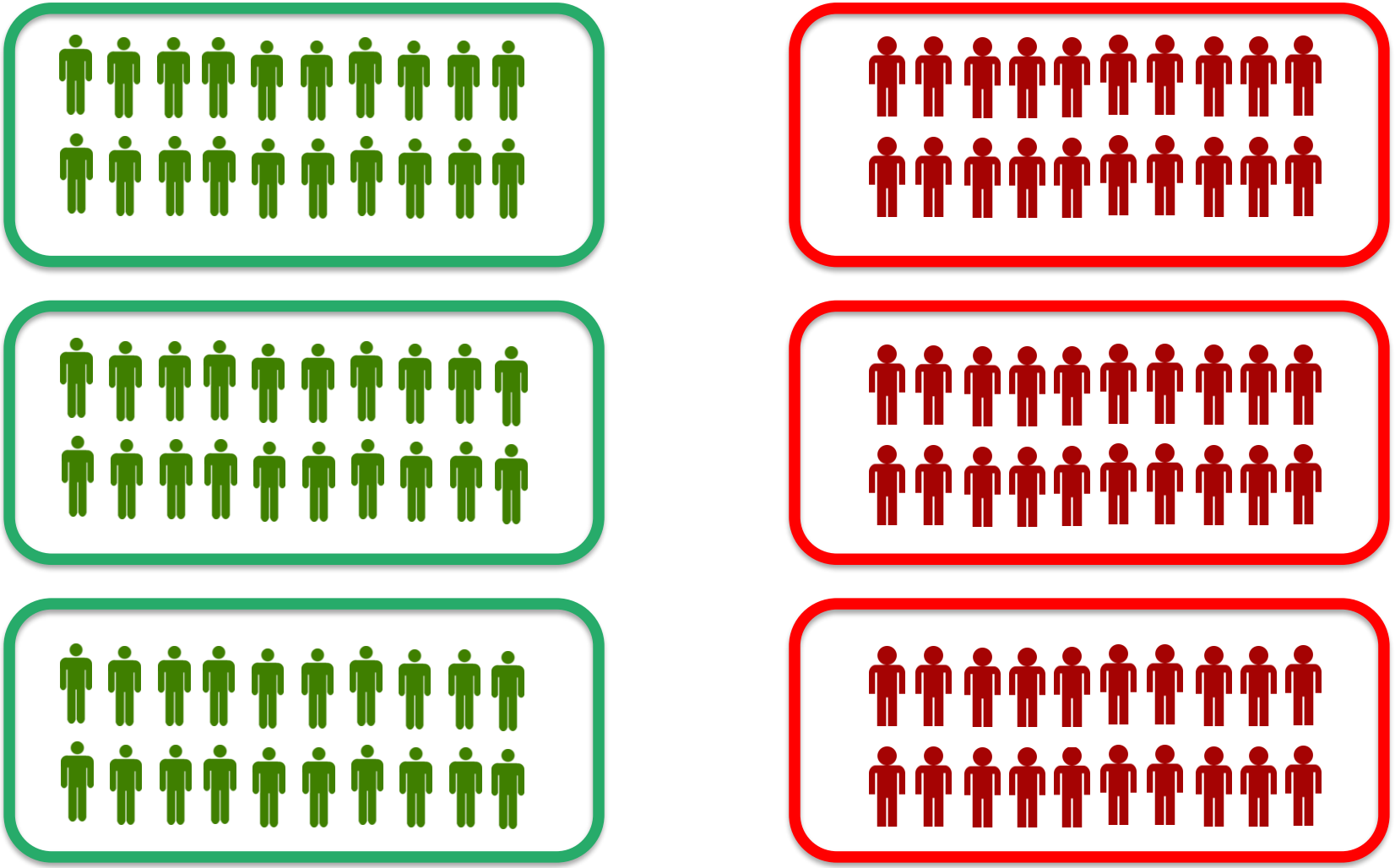
1. I claim, and hope to convince you in the next five minutes
2. It will be when we publish our program
3. Methods for the analysis of incidence rates in cluster randomized trials; S Bennett et al; International Journal of Epidemiology, Volume 31, Issue 4, August 2002, Pages 839–846, https://doi.org/10.1093/ije/31.4.839

Cluster-level analysis

- Summarise the outcome at the cluster level
- Compare the cluster proportions / means
- Adjust for confounders by first running an individual level analysis, omitting the treatment variable and ignoring clustering
- Then summarise the residuals (instead of the outcome)
- But this can be tricky to code, and involves data manipulation (which can be scary)



Unadjusted cluster-level analysis



Unadjusted cluster-level analysis

Unadjusted cluster-level analysis

p = 3/20 = 0.15

p = 10/20 = 0.50

p = 5/20 = 0.25

p = 5/20 = 0.25

p = 4/20 = 0.20

p = 3/20 = 0.15

Unadjusted cluster-level analysis

$\overline{P} = 0.30$

$\overline{P} = 0.20$

Unadjusted cluster-level analysis

$\overline{P} = 0.30$

$\overline{P} = 0.20$

Use a t-test to compare the cluster-level prevalences

(0.15, 0.50, 0.25) vs (0.25, 0.20, 0.15)

There will be 6-2=4 degrees of freedom

The Hayes & Moulton method

j = cluster
k = individual within cluster
m = coefficients in linear predictor (from logistic regression)

π_{jk} = probability of event / outcome
γ_m = Coefficient from logistic regression
z_{jkm} = Covariate value for individual
d_j = Number of events in j^{th} cluster

STAGE 1

- Before summarising at cluster level, run a logistic regression
  - ***exclude*** *the intervention & clustering*
  - *include all other covariates (individual and cluster-level)*

$$\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha + \sum_m \gamma_m z_{jkm}$$

- Use the predicted log-odds for each individual to calculate the predicted probability

Cluster Randomised Trials by Hayes & Moulton, pages 223-224

The Hayes & Moulton method

STAGE 2

- Sum these probabilities to calculate the expected number of events per cluster

$$e_j = \sum_k \hat{\pi}_{jk}$$

- Ratio of observed to expected events for each cluster *j* is

$$R_j = \frac{d_j}{e_j}$$

- Calculate the mean R in each arm
- Adjusted risk ratio is then

$$\beta_{effect} = \frac{\bar{R}_1}{\bar{R}_0}$$

The Hayes & Moulton method

- Use a t-test to compare these R_j between the two arms
- One additional degree of freedom is lost for each cluster-level variable included in the individual level logistic regression (stage 1)
- Hence Stata's t-test command will not be valid

- For a risk difference we calculate the difference between observed and expected outcomes
- Works in a similar way for continuous and count outcomes

clan: syntax

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

```
clan depvar [indepvars] [if] [in] , [options]
```

Compulsory options

- `arm(varname)` must be binary
- `cluster(varname)`
- `effect(string)` one of rr, rd, rater, means

Optional options

- `strata(varname)`
- `fuptime(varname)`
- `plot`
- `level(cilevel)`

`clan`: a nascent Stata program

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

- Five types of effect measure
  - Risk ratio
  - Risk difference
  - Difference in means
  - Rate ratio
  - Rate difference
- Follows standard Stata syntax (as much as possible)
  - eg allows i.varname for categorical variables

`clan`: a nascent Stata program

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

- Covariate adjustment
- Automatically adjusts degrees of freedom
- Can cope with zero prevalence clusters
- Plot of cluster-level summaries
- Stores key results in a return list
- Doesn't change the data in memory

A live demonstration

- <Switch to Stata>

What happens next

- Finish the program (& write a help file)
- The code is available for beta testing
  - Email **stephen.nash@lshtm.ac.uk**
- Improve the program (effect modification, cluster weights, stratification)
- Make available on Stata

```
ssc install clan
```

- Stata journal paper

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Thoughts and suggestions?

stephen.nash@lshtm.ac.uk