Cluster randomised trial analysis made easy: the clan Stata command

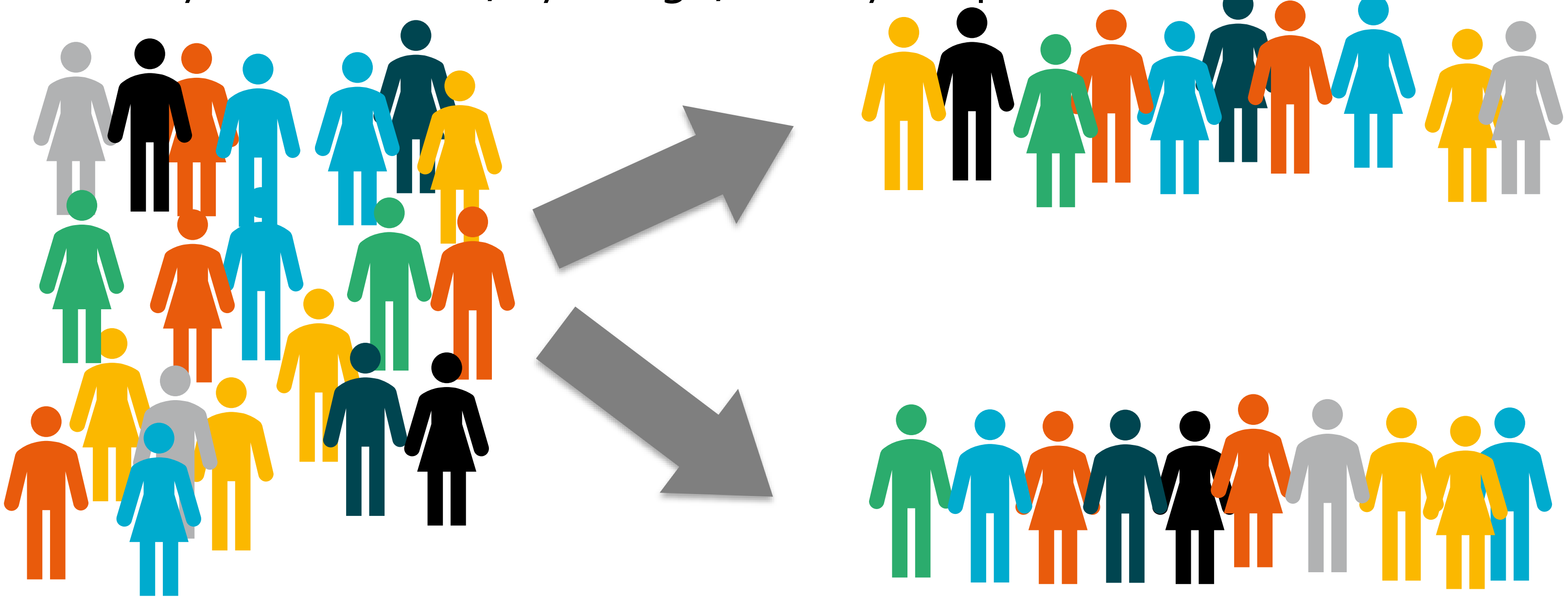
Jennifer Thompson

LONDON
SCHOOL of
HYGIENE
&TROPICAL
MEDICINE

International
Statistics &
Epidemiology
Group

What is a trial?

The analysis of a trial is, by design, usually simple.

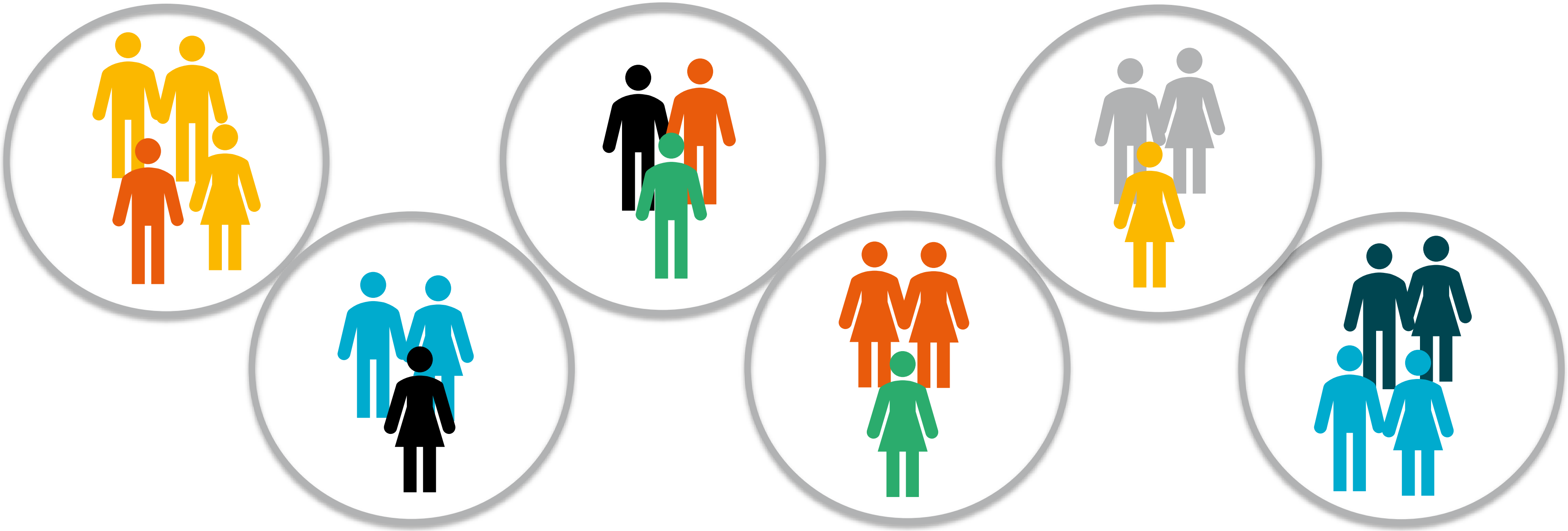


We want to compare conditions, say a standard or care to a new intervention. We randomly assign patients between the conditions.

Analysis can be very simple because the design, if well conducted, ensures all potential confounders are balanced between the arms.

What is a cluster randomised trial?

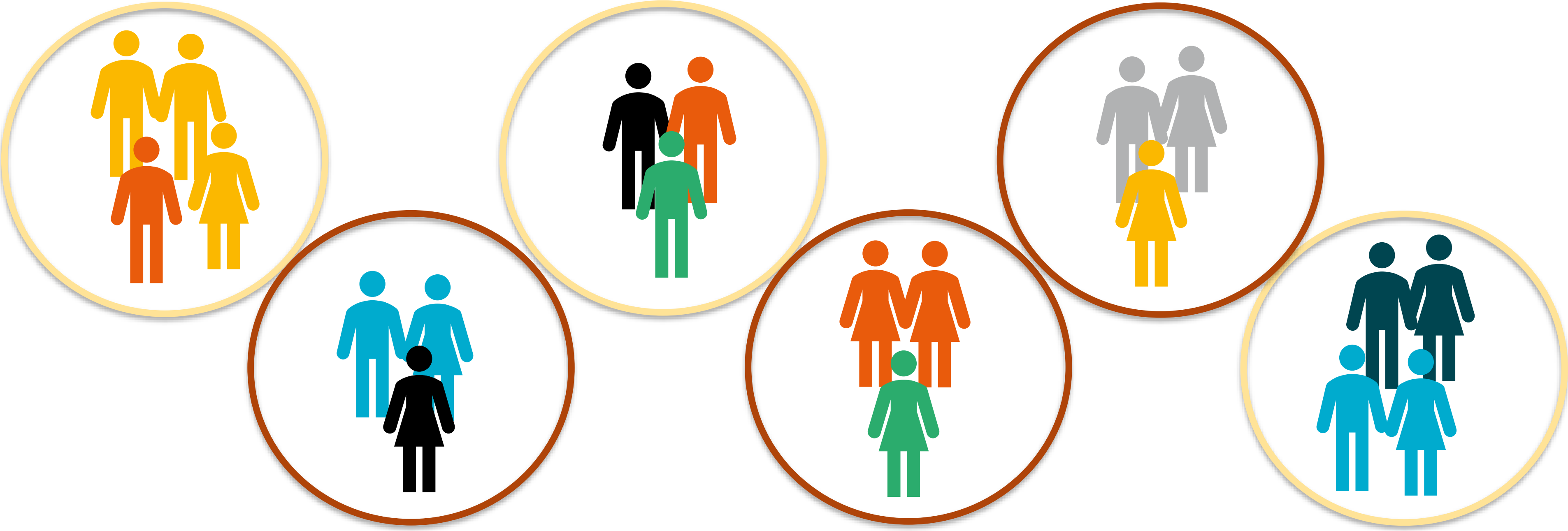
Randomise groups of individuals (clusters) like villages or hospital wards. Individuals in the same clusters are likely to be more similar to one another than individuals in different clusters.



This is clustering or correlation

What is a cluster randomised trial?

Randomise groups of individuals (clusters) like villages or hospital wards. Individuals in the same clusters are likely to be more similar to one another than individuals in different clusters.



This is clustering or correlation

Analysis of a cluster randomised trial

Analysis of these trials must account for this correlation.

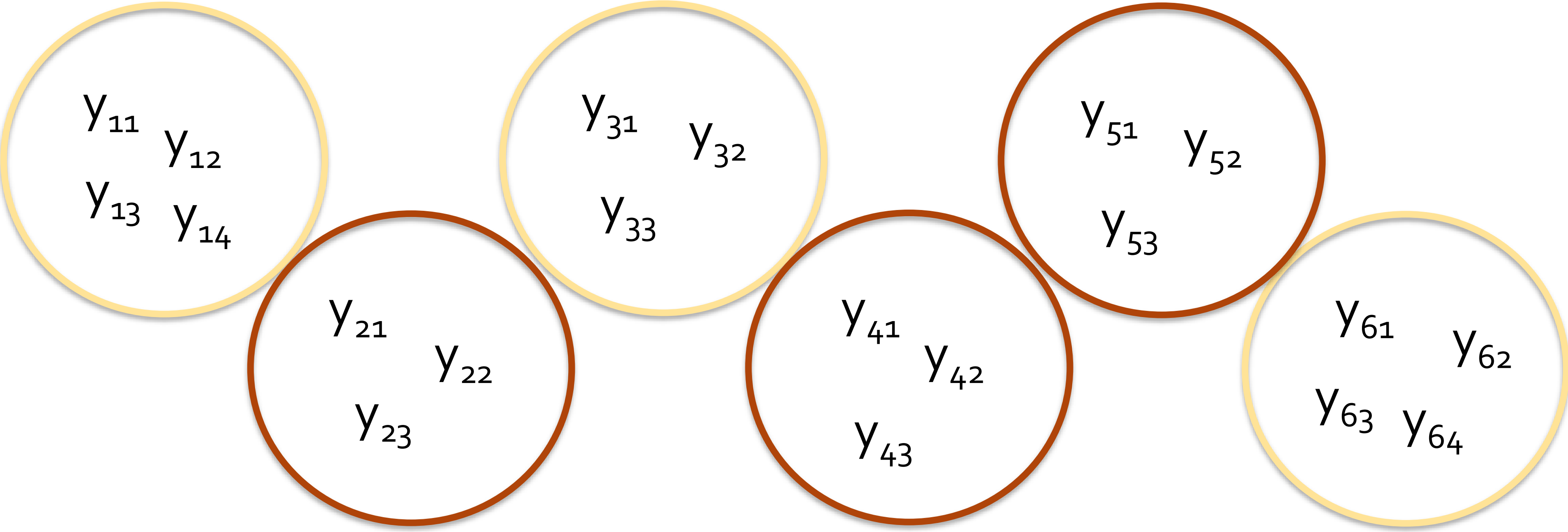
Common options for analysis:

- Generalised linear mixed models
  - mixed, melogit, xtlogit, xtpoisson, etc
- Generalised estimating equation
  - xtgee, xtgeebcv
- Cluster level analysis
  - clan

Cluster-level analysis

The comparison of trial conditions is a between cluster comparison.

- Collapse the data into a summary measure of each cluster, e.g. mean, proportion, or rate
- Analyse these as independent data points e.g. with a t-test



y_{11} y_{12} y_{13} y_{14}

y_{21} y_{22} y_{23}

y_{31} y_{32} y_{33}

y_{41} y_{42} y_{43}

y_{51} y_{52} y_{53}

y_{61} y_{62} y_{63} y_{64}

Cluster-level analysis

The comparison of trial conditions is a between cluster comparison.

- Collapse the data into a summary measure of each cluster, e.g. mean, proportion, or rate
- Analyse these as independent data points e.g. with a t-test

$\bar{y}_1$ $\bar{y}_3$ $\bar{y}_5$

$\bar{y}_2$ $\bar{y}_4$ $\bar{y}_6$

Example: Mema Kwa Vijana trial

In 2003, half of all new HIV infections in sub-Saharan Africa were in adolescents ages 15-24.

Educational intervention given to school children to reduce transmission of HIV in adolescents.

20 communities in Tanzania randomised to the intervention or control.

A secondary outcome was knowledge of HIV acquisition assessed separately in boys and girls. Binary outcome 3/3 questions correct

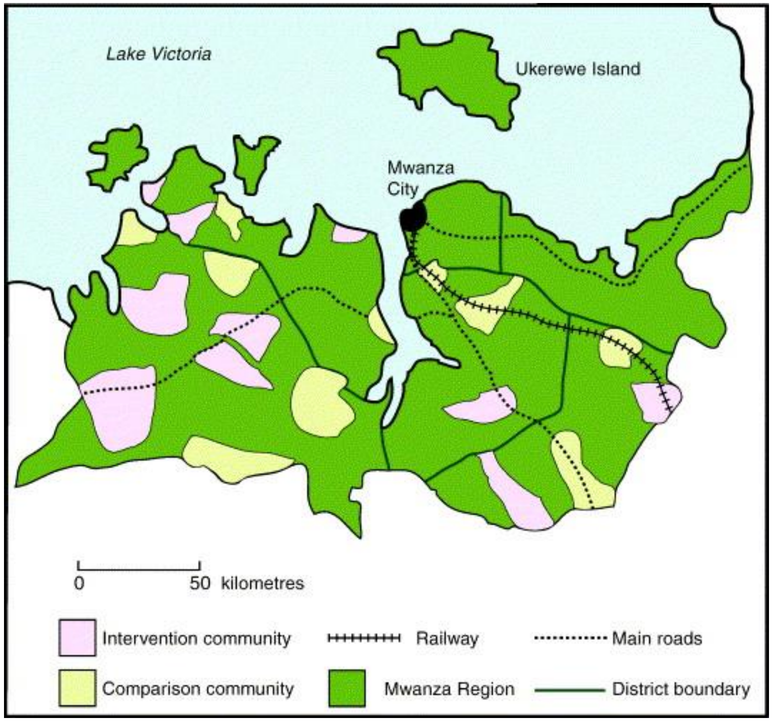
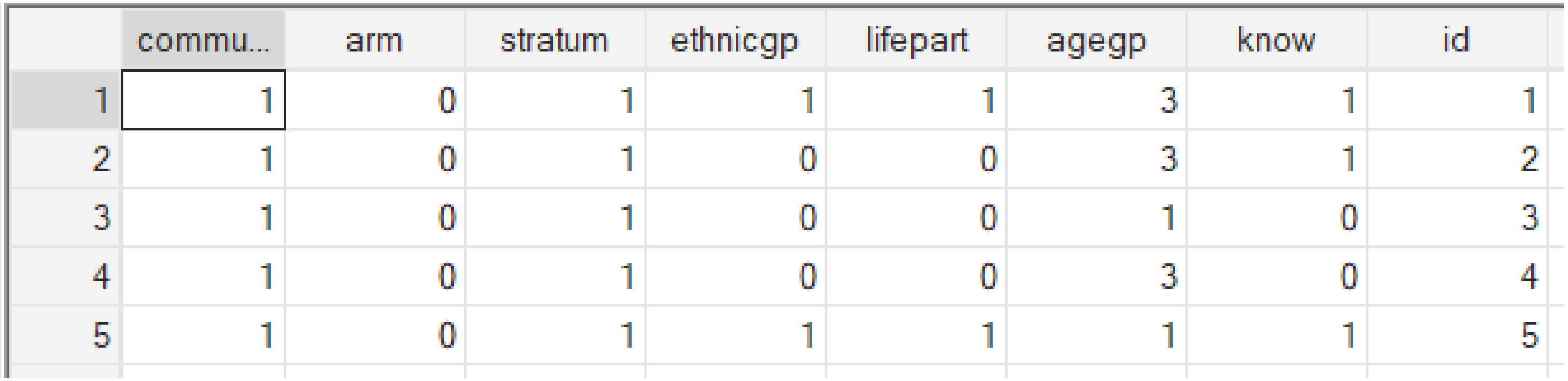


Image from Hayes, Richard J., et al. "The MEMA kwa Vijana project: design of a community randomised trial of an innovative adolescent sexual health intervention in rural Tanzania." *Contemporary clinical trials* 26.4 (2005): 430-442.

Example: Mema Kwa Vijana trial

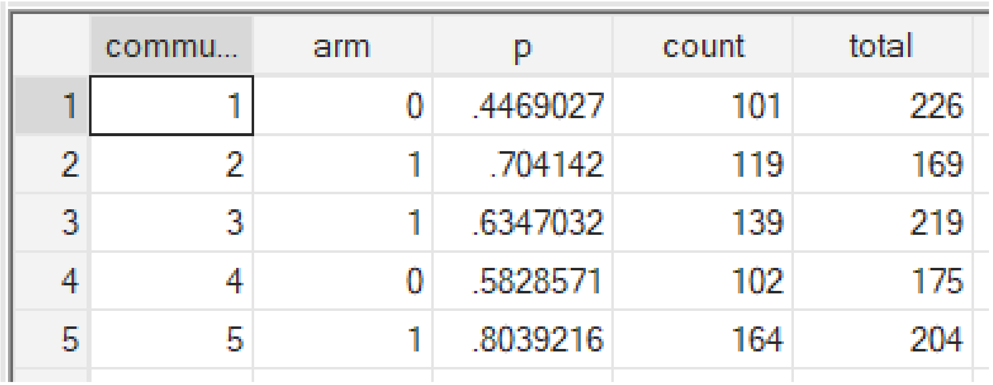
Control	Intervention
110/226 (49%)	164/204 (80%)
65/171 (38%)	141/206 (68%)
69/178 (39%)	111/171 (65%)
87/194 (45%)	139/219 (64%)
102/229 (45%)	115/207 (56%)
84/243 (35%)	172/237 (73%)
121/196 (62%)	111/187 (59%)
101/226 (45%)	119/169 (70%)
102/175 (58%)	157/219 (72%)
67/186 (36%)	127/257 (49%)

Individual level data:



	commu...	arm	stratum	ethnicgp	lifepart	agegp	know	id
1	1	0	1	1	1	3	1	1
2	1	0	1	0	0	3	1	2
3	1	0	1	0	0	1	0	3
4	1	0	1	0	0	3	0	4
5	1	0	1	1	1	1	1	5

Or summarise for each cluster:



	commu...	arm	p	count	total
1	1	0	.4469027	101	226
2	2	1	.704142	119	169
3	3	1	.6347032	139	219
4	4	0	.5828571	102	175
5	5	1	.8039216	164	204

Example: Mema Kwa Vijana trial

Control arm: 908/2024 (45%)

Intervention arm: 1356 / 2076 (65%)

Risk difference (95% confidence interval): 21% (12%, 29%)

P-value: 0.0001

Which analysis should I use?

LONDON
SCHOOL *of*
HYGIENE
&TROPICAL
MEDICINE

Number of Clusters

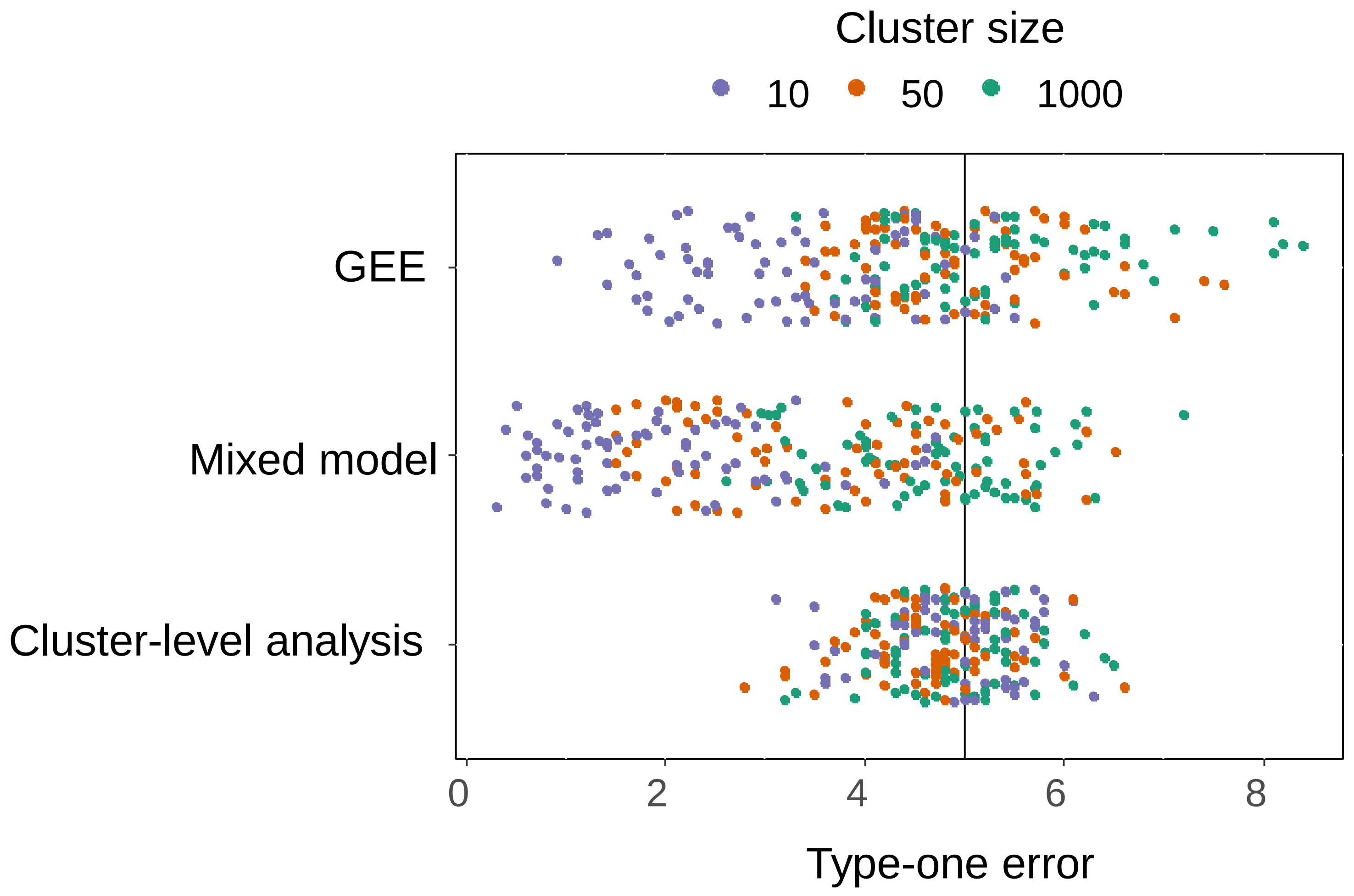
Common to have less than 30 cluster

Individual level methods need adaptations

- Degree of freedom corrections are needed.
- GEE need small sample corrections applied to the robust standard errors.
- Generalised linear mixed models must use REML corrections, but these aren’t always available for non-normal outcomes: I don’t think Stata has a method that applies a REML-type correction for binary outcomes.

Cluster-level analysis performs well regardless of number of clusters.

Simulation study type-one error



Each point a different scenario with 12 clusters varying cluster size, ICC, outcome prevalence, cluster mean distribution

MKV trial odds ratios

Method	Odds ratio (95% confidence interval)	P-value
Cluster-level	2.40 (1.66, 3.47)	0.0001
Mixed effect model *	2.38 (1.68, 3.39)	0.00006
GEE	2.33 (1.60, 3.38)	0.0002

```
collapse (mean) p=know (sum) count = know (count) total = know,
by(community arm)
gen lodds = log(p / (1-p))
ttest lodds, by(arm)

xtlogit know i.arm, i(community) re
di exp(_b[1.arm] - invttail(18, 0.025) * _se[1.arm])
di exp(_b[1.arm] + invttail(18, 0.025) * _se[1.arm])
di 2 * ttail(18, abs(_b[1.arm] / _se[1.arm]))

xtgeebcv know i.arm, cluster(community) stderr(fg)
```

*Uses Adaptive quadrature, which is NOT a restricted maximum likelihood approach

Odds ratio or Risk ratio?

Meaningful effects estimated

Cluster level analysis makes it simple to estimate a risk ratio and risk difference for binary outcomes. Just by changing the summary statistic calculated for the clusters:

- Risk difference: calculate the risk
- Risk ratio: calculate the log risk
- Odds ratio: calculate the log odds

It is possible for some individual level analysis but often struggle to converge or methods are not implemented in software. Change the link function of the analysis model:

- Risk difference: Identity link
- Risk Ratio: Log link

MKV trial risk difference

Risk difference			
Cluster level analysis	21%	(12%, 29%)	0.0001
GEE	21%	(12%, 29%)	0.00008
Mixed effect model	Identity link not allowed in meglm		

```
clan know , arm(arm) cluster(community) effect(rd)

xtgeebcv know i.arm, cluster(community) family(binomial)
link(identity)
```

MKV trial risk ratio

Risk ratio			
Cluster level analysis	1.47	(1.25, 1.73)	0.0001
GEE	1.46	(1.24, 1.72)	0.0001
Mixed effect model	Log link not allowed in meglm		

```
clan know, cluster(community) arm(arm) effect(rr)

xtgeebcv know i.arm, cluster(community) family(binomial)
link(log)
```

Loss of power?

With a large number of clusters, the cluster-level analysis will have less power than a mixed effect model if cluster size varies.

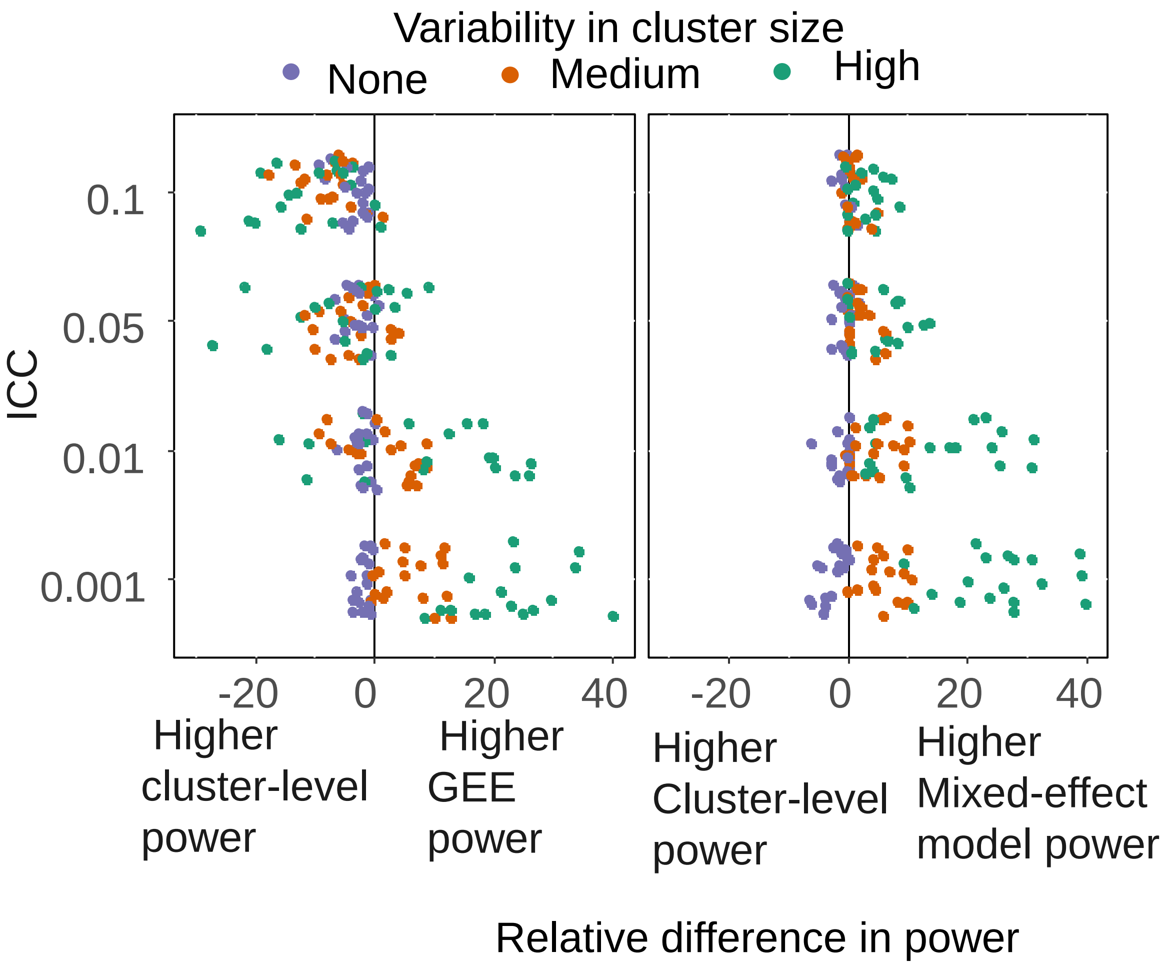
It is common for cluster size to vary

We don't lose much power (if any) with a small number of clusters

MKV trial:

Clusters varied in size from 169 to 257 and we saw little difference in the confidence interval width and p-value

Simulation study power



Each point a different scenario with 20 clusters varying cluster size, ICC, outcome prevalence, cluster mean distribution

Adjusting for covariates

Is it more difficult to adjusting for individual level covariates in a cluster-level analysis?

This is important in a cluster randomised trial. There are fewer units of randomisation (average ~30), so randomisation doesn't ensure balance. So adjusted analysis is commonly used as the primary analysis.

Adjusting for individual level covariates with an individual level analysis is simple: just add them into the model

With a cluster-level analysis, adjusting for individual level covariates is more difficult, but can be done.

With clan command, its no more difficult that with GEE or mixed effect model

MKV trial adjusted analysis

Risk ratio	Unadjusted	Adjusted
Cluster level analysis	1.47 (1.25, 1.73) p=0.0001	1.44 (1.26, 1.66) p=0.00003
GEE	1.46 (1.24, 1.72) p=0.0001	1.41 (1.21,1.64) p=0.0001

```
clan know i.ethnicgp i.agegp,  cluster(community)
arm(arm) effect(rr)
```

```
xtgeebcv know i.arm i.ethnicgp i.agegp,
cluster(community) family(binomial) link(log)
stderr(fg)
```

Manual adjusted cluster-level analysis

Describe 2 stage adjustment for covariates

1. Regress outcome on covariates ignoring clustering and the intervention

2. Use this to calculate a residual for each cluster

Risk difference residual = observed risk – expected risk

Risk ratio residual = observed risk / expected risk

3. Use a t-test or other simple analysis method to analyse these residuals

MKV trial manual adjustment

```
logit know i.ethnicgp i.agegp
predict expected
list community arm know expected in 1/10
```

```
     +------------------------------------+
     | commun~y   arm   know   expected |
     |------------------------------------|
  1. |        5     1      1   .5622647 |
  2. |       14     0      0   .4910673 |
  3. |       19     0      0   .5016555 |
  4. |       14     0      0   .5622647 |
  5. |       18     1      0   .5016555 |
     +------------------------------------+
```

MKV trial manual adjustment

```
collapse (sum) know expected (count) clustersize = know,
by(community arm)
gen residual = (know/clustersize) / (expected/clustersize)

list arm know expected clustersize residual in 1/5
```

```
     +-----------------------------------------------+
     | arm   know   expected   cluste~e   residual |
     |-----------------------------------------------|
  1. |   0    101   130.6926        226   .7728055 |
  2. |   1    119   99.74625        169   1.193027 |
  3. |   1    139    119.225        219   1.165863 |
  4. |   0    102   101.9492        175   1.000499 |
  5. |   1    164   115.6075        204   1.418593 |
     +-----------------------------------------------+
```

MKV trial manual adjustment

```
ttest residual, by(arm)
```

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
       0 |      10    .8221376    .0504584    .1595636    .7079927    .9362826
       1 |      10    1.175819    .0399415     .126306    1.085465    1.266173
---------+--------------------------------------------------------------------
Combined |      20    .9989783    .0512521    .2292064    .8917064     1.10625
---------+--------------------------------------------------------------------
    diff |           -.3536814    .0643535               -.4888831   -.2184797
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -5.4959
H0: diff = 0                                     Degrees of freedom =       18

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Conclusions

There are many approaches to analysing cluster randomised trials.

The cluster-level analysis isn’t used as frequently as other approaches (mixed models are by far the most common!)

But… the clan command makes implementing the cluster-level analysis just as easy to implement with very reliable type-one error, reasonable power, and easy estimation of meaningful effects

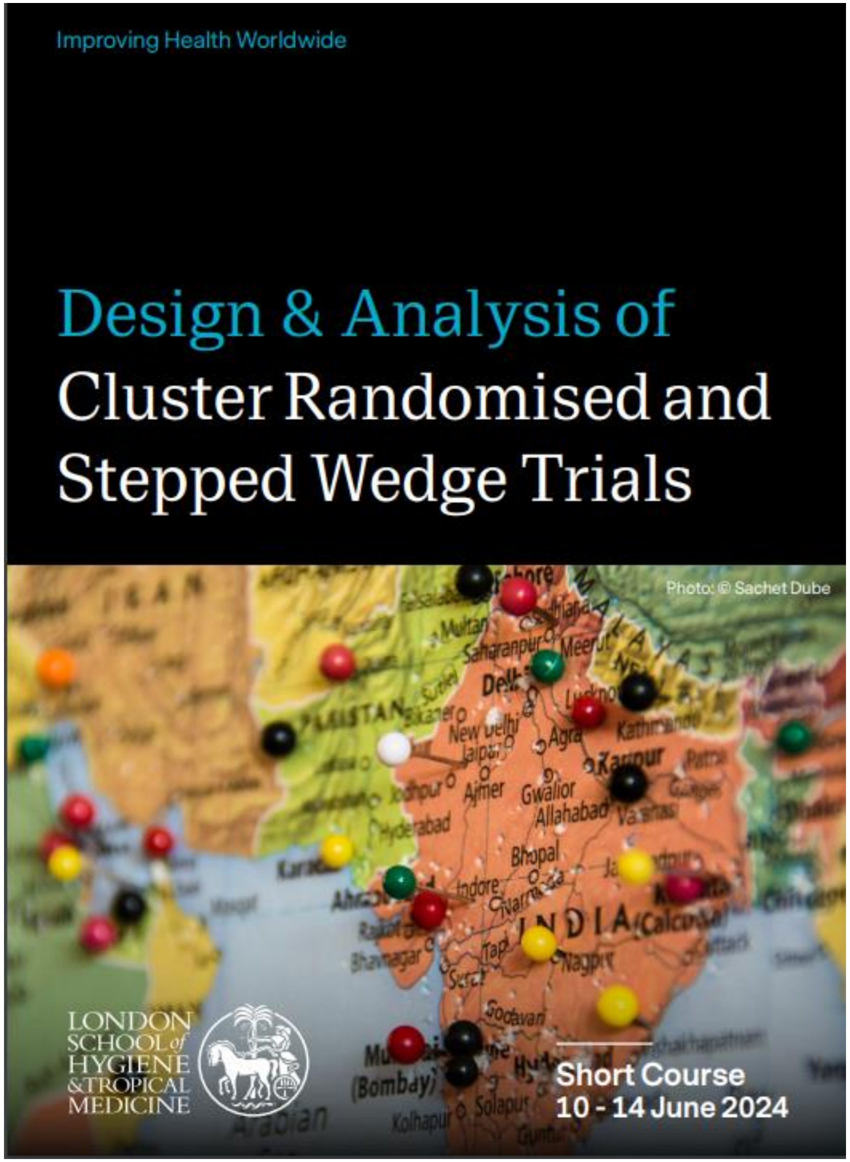
Thank you

Acknowledgements: Stephen Nash, Baptiste Leurent, Larry Moulton, Richard Hayes, Katherine Fielding, Clemence Leyrat

jennifer.thompson@lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

International Statistics & Epidemiology Group



Improving Health Worldwide

Design & Analysis of Cluster Randomised and Stepped Wedge Trials

Photo: © Sachet Dube

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Short Course
10 - 14 June 2024

https://www.lshtm.ac.uk/study/courses/short-courses/cluster-randomised-trials

References

Thompson, Jennifer A., et al. "Cluster randomized controlled trial analysis at the cluster level: The clan command." *The Stata Journal* 23.3 (2023): 754-773.

Thompson, Jennifer A., et al. "Cluster randomised trials with a binary outcome and a small number of clusters: comparison of individual and cluster level analysis method." *BMC Medical Research Methodology* 22.1 (2022): 1-15.

Hayes, Richard J., et al. "The MEMA kwa Vijana project: design of a community randomised trial of an innovative adolescent sexual health intervention in rural Tanzania." *Contemporary clinical trials* 26.4 (2005): 430-442

Kahan, Brennan C., et al. "Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study." Trials 17 (2016): 1-8.

Gallis JA, Li F, Turner EL. xtgeebcv: A command for bias-corrected sandwich variance estimation for GEE analyses of cluster randomized trials. Stata J. 2020 Jun;20(2):363-381. doi: 10.1177/1536867x20931001. Epub 2020 Jun 19. PMID: 35330784; PMCID: PMC8942127.