Pragmatic trial designs

Topics covered in this course

- Cluster randomized trials
- Stepped Wedge Trials
- Real world effectiveness of vaccines

Research phases and types of studies

Research Phase	Basis Science Research	Translational Research	Clinical Research	Implementation Research	Clinical Practice
Objectives	Understand biology and mechanism of action of potential treatment	Translate basic science knowledge to humans	Evaluate efficacy	Translate research to practice, Evaluate effectiveness	Opportunity to identify new clinical questions and gaps in patient care
Type of studies	Laboratory studies, Preclinical studies, Animal studies	Phase I and II clinical trials	Observational studies, Phase III clinical trials (can include some pragmatic elements)	Phase IV clinical trials, Pragmatic clinical trials	

A recap on explanatory clinical trials

- Gold standard to evaluate efficacy of an intervention Individuals are randomized to treatments (placebo or active control)
- Treatments are specified according to a carefully devised trial protocol, and patients enrolled in the trial must meet a set of prespecified eligibility criteria
- Randomization and blinding - High internal validity
- Resource intensive
- External validity? Generalizability?

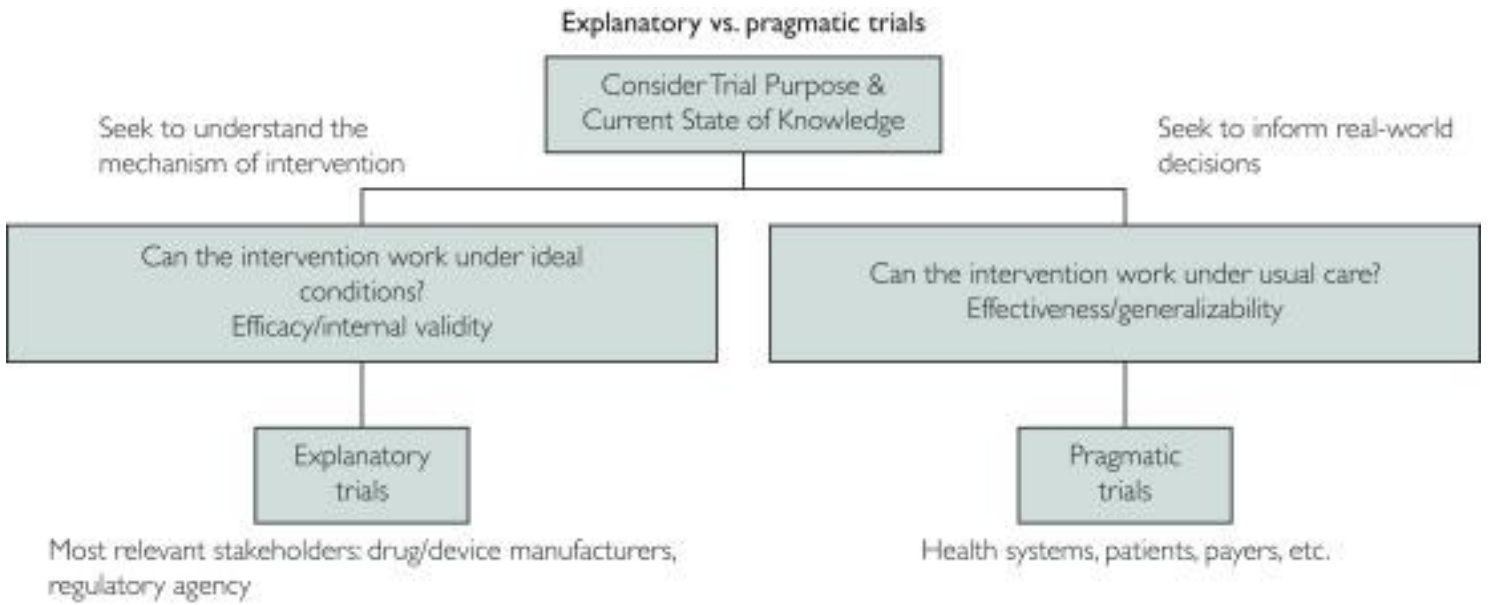
Pragmatic Clinical Trials

- Pragmatic clinical trials are often conducted as part of implementation science, which connects clinical research and clinical practice
- Pragmatic research is designed with input from health systems—and produces evidence that can be readily used to improve care.
- By engaging health systems, providers, and patients as partners, pragmatic research accelerates the integration of research, policy, and practice
- The PRECIS-2 (Pragmatic Explanatory Continuum Indicator Summary-2) tool is useful to understand the level and the aspect of pragmatism of a trial’s design

Pragmatic Clinical Trials

- Two common pragmatic research features
- Use of electronic health records (EHRs)
- Randomization of treatment alternatives based on normal health care operations (ie. clinic or provider level)

Decision tree for explanatory vs pragmatic trials



Design elements of pragmatic vs explanatory trials

Feature	**Explanatory trials**	**Pragmatic trials**
Eligibility criteria	Narrow	Broad
Recruitment strategies	Targeted	General
Clinic setting	Academic	Community
Expertise and resources	Research	Clinical practice
Intervention delivery	Prescriptive	Flexible
Intervention adherence	Closely monitored	As usual care
Follow-up requirements	Closely followed	As usual care
Primary outcomes	Relevant to physicians	Relevant to patients
Primary analysis	Can be per protocol	Intention-to-treat

Key terms

Term	Definition
Efficacy	The biological effect of a treatment on outcomes. Treatment efficacy is evaluated in a well-controlled environment to isolate the direct effect of the treatment from the influence of external factors
Effectiveness	The effect of a treatment on outcomes in an uncontrolled real-world setting of clinical practice
Explanatory clinical trials	Trials that aim to document the safety and efficacy of an intervention under ideal conditions
Pragmatic clinical trials	A spectrum of trials with design elements that closely resemble how patients are treated in clinical practice and are intended to evaluate how treatments affect outcomes in the real-world setting clinical practice

References

- Thorpe KE, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. Can Med Assoc J, 2009, 180: E47-57.
- Tunis SR, et al. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA, 2003;290:1624-1632.
- Glasgow RE, et al. Practical clinical trials for translating research to practice: design and measurement recommendations. Med Care, 2005;43(6):551-557.
- Zwarenstein M, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ, 2008 Nov 11;337:a2390

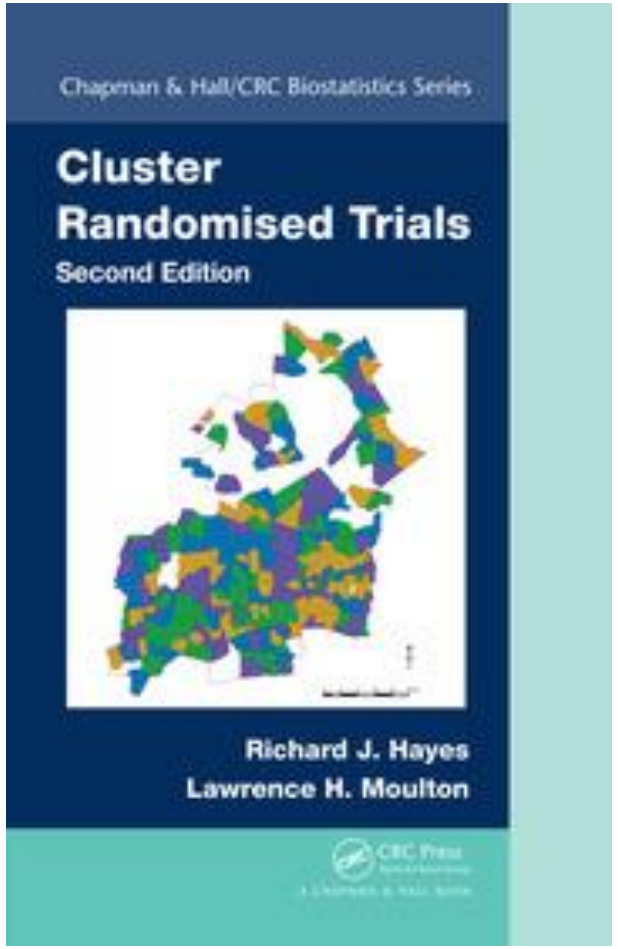
Topics covered in this course

- Cluster randomized trials
- Stepped Wedge Trials
- Real world effectiveness of vaccines

Cluster Randomized Trials

Introduction

Key text for this course



Design and analysis of CRTs – course at LSTHM

Books on CRT's

- Donner A & Klar N. (2000) Design and Analysis of Cluster Randomization Trials in Health Research. Arnold
- Murray D M (1998) Design and Analysis of Group-Randomised Trials. Oxford
- Richard J. Hayes, Lawrence H. Moulton (2009) Cluster Randomised Trials. Chapman & Hall/CRC (2nd edition 2017)
- Spiegelhalter D, Abrams K, Myles J (2004) Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Wiley

Rationale for randomizing by cluster

- The intervention by its nature has to be applied to entire communities
- or other groupings of individuals, or it is more convenient or acceptable to apply it in this way.
- We wish to avoid the contamination that might result if individuals in
- the same community were to be randomised to different treatment arms.
- We wish to capture the population-level effects of an intervention applied to a large proportion of a population, for example an intervention designed to reduce the transmission of an infectious agent.

Some case studies

A. Reducing adolescent tobacco use

- 12 schools randomly assigned to health promotion intervention called
  Smoke-free generation
- 12 schools acted as controls
- Primary outcome was proportion of children smoking after 2 years

Some case studies

- ***Data from Smoke free generation trial***

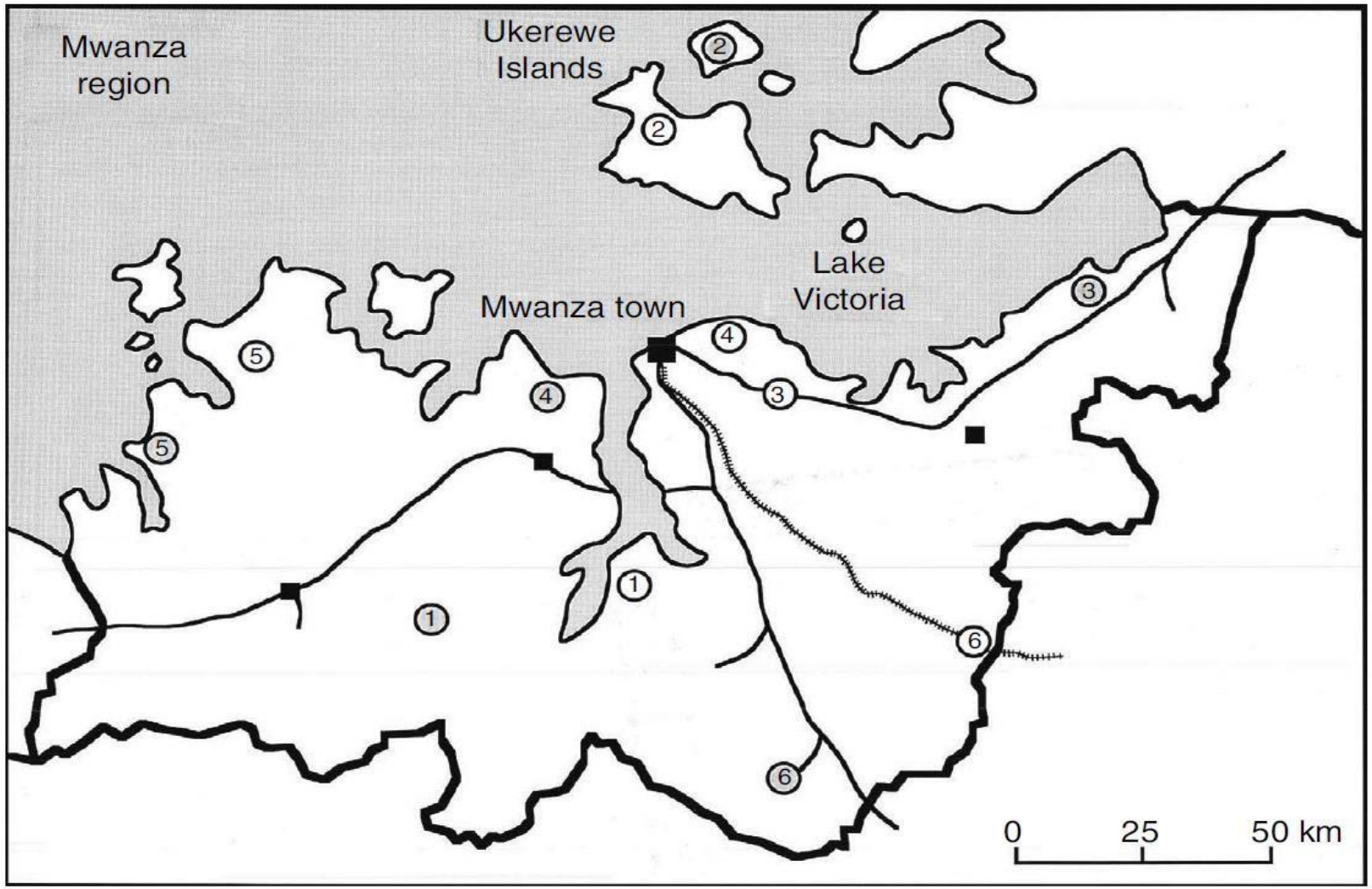
	Intervention		Control	
	0 / 42	0	5 / 103	0.049
	1 / 84	0.012	3 / 174	0.017
	9 / 149	0.060	6 / 83	0.072
	11 / 136	0.081	6 / 75	0.080
	4 / 58	0.069	2 / 152	0.013
	1 / 55	0.018	7 / 102	0.069
	10 / 219	0.046	7 / 104	0.067
	4 / 160	0.025	3 / 74	0.041
	2 / 63	0.032	1 / 55	0.018
	5 / 85	0.059	23 / 255	0.102
	1 / 96	0.010	16 / 125	0.128
	10 / 194	0.052	12 / 207	0.058
Overall	58 / 1341	0.043	91 / 1479	0.062

Some case studies

B. Reducing HIV incidence by controlling STDs

- Tanzania: 6 pairs of rural communities were chosen
- One of each pair: Improved STD treatment services
- Random sample of 1000 adults selected in each community
- Primary outcome: HIV incidence over 2 years

Some case studies

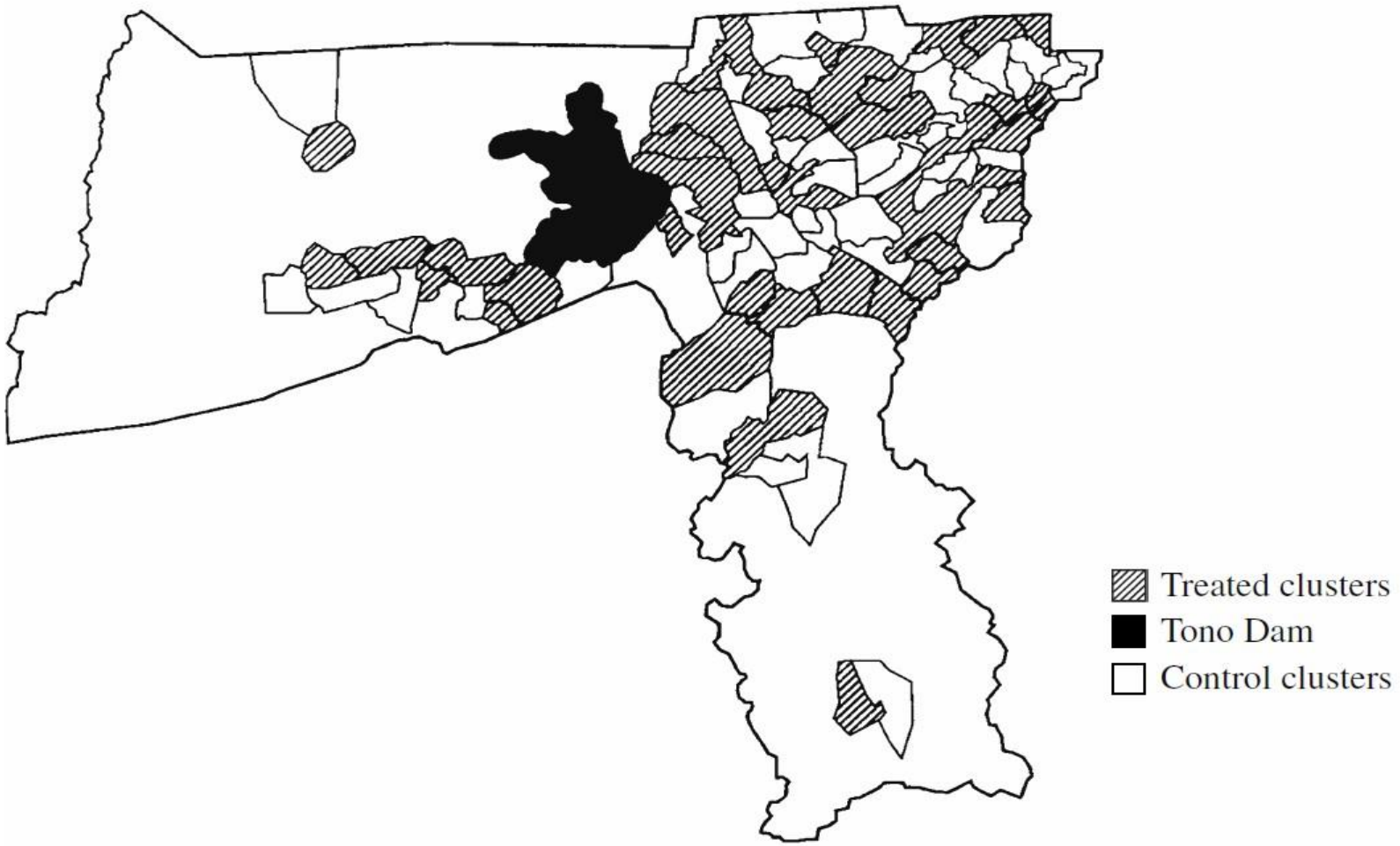


Some case studies

C. Impact of insecticide-treated bednets on child mortality

- Ghana: 96 clusters of compounds
- Clusters were arbitrary geographical zones
- 48 randomly selected to receive insecticide-treated nets
- Demographic surveillance used to record all births/deaths/migrations

Some case studies



Why do we need specialized methods to analyze CRTs?

- Observations on individuals in the same cluster are usually correlated.

- With small numbers of clusters, we cannot rely on randomisation to ensure adequate balance between arms.

Where does the Between Cluster Variation come from?

- Patients select the cluster to which they belong. Patient characteristics could be related to age and sex differences among doctors
- Important covariates at the cluster level affect all individuals within the cluster in the same manner. Example: Stock out of ARV drugs in certain clinics may be related to virologic failure rates
- Patients within clusters frequently interact, as a result respond similarly.
- Example: Promotion of exclusive breastfeeding in HIV positive mothers may be enhanced through positive feedback from mothers who follow this practice or vice versa.

Between-cluster variability and within-cluster correlation

Individually randomised trial: Observations generally assumed to be *independent* of one another

Cluster randomised trial: Observations on individuals in the same cluster are usually *correlated*

- Outcomes for individuals in *same* cluster likely to be more similar than outcomes for inidividuals in *different* clusters

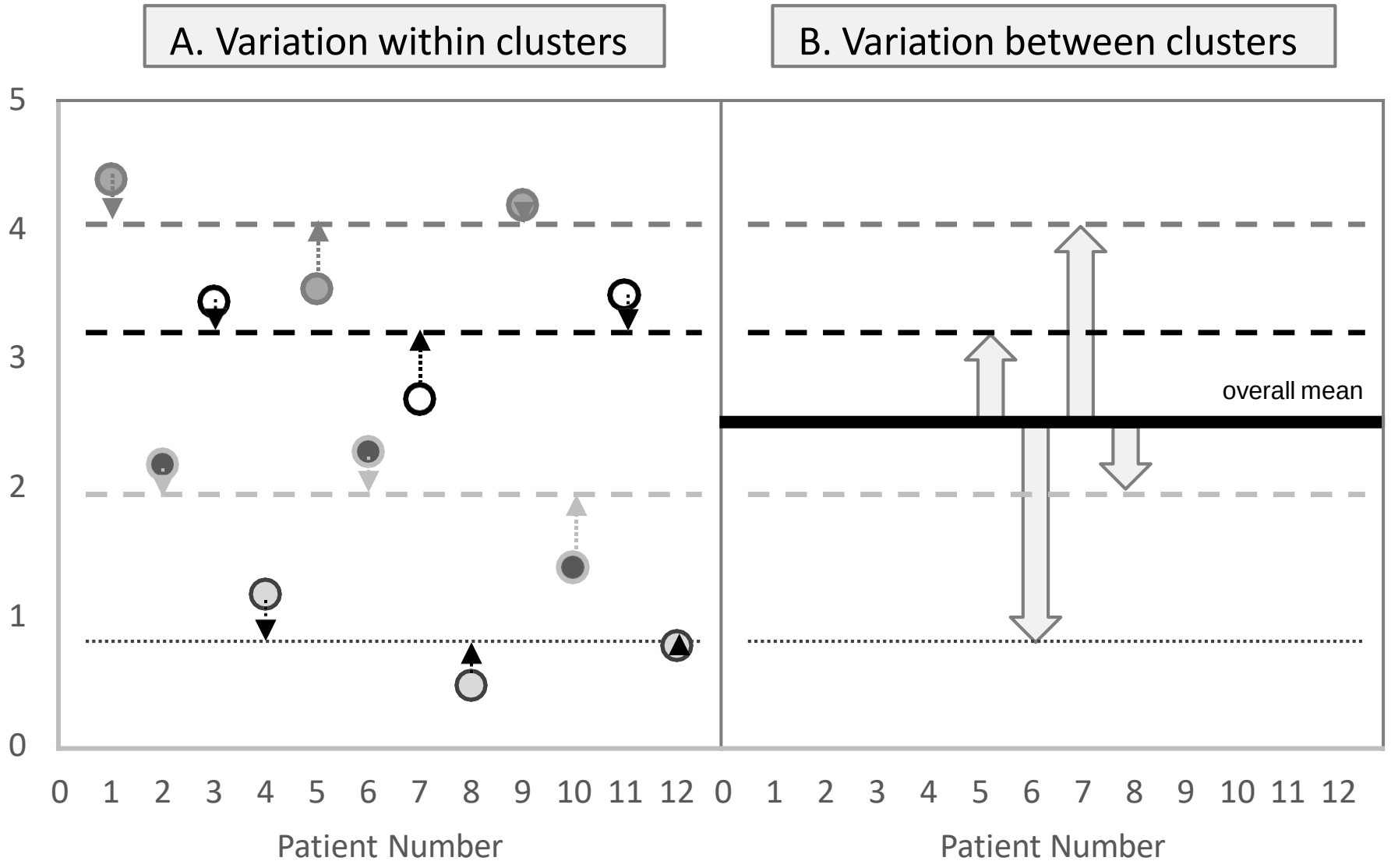
Note: While individuals *within* clusters are correlated, the clusters themselves are assumed to be independent

Example

- Training programme for GPs to improve treatment of back pain
- 8 GPs randomly allocated to intervention (4 GPs) or control (4 GPs)
- Performance score on each patient

Example

Data from control arm (data on 3 patients for each GP):



Between-cluster variability and within-cluster correlation

Some reasons for within-cluster correlation

- Individuals tend to *behave* or *respond* more like others in the same cluster than like others in a different cluster
- Individuals may have *level of exposure* more like others in the same cluster than like others in a different cluster
- *Transmission of infection* between individuals in the same cluster

Between-cluster variability and within-cluster correlation

Implications of within-cluster correlation

Total variation	=	Between-individual variation	+	Between-cluster variation
		Binomial, Poisson		Extra-binomial, Extra-Poisson, Overdispersion

→ Standard errors are increased

This means that:

- Sample size: Has to be increased to take account of clustered design

- Analysis: Has to take account of clustered design
  Note: standard methods will give CIs that are too wide and p-values that are too small

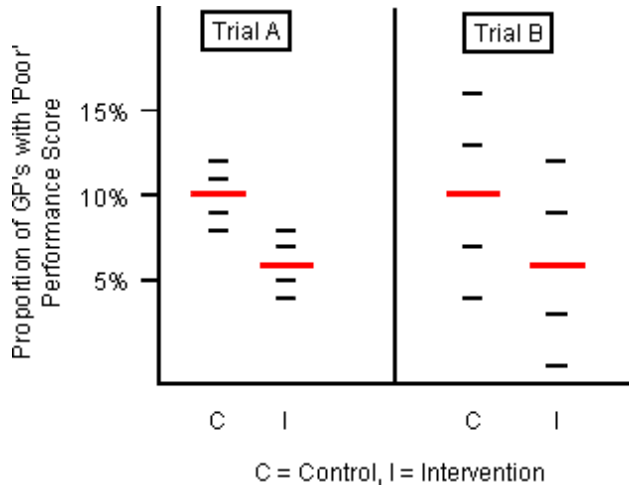
Between-cluster variability and within-cluster correlation

Example (back pain trial): 2 scenarios to compare

Trial A		Trial B	
Control	**Intervention**	**Control**	**Intervention**
8/100 (8%)	4/100 (4%)	4/100 (4%)	0/100 (0%)
9/100 (9%)	5/100 (5%)	7/100 (7%)	3/100 (3%)
10/100 (10%)	6/100 (6%)	10/100 (10%)	6/100 (6%)
11/100 (11%)	7/100 (7%)	13/100 (13%)	9/100 (9%)
12/100 (12%)	8/100 (8%)	16/100 (16%)	12/100 (12%)
50/500 (10%)	**30/500 (6%)**	**50/500 (10%)**	**30/500 (6%)**

- Analysis ignoring clustering
  - Chi-square p = 0.02
  - Trials A and B!

- Correct analysis:
  - Trial A: p = 0.004
  - Trial B: p = 0.22



Measures of between-cluster variability k and ρ

Coefficient of variation, *k*

Notation: *c* clusters, (true) rate in *jth* cluster is defined as λ_j

Assume the clusters have been randomly sampled from some wider population of clusters, with mean λ and variance σ_B^2

Then the *coefficient of variation* is defined as

$$k = \sigma_B / \lambda$$

Note: Similar definitions for proportions and means:

$k = \sigma_B / \pi$, $k = \sigma_B / \mu$

Example

- Suppose we are planning a CRT to investigate the effect of providing insecticide treated bednets on the prevalence of malaria parasitaemia in village children aged under 5 years. We plan to randomly allocate entire villages to the intervention andcontrol arms, and need an estimate of between-village variability to assist with the study design.Malaria parasitaemia is a binary outcome, and we are interested in making inferences about the proportion, π, with this outcome. Malariologists who are familiar with the area tell us that the prevalence of malaria parasitaemia averages around 30% but could easily vary between 15 and 45% in individual villages. If we are prepared to assume that the village prevalences are approximately normally distributed, this would imply a σB of around 0.075, since roughly 95% of prevalences would fall within two standard deviations of the mean (30% ± 2 × 7.5% gives the interval 15–45%).
- Thus we could obtain an estimate of k as:

$$k = \sigma B/\pi = 0.075/0.30 = 0.25$$

Key concepts and notation

- Assuming that x is a binary or continuous variable, the intracluster correlation coefficient is defined as

$$\rho = \frac{E(x_{jk}-\mu)(x_{jk`}-\mu)}{E(x_{ik}-\mu)^2}$$

- Where x_{jk} is the observed outcome for the kth individual in the jth cluster (j = 1, ..., c).

In this equation, the expectation in the numerator is over all distinct pairs of individuals (k ≠ k') taken from the same cluster and over all clusters; the expectation in the denominator is over all individuals and all clusters; and μ is the true mean over the entire population.

Measures of between-cluster variability k and ρ

Intra-cluster correlation coefficient, ρ

ρ measures what proportion of total variance is accounted for by variation between clusters

For quantitative outcomes:

$$\rho = \frac{\sigma_B^2}{\sigma^2} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$$

For binary outcomes:

$$\rho = \frac{\sigma_B^2}{\pi(1-\pi)}$$

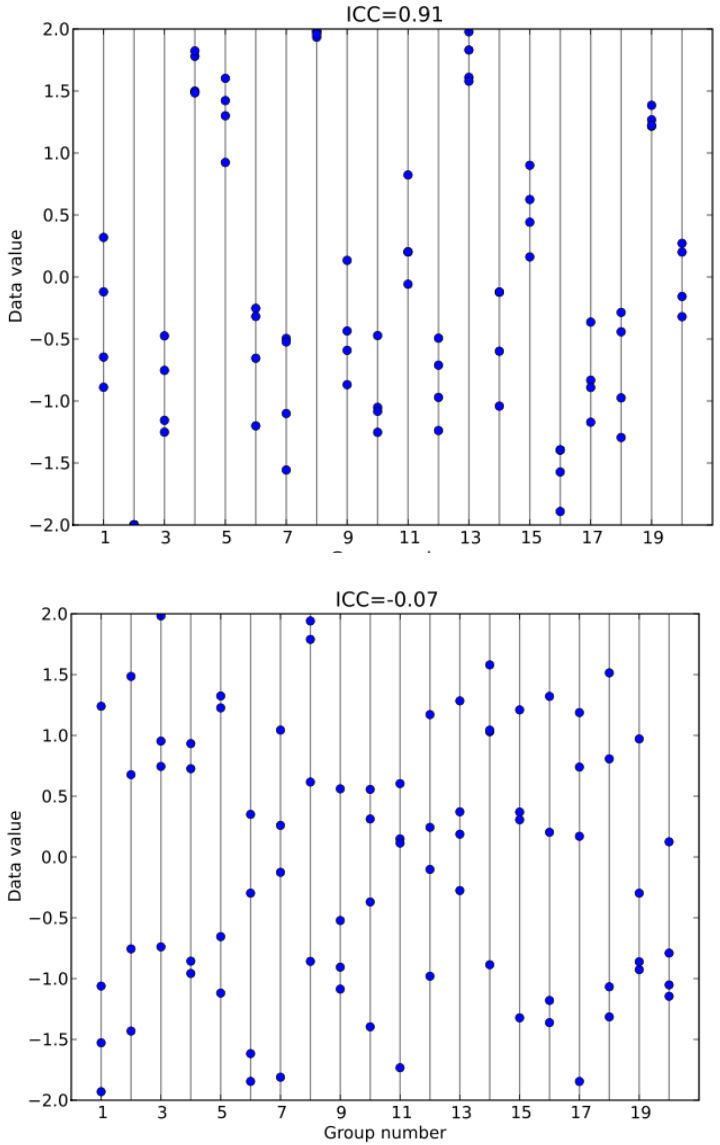
Note: ρ is not defined for rates, so use k instead for sample size calculations

ICC

- It describes how strongly units in the sam
- IF>0 if m large and ICC small
- m=500 , ICC=.001 >> IF=1.5

- IF>0 if m small and ICC moderate
- m=20, ICC=.05 >> IF=2

- Typical ICC
- <.01 for large facilities
- <.05 for moderate size facilities
- <.15 for husband-wife pair



Measures of between-cluster variability k and ρ

Design effect = Var_{CRT} / Var_{RCT} = $\frac{\text{Sample size for CRT}}{\text{Sample size for RCT}}$

A simple rule: Design effect for clusters of size m is given by

$$DEff = 1 + (m - 1)\,\rho$$

m sample size within cluster
c number of clusters
If m varies use average- harmonic mean
since p is undefined for rates, one can use k

Increase in variance due to clustering. CRT have larger variance

Reasons for using cluster randomisation

- **Interventions designed to be implemented at cluster level**
  - Water/sanitation project
  - School-based health education
  - Training of health care workers at clinic

- **Logistical convenience and acceptability**
  - Delivering intervention to some households but not to their neighbours in a rural village
  - Practical considerations: e.g. Vitamin A supplementation trial in Ghana

- **Contamination**
  - Intervention messages or products may be shared between individuals in the community
  - Note: CRT design helps to reduce contamination but may not eliminate it completely

Reasons for using cluster randomisation

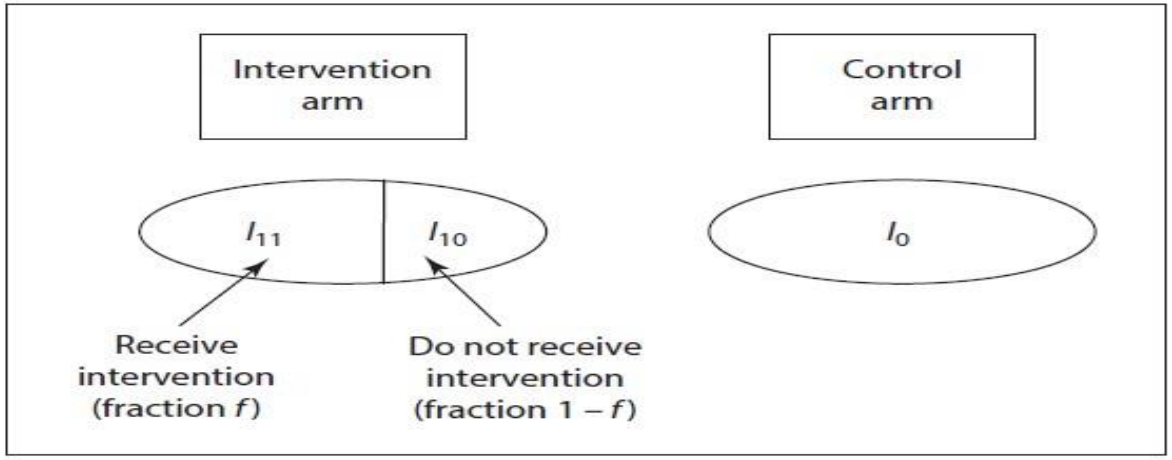
- **Capturing the *indirect effects* of an intervention**
  - Effects of intervention on *infectivity* of individuals
  - Herd immunity
  - Mass effects of intervention

Note: *Indirect effect* means that individuals who do not receive the intervention get some benefit from the intervention

e.g HIV prevention trials

Reasons for using cluster randomisation

- **Direct, indirect, total and overall effects**



$RR_{Direct} = I_{11}/I_{10}$ $\quad\quad$ $RR_{Total} = I_{11}/I_0$

$RR_{Indirect} = I_{10}/I_0$ $\quad\quad$ $RR_{Overall} = I_1/I_0 = [fxI_{11} + (1\text{-}f)xI_{10}] / I_0$

Note: RR_{Direct}, $RR_{Indirect}$ and RR_{Total} are all subject to confounding

Disadvantages of cluster randomisation

- Efficiency
  - CRT has greater variance for the same sample size
- Blinding
  - Often difficult to achieve blinding in a CRT of a cluster-wide intervention, so there is potential for *selection bias*
- Imbalance
  - Often have small number of clusters so randomisation cannot be relied on to ensure *balance* between study arms
- Generalisability
  - Measured effects of intervention may vary due to differences in implementation of intervention in different places
  - Or differences in take-up of intervention
  - Or differences in indirect effects

Summary

- Cluster randomisation may be the method of choice for trials of interventions delivered to groups or communities or where indirect effects are expected
- Between-cluster variability means that observations within clusters are correlated and this must be accounted for in both the design and analysis
- The between-cluster coefficient of variation or the intra-cluster correlation coefficient may be used to measure the extent of between- cluster variability
- The overall effect is the main measure of effect in a CRT but it may also be possible to obtain estimates of direct, indirect and total effects
- CRTs are usually larger than an equivalent individually randomised trial and so the advantages and disadvantages of using this design should be weighed carefully

Goals for this lecture

By the end of this session you will:

- list different **types of clusters** and discuss when they may be used
- discuss what factors should be considered when deciding on the **size of the entire cluster**
- compare different strategies used to reduce **contamination**
- discuss the **rationale behind matching, stratification, and restricted randomisation**
- review the **advantages and disadvantages** of matching, stratification and restricted randomisation
- discuss the **variables** that may be used for matching, stratification or as balance criteria in restricted randomisation
- compare the advantages and disadvantages **of cross-sectional sampling and follow-up of cohorts** to evaluate intervention effects
- evaluate which **design strategy is most suitable** for a variety of situations

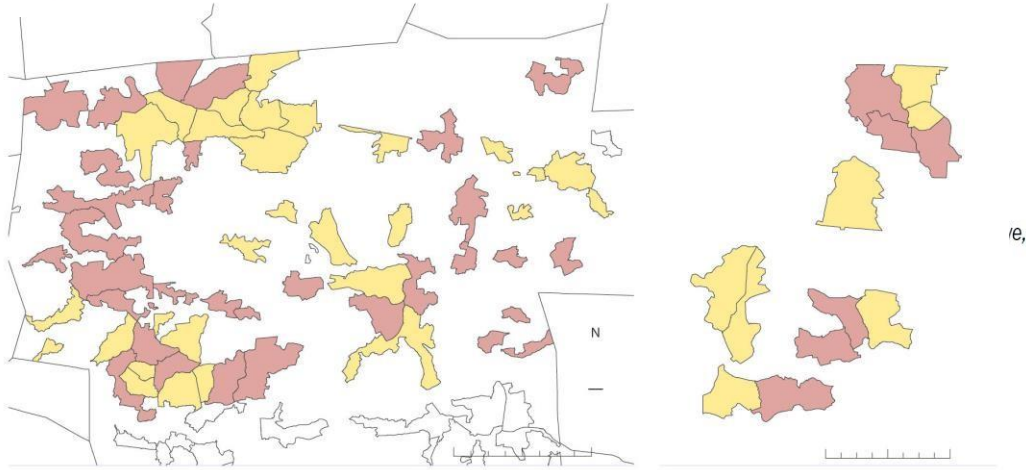
Type of cluster

- Geographical
  - Deliver intervention to the entire target populations in an area
  - Population-level impact of intervention measured - combines the direct and indirect effects of the intervention
  - Defined Communities or Arbitrary geographical zones

Effectiveness and cost-effectiveness of reactive, targeted indoor residual spraying for malaria control in low-transmission settings: a cluster-randomised, non-inferiority trial in South Africa

David Bath, Jackie Cook*, John Govere, Phillemon Mathebula, Natashia Morris, Khumbulani Hlongwana, Jaishree Raman, Ishen Seocharan, Alpheus Zitha, Matimba Zitha, Aaron Mabuza, Frans Mbokazi†, Elliot Machaba, Erik Mabunda, Eunice Jamesboy, Joseph Biggs, Chris Drakeley, Devanand Moonasar, Rajendra Maharaj, Maureen Coetzee, Catherine Pitt‡, Immo Kleinschmidt‡*

- Census wards were mapped and formed into clusters comprising populations of about 5000–10 000 people
- Objective was to determine whether targeted vector control is non-inferior to the standard strategy of Indoor Residual Spraying



Type of cluster (2)

- Institutional
  - Schools

Impact of a theoretically based sex education programme (SHARE) delivered by teachers on NHS registered conceptions and terminations: final results of cluster randomised trial

M Henderson, D Wight, G M Raab, C Abraham, A Parkes, S Scott, G Hart

Project SHARE
- Cluster – secondary schools in Scotland
- School attendees
- Effect of a sex education programme delivered by teachers

  - Health Unit

Effectiveness of a Strategy to Improve Adherence to Tuberculosis Treatment in a Resource-Poor Setting
A Cluster Randomized Controlled Trial

- Cluster – district health centres in Senegal
- Smear+ adult TB patients
- Complex intervention aiming to improve TB treatment adherence

  - Workplaces

A Trial of Mass Isoniazid Preventive Therapy for Tuberculosis Control

Gavin J. Churchyard, M.B., B.Ch., Ph.D., Katherine L. Fielding, Ph.D., James J. Lewis, Ph.D., Leonie Coetzee, D.Soc.Sc., Elizabeth L. Corbett, M.B., B.Chir., Ph.D., Peter Godfrey-Faussett, F.R.C.P., Richard J. Hayes, D.Sc., Richard E. Chaisson, M.D., and Alison D. Grant, M.B., B.S., Ph.D., for the Thibela TB Study Team

Thibela TB study
- Cluster – mine shaft(s)/hostel(s) in South Africa
- All workers
- Examining effect community-wide TB preventive therapy on TB incidence and prevalence

Combines medical centre, hostel and mine shaft

Size of cluster

- Size of cluster could be a choice when defining the cluster
- Distinguish between the population to whom the intervention (treatment) is delivered, and the (sub)-population among whom the outcome is measured
- May chose to measure the outcome of interest among a sample of individuals in each cluster, rather than the entire cluster
  - Considered in more detail in session 3 (sample size)
- Statistical considerations
  - A large number of small clusters is statistically more efficient than a small number of large clusters

$$DEff = 1 + (m - 1)\rho$$

where m=cluster size and ρ is the intracluster correlation coefficient

Keep ρ fixed, then as m increases, *DEff* increases

  - Example: ρ =0.02 and m=50 or 100 → *DEff* =1.98 and 2.98, respectively

  - However, in practice ρ will vary with cluster size: ρ will be smaller for larger clusters

Size of cluster (2)

- Statistical considerations (continued)
  - Important to have estimates of the correlation in responses netween individuals from the same cluster / the variation in responses between clusters when estimating the sample size of a CRT

- Logistic and financial issues
  - May influence the size of the cluster
  - Cost of intervention in many small clusters versus few large clusters
  - Costs for carrying out fieldwork, transport and supervision generally higher in many small clusters compared to fewer large clusters

Contamination

- Contamination (spillover) typically leads to a dilution of the intervention effect
- Can arise as a result of interventions in treatment clusters 'spilling over' into control clusters due to social and/or biological processes :
  1. Contacts between **intervention and control clusters and the wider community**
     - travel or migration between clusters; social network resides in different trial arms
     - Movements across cluster boundaries or when individuals live close to cluster boundaries, so called edge-effects
     - Eg., health education message intervention: in-migration may dilute out the effects of the intervention. In-migration may also bring in individuals with infectious disease, which will affect the results of trials evaluating interventions against them
  2. Biological spillover: if the intervention is vector control (e.g. control of malaria or dengue vector mosquitoes), movement of the vector may lead to contamination

Contamination (2): biological spillover exa

The NEW ENGLAND JOURNAL of MEDICINE
ESTABLISHED IN 1812 JUNE 10, 2021 VOL. 384 NO. 23

Efficacy of Wolbachia-Infected Mosquito Deployments for the Control of Dengue

A. Utarini, C. Indriani, R.A. Ahmad, W. Tantowijoyo, E. Arguni, M.R. Ansari, E. Supriyati, D.S. Wardana, Y. Meitika, I. Ernesia, I. Nurhayati, E. Prabowo, B. Andari, B.R. Green, L. Hodgson, Z. Cutcher, E. Rancès, P.A. Ryan, S.L. O'Neill, S.M. Dufault, S.K. Tanamas, N.P. Jewell, K.L. Anders, and C.P. Simmons, for the AWED Study Group*

- Randomly assigned 12 geographic clusters to receive deployments of wolbachia-infected *A. aegypti* (intervention) and 12 clusters to control (no deployment)
- Endpoint: virologically confirmed dengue
- W. infection was monitored in all clusters (see graph)
- Wolbachia infection gradually spread from intervention to control clusters (graph B)

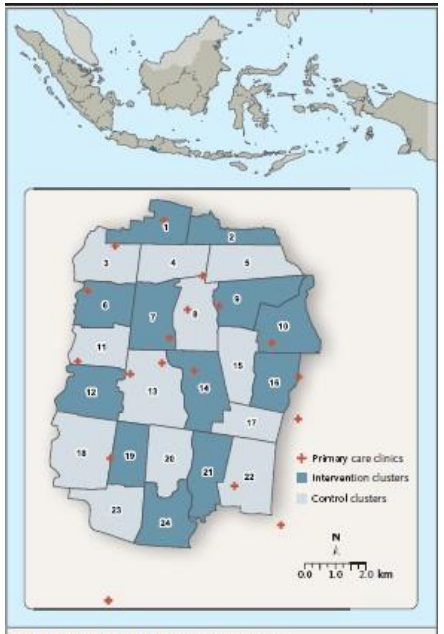


Figure 1. Map of the Trial Location and Clusters.
A map of Indonesia is shown at the top, with the location of Yogyakarta Province shaded in dark blue. The enlarged area at the bottom shows the trial area in Yogyakarta City, which includes a small area of neighboring Bantul District (clusters 23 and 24). Intervention clusters (which received deployments of *Aedes aegypti* mosquitoes infected with the *w*Mel strain of *Wolbachia pipientis*) are shaded in dark blue, and control clusters (which received no deployments) are shaded in light blue. Red crosses indicate the locations of the primary care clinics where enrollment was conducted.

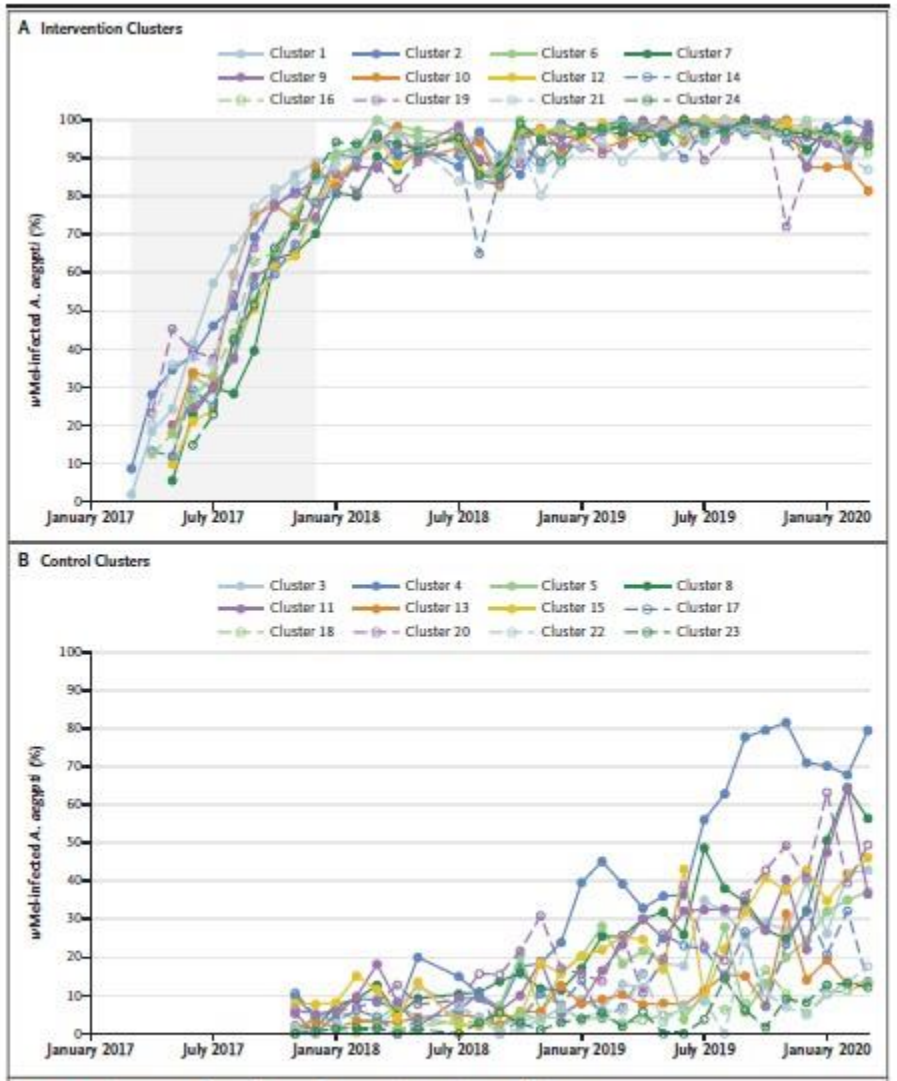


Figure 2. Introgression of *w*Mel into Local *A. aegypti* Mosquito Populations.
The lines in each panel indicate the percentage of *A. aegypti* infected with *w*Mel collected from traps in intervention clusters (Panel A) and control clusters (Panel B) each month from the start of deployments (March 2017) to the end of participant enrollment (March 2020). The shaded area in Panel A indicates the period from the first release in the first intervention cluster (March 2017) to the last release in the last intervention cluster (December 2017). There were 9 to 14 release rounds, with each round lasting 2 weeks, in each intervention cluster.

Reducing contamination

- Cluster size
  - Edge-effects tend to be smaller when cluster sizes are larger as a smaller proportion of individuals live near cluster boundaries
- Separation of clusters
  - Make sure clusters are well separated – useful for reducing the level contact between intervention and control clusters
  - Degree of separation depends on intervention & transmission dynamics (for infectious diseases)

Example: Thibela TB study

- Clusters were defined as a mine shaft or shafts, and associated hostels
- Geographically or structurally separated from other clusters

Thibela TB: Design and methods of a cluster randomised trial of the effect of community-wide isoniazid preventive therapy on tuberculosis amongst gold miners in South Africa

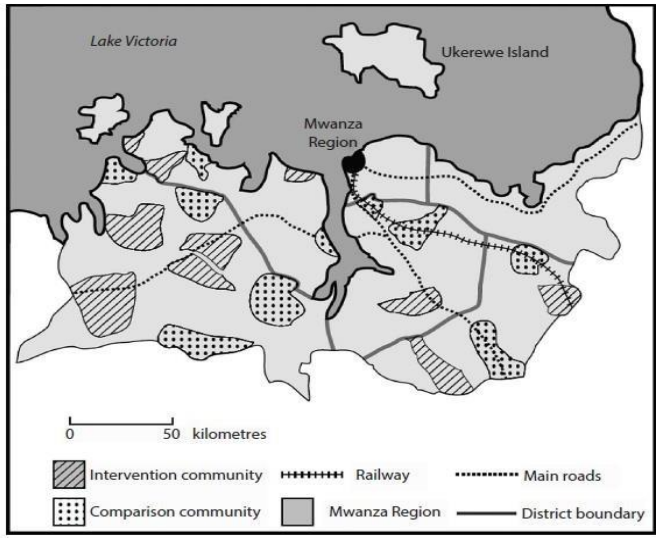
Katherine L. Fielding [a,*], Alison D. Grant [b], Richard J. Hayes [a], Richard E. Chaisson [c], Elizabeth L. Corbett [b], Gavin J. Churchyard [a,d]

Reducing contamination

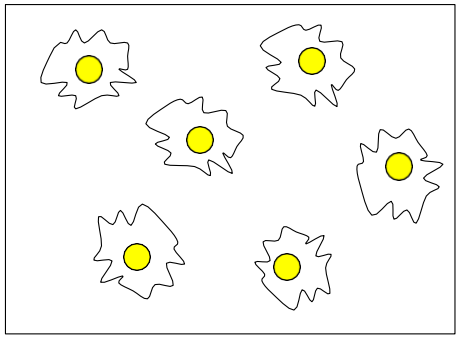
- Buffer zones

Trial clusters separated by buffer zones, so that there is little or no common boundary between any two clusters

Example: MEMA kwa Vijana trial
- Adolescent sexual health intervention
- Randomised 20 rural communities in Tanzania



- 'Fried egg' design



- Intervention or control condition applied to all individuals in each cluster
- Outcomes measured in the central (yellow) part of the cluster - this is surrounded by a buffer zone receiving the same condition as the evaluation sample
- Disadvantage: central areas may not be representative of the whole cluster

MulebaTrial: comparison of Combined use of IRS and ITNs vs ITN alone against Malaria

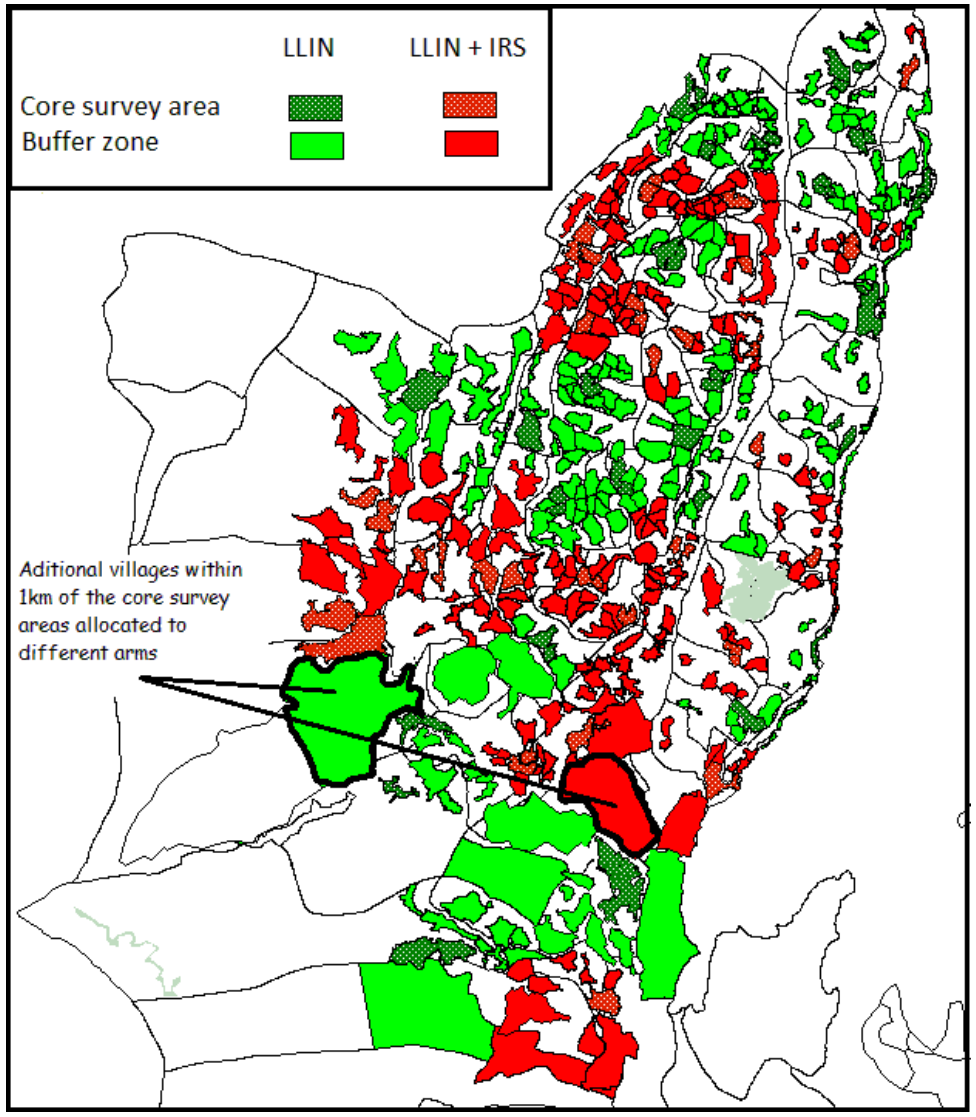
OPEN ACCESS Freely available online

PLOS | MEDICINE

Indoor Residual Spraying in Combination with Insecticide-Treated Nets Compared to Insecticide-Treated Nets Alone for Protection against Malaria: A Cluster Randomised Trial in Tanzania

Philippa A. West[1]*, Natacha Protopopoff[2], Alexandra Wright[2], Zuhura Kivaju[3], Robinson Tigererwa[4], Franklin W. Mosha[5], William Kisinza[3], Mark Rowland[2], Immo Kleinschmidt[6]

1 Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, United Kingdom, 2 Department of Disease Control, London School of Hygiene & Tropical Medicine, London, United Kingdom, 3 National Institute for Medical Research, Amani Medical Research Centre, Muheza, Tanzania, 4 Muleba District Medical Office, Department of Health, Muleba, Tanzania, 5 Kilimanjaro Christian Medical College, Tumaini University, Moshi, Tanzania, 6 Medical Research Council Tropical Epidemiology Group, London School of Hygiene & Tropical Medicine, London, United Kingdom



Design choices for impact evaluation

- Repeated cross sectional surveys
  - A cross-sectional random sample of the population
  - Measures prevalence (risk) as an outcome (not incidence)
  - Usually two surveys: baseline and at follow-up
    - Baseline data can be used for randomisation or cluster-level adjustment
  - Advantages
    - Logistically easier than cohort follow-up
    - If >2 surveys, time-specific effects can be measured
  - Disadvantages
    - Unable to measure incidence
- Cohort
  - A random sample recruited at start followed up over time,
  - Gives direct measure of case incidence
  - Disadvantages: attrition, logistically more complex, incidence measure could be distorted in a cohort of individuals that are surveyed repeatedly
  - Advantage: Incidence often a preferred outcome measure

Content

1. Type of cluster
2. Size of cluster
3. Contamination
4. Design choices for impact evaluation

5. Matching and stratification
6. Restricted (covariate constrained) randomisation
7. Choice of matching, stratification or balance variables
8. Randomisation procedures

Matching and stratification

- Unmatched CRT
  - clusters are randomly allocated to trial arms without restriction
- Matched CRT
  - clusters are split into groups so that within each group one cluster is randomised to each trial arm
  - For two arms the groups are matched pairs
- Stratified CRT
  - clusters are split into groups (strata) so that within groups, more than one cluster may be randomised to each trial arm and all trial arms are represented within each group

Rationale for matching and stratification

- Matching and stratification involves grouping clusters based on their baseline characteristics, so that within each group the clusters share similar baseline characteristics

- Two main reasons:
  - **Minimises baseline imbalance** between treatment arms
    - In CRTs with a small number of clusters, randomisation cannot necessarily be relied upon to ensure baseline balance
    - Matching/stratification can help reduce baseline imbalance

  - **Improves study power and precision**
    - Variation in the outcome between clusters effects precision and power
    - Matching/stratification can reduce the between cluster coefficient of variation (k)

Comparison of the three designs

Degrees of freedom

- Power depends on the test statistic and the critical value of the t-distribution; latter depends on the degrees of freedom (df)
  - df: pair-matched << stratified < unmatched
  - critical value of the t-distribution: pair-matched > stratified >unmatched
- pair-matched design - weighing up reduction in k versus losses in df
  - When there are a few clusters matched designs only give increased power if the matching is highly effective

Design	Formula for df	Example df (16 clusters in total, c=8 clusters/arm)	$t_{0.05}$
Unmatched	2(c – 1)	2 x (8 – 1) = 14	2.14
Stratified into two strata of equal size	2(c – 2)	2 x (8 – 2) = 12	2.18
Pair-matched	c – 1	8 – 1 = 7	2.36

Comparison of the three designs (2)

Drop-out of clusters

- Power and precision may be affected and selection bias may occur
- For pair-matched trial, drop-out will result in the entire matched-pair not contributing to the analysis

Statistical inference

- Use of a matched design imposes a number of limitations on the statistical inference that can be carried out

- Stratification is a less rigid method for reducing baseline imbalance between treatment arms and reducing the variance in the intervention effect, but has less limitations in terms of statistical inference than matching

- Baseline imbalance can impact the validity and credibility of the estimates of the intervention effect, so it is important to ensure balance on appropriate factors

Feature or statistical issue	Design		
	Unmatched	Stratified	Pair-matched
Between-cluster variation	The variance of the estimated treatment effect may be higher than with other designs	Aims to achieve a reduction in the beween-cluster variation but not as substantial as a pair-matched design	Effective matching should result in lower between-cluster variation than unmatched or stratified trials
Adjustment for covariates	Able to adjust for covariates using a range of methods	Able to adjust for covariates using a range of methods	Regression methods do not work well for matched design and alternative methods based on cluster-level summaries are the main option (covered in session 6)
Testing for variation in the intervention effect between pairs/ strata	n/a	Able to assess due to replication of clusters within strata	Unable to assess if the intervention effect varies between matched pairs as there is no replication within matched pairs
Estimating coefficient of variation (k) and intra cluster correlation (ρ)	Able to estimate k and ρ as per Session 3	k or ρ may be estimated in each stratum as for an unmatched study and then a pooled estimate of these is calculated (see session 3 for more detail)	Unable to estimate k or ρ because difference in outcome between two clusters in a matched pair results both from the intervention effect and the between-cluster variability

Restricted randomisation

- Using matching/ stratification to reduce the risk of imbalance between study arms with a large number of variables, is limited
  - May need a *large number of strata on which balance* is required

- Restricted (or covariate-constrained) randomisation is a technique that can *minimise imbalance* between study arms

- Restricted randomisation restricts how clusters can be allocated to study arms
  - **Unrestricted** randomisation allows clusters to be allocated to any study arm irrespective of their characteristics
  - **Restricted** randomisation involves selecting randomly from a subset of the allocations that satisfy certain pre-defined criteria of baseline balance

Basic principles of restricted randomisation

Unrestricted randomisation:

- 2c clusters are allocated in a ratio of 1:1, to one of 2 arms
- The number of unique combination of 2c clusters allocated to 2 arms:

$$^{2c}C_c = \frac{(2c)!}{c!\,c!}$$

(where ! represents the factorial function*)

- Example: 12 clusters are allocated to 2 arms (ratio 1:1)

$$^{12}C_6 = \frac{(2 \times 6)!}{6!\,6!} = \frac{(12)!}{6!\,6!} = \frac{4.79 \times 10^8}{720 \times 720} = 924$$

- There are 924 possible allocations

* 6!=6x5x4x3x2x1

Basic principles of restricted randomisation (2)

- Restricted randomisation can help to ensure an acceptable level of overall balance on important variables between trial arms
- Restricting the list of all possible allocations to a subset which are "acceptable" because they meet the balance criteria
- One allocation is then randomly selected from the subset of acceptable allocations

Restricted randomisation – general steps

- List the variables on which balance is required
- Define a list of proposed balance criteria for these variables, e.g. the maximum difference between study arms of important cluster level variables such as the outcome variable and other variables which may have an effect.
- Either enumerate:
  - full list of possible allocations (if this is not too large) or
  - generate a large number of possible allocations (10-15,000 may be enough)
- Identify the subset of the possible allocations which met the restriction criteria as the list of acceptable allocations

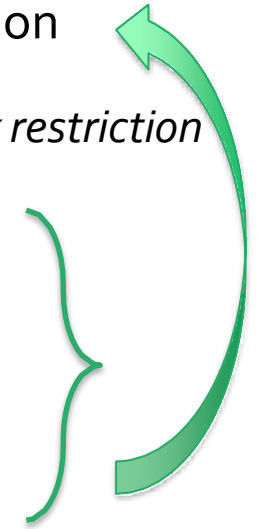
Restricted randomisation – general steps (2)

- Check validity of restricted randomisation:
  - If the list of all allocations is large but the number of allocations after restriction is small (say less than 100 allocations), refine the restriction criteria until the list of acceptable allocations is large enough
  - Also check on how many times each pair of clusters are assigned to the same treatment arm –this should be about 50%. If pairs rarely occur or almost always occur together they are not independently allocated to a study arm. The balance criteria need to be relaxed and the process repeated.
- Once a list of acceptable allocations has been produced that meets the above conditions, select one of the acceptable allocations using simple random sampling.

Restricted randomisation – general steps

- List the variables on which balance is required
- Define a list of proposed balance criteria for these variables, i.e. maximum allowable difference between study arms
- Either enumerate the full list of possible allocations (if this is not too large) or simulate a data series of a large number of possible allocations (10-15,000 may be enough)
- Identify the subset of the possible allocations which met the restriction criteria as the list of acceptable allocations

Refine/relax restriction



- Check validity of the restricted randomisation:
  - If the list of all allocations is large but the number of allocations after restriction is small (say less than 100 allocations) and/or
  - pairs of clusters are always/never (or close to 1) assigned to the same treatment arm

- Select one of the acceptable allocations using simple random sampling

Choice of matching, stratification or balance variables

- Main outcome variable, e.g. existing prevalence of the disease/infection
- Covariates that area relevant to the outcome, e.g. use of bednets, urban/rural, presence of health clinic, etc
- Cluster size
- Socio-economic or logistical criteria

Main outcome variable

- Achieving baseline balance in risk factors for the outcome of interest is important since the intervention effect is estimated on the assumption that everything is the same between study arms except for the intervention
- Most important predictor of outcome is the baseline value of the outcome.This is often used using data from -
  - baseline cross-sectional survey to measure prevalence
  - cohort study during a baseline period prior to randomisation to measure incidence

Covariates

Investigators may also choose to

- balance on risk factors for the event of interest
  - may act as surrogate markers for the outcome of interest as they are closely correlated with the outcome
  - may include:
    - Cluster location and cluster size, or;
    - summary measures of individual-level variables such as socio-economic status, mean age or other risk factors

- balance on social and demographic variables even when they are not associated with the outcome
  - to balance the groups with respect to unknown confounders

Sample size and political/logistical reasons

- Sample size

  - Similar numbers of clusters and individuals across both arms will achieve the highest precision → precision is influenced by variation at both levels

- Political or logistical reasons
  - Achieving a design that is acceptable to stakeholders may be important for the success of the trial
  - Each political unit in a study area receives a ' fair share' of the intervention
  - Costs of trials can be substantial and so for example, limiting the distance field workers have to travel to visit different clusters in the same arm, may help make a trial more logistically feasible

Example: Misungwi Trial evaluating new types of insecticide treated bednets

Effectiveness and cost-effectiveness against malaria of three types of dual-active-ingredient long-lasting insecticidal nets (LLINs) compared with pyrethroid-only LLINs in Tanzania: a four-arm, cluster-randomised trial

Jacklin F Mosha*, Manisha A Kulkarni*, Eliud Lukole, Nancy S Matowo, Catherine Pitt, Louisa A Messenger, Elizabeth Mallya, Mohamed Jumanne, Tatu Aziz, Robert Kaaya, Boniface A Shirima, Gladness Isaya, Monica Taljaard, Jacklin Martin, Ramadhan Hashim, Charles Thickstun, Alphaxard Manjurano, Immo Kleinschmidt, Franklin W Mosha, Mark Rowland, Natacha Protopopoff

Randomisation and masking

An independent statistician conducted constrained randomisation to allocate the 84 clusters to the four study groups at a ratio of 1:1:1:1, ensuring that absolute differences in cluster means between study groups were within the specified ranges for each of the following variables: population size (range ± 9000 people); malaria infection prevalence at baseline (± 5 percentage points); socioeconomic status, measured as the percentage of households in the poorest tercile based on the wealth index of the entire study area (± 10 percentage points); LLIN use, measured as the percentage of residents who reported using a net last night (± 10 percentage points); and cluster suitability (predicted for each cluster using an ecological niche model and based on observed species composition in the study area)[13] for *Anopheles gambiae* sensu stricto (ss) and *Anopheles funestus* (± 15 percentage points). Using Stata (version 11), 200 000 random allocations were generated and tested against the restriction criteria; of the 10 077 acceptable allocations, one was selected at random. The validity of the randomisations was verified by checking the frequency of allocations to the same study group of all pairs of clusters

Randomisation procedures

- For geographical or community clusters with a large cluster size, a public ceremony can provide a platform for carrying out the randomisation of clusters to trial arms

- Three main components:
  - explaining the rationale for the trial,
  - explaining the need for randomisation and the methods that have been employed,
  - carrying out the randomisation with involvement of key stakeholders

Choice of design strategy

- Many cluster randomised trials only employ a small number of clusters
  - Randomising a small number of clusters can lead to imbalance in important baseline covariates between study arms
  - Unmatched CRTs may have substantial between-cluster variability, affecting precision and power

- Balance at baseline between trial arms can be improved, and power and precision can be increased by:
  - Matching of clusters,
  - Stratification of clusters
  - Restricted randomisation

Matching, stratification and restricted randomisation

- Matching
  - *Potential* to reduce imbalance
  - *Potential* to increase the power and precision of a CRT through a reduction in the between-cluster variability in the outcome
  - *However* does lead to a loss in degrees of freedom and for trials with only a few clusters there is a risk that on balance this will result in a loss of power, unless matching is highly effective
- Stratification
  - *Potential* to reduce imbalance
  - *Potential* to increase the power and precision of a CRT through a reduction in the between-cluster variability in the outcome
  - Leads to a loss of *fewer* degrees of freedom than matching consequently, stratification can be a more effective strategy than matching
- Restricted randomisation
  - *Improves* overall balance between trial arms
  - Can lead to an improvement in power and precision but *not as effectively* as other two

Guidelines for the choice of design strategy

- Small number of clusters (10-15 per arm)
  - Recommend stratification to increase power and precision
  - Restrict the number of strata to avoid loss of too many degrees of freedom
  - If balance is needed on a large number of variables then use a combination of stratification and restricted randomisation

- Large number of clusters
  - Tend not to have a problem with imbalance (randomisation itself will bring about balance)
  - Matching or stratification may still help with power and precision
  - Large number of clusters means that loss of degrees of freedom is less of an issue
  - Stratification with 3-4 strata will capture most of the additional power achieved by matching
    - therefore recommend stratification

Sample size for Cluster Randomised Trials

- In all trials, responses (outcomes) vary between individual participants
- In individually randomised trials, we randomise a sufficient number of participants to accurately measure the *average* outcome, and determine whether this varies between participants in different trial arms
- In a CRT, responses and outcomes also vary **between clusters**
- Even if each cluster contains a large number of individuals, the trial may be insufficiently powered if there are too few clusters
- i.e. need to randomise sufficient number of clusters to obtain sufficiently accurate estimates of intervention effects

Sample Size Calculations for Unmatched Studies

- We will consider sample size calculations for
  - means (μ), for quantitative outcomes
  - event rates (λ),
  - proportions (π), for binary outcomes
- In a CRT, the sample size calculation needs to take account of the between-cluster variability. There are two main measures
  - the coefficient of variation, k
  - the intra-cluster correlation coefficient (ICC), or ρ *(not defined for event rates)*

Sample Size Calculations:

For an *individually randomised* trial the sample size

$$n = \left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{(\sigma_0^2 + \sigma_1^2)}{(\mu_0 - \mu_1)^2}$$

- n is the number of participants required in each arm,
- μ_o and μ_1 are the means in the control and intervention arms,
- σ_o and σ_1 are the standard deviations of the outcome in each arm
- α is the significance level, or type 1 error rate, usually set to 5%
- $z_{\alpha/2}$ is the standard normal distribution value corresponding to upper tail probabilities of $\alpha/2$; For α= 5 % → $z_{\alpha/2}$=1.96
- β is the type 2 error rate, usually set at 10% or 20% . The Power of the study to show a significant effect of the intervention is 100(1–β)%.
- z_{β} is the standard normal distribution value corresponding to upper tail probabilities of β;
  for Power=80%, β =20% →z_{β}=0.84

Sample Size Calculations:

- For an equal number of individuals in each cluster (m) and coefficients of variation for each trial arm $k_0$and k_1, the number of clusters per arm (c) is given by:

$$c = 1 + \left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{(\sigma_0^2 + \sigma_1^2)/m + (k_0^2\mu_0^2 + k_1^2\mu_1^2)}{(\mu_0 - \mu_1)^2}$$

- where μ_o and μ_1 are the means in the control and intervention arms
- σ_o and σ_1 are the **within-cluster** standard deviations of the outcome in each arm.
- Assuming that the intervention has a similar effect in all the clusters then the coefficient of variation for the two trial arms will be similar (k) and

$$c = 1 + \left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{(\sigma_0^2 + \sigma_1^2)/m + k^2(\mu_0^2 + \mu_1^2)}{(\mu_0 - \mu_1)^2}$$

Note: m is the number of individuals where outcome will be mesured ad not individuals who will receive intervention

(2) Event rates

- Sample sizes for event rate data are the amount of person-time required in each arm of a trial
  - Sample size for an ***individually randomised*** trial with two trial arms

$$y = \left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{(\lambda_0 + \lambda_1)}{(\lambda_0 - \lambda_1)^2}$$

  - y is the number of person-years in each arm
  - λ_0 and λ_1 represent the rates in the control and intervention arms
- In a ***CRT*** this is modified to account for the between cluster variability to

$$c = 1 + \left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{(\lambda_0 + \lambda_1)/y + (k_0{}^2\lambda_0{}^2 + k_1{}^2\lambda_1{}^2)}{(\lambda_0 - \lambda_1)^2}$$

- Which simplifies to

$$c = 1 + \left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{(\lambda_0 + \lambda_1)/y + k^2(\lambda_0{}^2 + \lambda_1{}^2)}{(\lambda_0 - \lambda_1)^2}$$

Note: ***y*** is now the person years available in each cluster

(3) Proportions (binary outcomes)

- Compare the proportions of participants experiencing the outcome in the intervention arm, π_1, with proportion in the control arm, π_0. Sample size per arm for an individually randomised trial is given by

$$n = \left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{\pi_0\left(1-\pi_0\right)+\pi_1\left(1-\pi_1\right)}{\left(\pi_0-\pi_1\right)^2}$$

- For CRT, the number of clusters required per arm, given *m* participants per cluster is

$$c = 1 + \left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{\pi_0(1-\pi_0)/m + \pi_1(1-\pi_1)/m + (k_0^2\pi_0^2 + k_1^2\pi_1^2)}{(\pi_0-\pi_1)^2}$$

- Simplifies to

$$c = 1 + \left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{\pi_0(1-\pi_0)/m + \pi_1(1-\pi_1)/m + k^2(\pi_0^2 + \pi_1^2)}{(\pi_0-\pi_1)^2}$$

If c is bigger and not feasible to run trial, rather increase minimum difference and also try various values of k if you want to keep power at the same level.

Example –Kilifi bednet trial

- CRT of effect of insecticide-treated bednets (ITNs) on malaria and hence child mortality [1]
- Primary objective was to measure the impact of ITNs on all-cause mortality among children aged 1-59 months
- Study area was divided into clusters of approx. 1000 people – included approx. 200 children aged 1-59 months in each cluster
- The intervention was randomised to clusters
- Deaths of children recorded over a two year follow up period
- Over the study duration there would be 424 child-years follow up per cluster
- Data available for the two years prior to the study showed that the mortality rate (λ_0) = 14.8 per 1000 person-years

Baseline data is critical in CRT studies, especially if there isn't much background data

[1] Nevill et al. 1996

Example continued...

- Trial was to have 80% power to show reduction in mortality rate in the intervention arm (λ_1) of at least 30% lower compared to the control arm, i.e. $\lambda_1 = 0.0148 \times 0.7 = 0.0104$ or 10.4 per 1000 person-years
- the coefficient of variation (k) was estimated as 0.29 and was assumed to be equal in the two trial arms
- Therefore using the equation

$$c = 1 + \left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{(\lambda_0 + \lambda_1)/y + k^2(\lambda_0{}^2 + \lambda_1{}^2)}{(\lambda_0 - \lambda_1)^2}$$

- For 80% power, $\beta = 20\%$; hence with 5% significance (α) we get

$$c = 1 + (1.96 + 0.84)^2 \frac{(0.0148 + 0.0104)/424 + 0.29^2(0.0148^2 + 0.0104^2)}{(0.0148 - 0.0104)^2}$$

$$= 36.2$$

- Therefore we would need 37 clusters per arm , i.e. 74 in total

The Design Effect (DEff)

Recall, from lecture 1

Design effect (Deff) = Var_{CRT} / Var_{RCT} = $\frac{\text{Sample size for CRT}}{\text{Sample size for RCT}}$

Therefore, Sample size for CRT= (Sample size for RCT) X Deff

Design effect can also be expressed as

$$DEff = 1 + (m - 1)\,\rho$$

Where ρ is the intra-cluster coefficient of variation

and m is the cluster size

Sample size calculations based on the intra-cluster correlation coefficient, ICC or ρ

Using

Sample size for CRT= (Sample size for RCT) X Deff

and

$$Deff = 1 + (m - 1)\,\rho$$

- We can express the sample size for CRT in terms of ρ

- For means we get $c = 1 + \left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{[\sigma_0^2+\sigma_1^2]\times[1+(m-1)\rho]}{m(\mu_0-\mu_1)^2}$

- For proportions we get $c = 1 + \left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{[\pi_0(1-\pi_0)+\pi_1(1-\pi_1)][1+(m-1)\rho]}{m(\pi_0-\pi_1)^2}$

- The intra-cluster correlation coefficient is not defined

  for event rates

Variable cluster size

- So far we have assumed a fixed number of participants per cluster
- If the number of participants in each cluster varies we need to make a modification to the sample size calculation. In the formula for sample size we use in place of cluster size m, the **harmonic mean of cluster size**
  - The harmonic mean of cluster sizes is defined as $\bar{m}_H = \frac{1}{\sum(1/m_j)/c}$
  - For event rates; $\bar{y}_H = \frac{c}{\sum(1/y_j)}$
  - The effect of using the harmonic mean instead of the arithmetic mean is to increase the number of clusters needed

Sample Size Calculations for Matched and Stratified Studies

- If a matched or stratified design is chosen, a simple adjustment is made to the sample size formulae.
  - Two extra clusters are added, instead of one as in unmatched designs to account for loss of degrees of freedom.
  - A matched coefficient of variation k_m is used instead of k. This is a coefficient of variation in true rates between clusters **within matched pairs**
  - For stratified trials k will generally lie between k_m for matched trials and the k for the unmatched design

$$c = 2 + (z_{\alpha/2} + z_{\beta})^2 \frac{(\sigma_0^2 + \sigma_1^2)/m + k_m^2(\mu_0^2 + \mu_1^2)}{(\mu_0 - \mu_1)^2}$$

$$c = 2 + (z_{\alpha/2} + z_{\beta})^2 \frac{(\lambda_0 + \lambda_1)/y + k_m^2({\lambda_0}^2 + {\lambda_1}^2)}{(\lambda_0 - \lambda_1)^2}$$

$$c = 2 + (z_{\alpha/2} + z_{\beta})^2 \frac{\pi_0(1 - \pi_0)/m + \pi_1(1 - \pi_1)/m + k_m^2(\pi_0^2 + \pi_1^2)}{(\pi_0 - \pi_1)^2}$$

Example: The Well London program

- The Well London project is a set of interventions focussed on healthy eating, exercise, mental health and wellbeing
- A CRT was conducted to evaluate effectiveness of the program at improving physical activity, diet and mental health
- Clusters were geographical areas comprising between **1,000 and 1,500 residents**
- **Two clusters** were **identified in each London borough** forming a matched pair, randomised to opposite study arms
- Primary outcome: proportion of adults who ate **five or more pieces of fruit and vegetables a day** assessed by a survey at baseline and after follow up

Example continued

- A national health survey indicated that on average 27% of adults ate 5 portions a day ($\pi_0 = 0.27$)
- **No baseline data** were available to estimate k_m
- Hence a range of sample sizes were calculated for **plausible estimates of k_m and difference in the outcome between arms**.
- One plausible estimate was k_m **= 0.10** SD is 10% of the mean
- This would imply that, within matched pairs, the **cluster-level proportion who ate 5 portions a day** in the absence of intervention would **vary from 80% to 120%** of an average stratum-specific value (1 ± 2 x 0.10)
- Planned to survey **100 adults per cluster** at the follow-up survey (m)
- Assumed that intervention would **increase proportion eating at least 5 portions a day by 50%.**
- Therefore π_1 = 1.5X0.27=0.405

Example continued

- For **80% power** and a significance level of 5%

$$c = 2 + \left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{\pi_0(1-\pi_0)/m + \pi_1(1-\pi_1)/m + k_m^2(\pi_0^2 + \pi_1^2)}{(\pi_0 - \pi_1)^2}$$

$$= 2 + (1.96 + 0.84)^2 \frac{0.27(1-0.27)/100 + 0.405(1-0.405)/100 + 0.10^2(0.27^2 + 0.405^2)}{(0.27 - 0.405)^2}$$

$$= 4.9$$

- would therefore require a minimum of 5 matched pairs

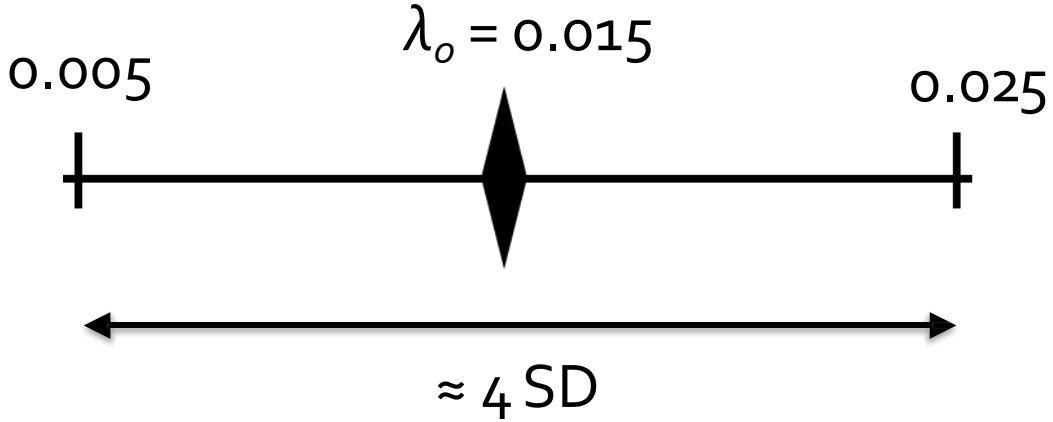
Estimating the Between-cluster Coefficient of Variation in the outcome: unmatched studies

Outcome	Coefficient of variation in i^{th} trial arm k_i
Event rates	$\frac{\sigma_{Bi}}{\lambda_i}$
Proportions	$\frac{\sigma_{Bi}}{\pi_i}$
Means	$\frac{\sigma_{Bi}}{\mu_i}$

- In addition to estimating λ_i, π_i or μ_i also need to estimates the between-cluster standard deviation σ_{Bi}
- In the lack of empirical evidence we can use informed judgement for assumptions regarding k

Example – returning to the Kilifi bednet trial

- From mortality rates in a similar area, average mortality ≈ 15/1000 person-years in the control arm (λ_o = 0.015)
- Mortality was thought to vary between 5/1000 person-years (0.005 per person year) and 25/1000 person-years (0.025 per person year) in different clusters.
- Remembering that *95%* of data *lie between* ± 2 SD, we can assume that roughly 95% of cluster-level mortality rates lie between 0.005 and 0.025
- Since 0.025-0.005 = 0.02, one SD = 0.02/4 = 0.005; σ_{Bi} =0.005
- $k = \sigma_{Bi} / \lambda_o$ = 0.005/0.015 = 0.33



Calculating k with prior data on between-cluster variation

- A more accurate estimate of σ_B can be obtained from the empirical variance s^2 of cluster-specific results, e.g. from baseline data
- Recall from lecture 1:

  total variation = between individual variation + Between cluster variation

  →Between cluster variation = total variation – between individual variation

- Therefore, for rates: $\hat{\sigma}_B^2 = s^2 - \frac{r}{\bar{y}_H}$ where r is the overall rate computed from all clusters combined.

	Estimate of the true between-cluster variation ($\hat{\sigma}_B^2$)	Estimate of the coefficient of variation ($\hat{k}$)
Event rates	$s^2 - \frac{r}{\bar{y}_H}$	$\hat{\sigma}_B / r$
Proportions	$s^2 - \frac{p(1-p)}{\bar{m}_H}$	$\hat{\sigma}_B / p$
Means	$s^2 - \frac{\hat{\sigma}^2}{\bar{m}_H}$	$\hat{\sigma}_B / \bar{x}$

$$s^2 = \sum (r_j - \bar{r})^2 / (c-1)$$

Example: Kilifi trial of insecticide-treated bednets

- Data were available for the overall child mortality rate: there were 321 deaths over 21,646 person-years of observation, giving overall child mortality rate r= 14.8 per 1000 person years (0.0148)
- The empirical standard deviation (s) of the observed mortality rates across clusters was *s* = 0.00758, and the harmonic mean of the person-years ($\bar{y}_H$) per cluster was 379.

$$\hat{\sigma}_B^2 = 0.00758^2 - \frac{0.0148}{379} = 1.84 \times 10^{-5}$$

$$\sigma_B = \sqrt{1.84 \times 10^{-5}} = 4.29 \times 10^{-3}$$

$$\hat{k} = \hat{\sigma}_B / r$$

$$\hat{k} = \frac{4.29 \times 10^{-3}}{0.0148} = 0.29$$

Estimating the coefficient of variation in matched and stratified trials

- For matched designs, sample size calculation requires **an estimate of k_m which represents the within-pair coefficient of variation in the outcome** between clusters in the absence of intervention
- For stratified designs sample size calculation requires an **estimate of k_m which represents the within-stratum coefficient of variation** in the outcome between clusters in the absence of intervention.
- As in unmatched trials, often no data are available on k_m so range of plausible values used to estimate the required sample size.
- If empirical data are available on the endpoint of interest from clusters grouped according to the matching or stratification variables, this can be used to estimate k_m

Estimating the coefficient of variation in matched and stratified trials

	Estimate of the between-cluster variance in the s^{th} stratum (or matched pair), σ_{Bs}^2.	Estimate of the matched coefficient of variation in the s^{th} stratum (or matched pair), k_{ms}.
Event Rates	$\hat{\sigma}_{Bs}^2 = s_s^2 - \frac{r_s}{\bar{y}_{Hs}}$	$\hat{k}_{ms} = \hat{\sigma}_{Bs}/r_s$
Proportions	$\hat{\sigma}_{Bs}^2 = s_s^2 - \frac{p_s(1-p_s)}{\bar{m}_{Hs}}$	$\hat{k}_{ms} = \hat{\sigma}_{Bs}/p_s$
Means	$\hat{\sigma}_{Bs}^2 = s_s^2 - \frac{\hat{\sigma}_s^2}{\bar{m}_{Hs}}$	$\hat{k}_{ms} = \hat{\sigma}_{Bs}/\bar{x}_s$

Assuming that the within-stratum coefficient of variation, k_{ms} is roughly constant across strata, a simple estimate of the value of k_m can be obtained as:

$$\hat{k}_m = \sum k_{ms}/S$$

If the number of clusters varies substantially across the strata, it may be preferable to use a **weighted average**, weighting by the number of clusters in a stratum.

Example: The Well London program

- Data from the baseline survey showed that there was **considerable variation of the primary outcome** (eating at least 5 portions of fruit and veg every day):
- assuming **no matching** the coefficient of variation (*k*) was **0.20,** and this r**educed to 0.14** using the matched design (k_m).
- This is considerably **higher than the estimate for k_m of 0.10** that was used in the initial power calculations
- Furthermore, the baseline survey indicated the **proportion of adults in the control arm eating 5 portions a day** would be higher than the initial estimate of 27% at **37%**
- Consequently, having randomised **20 matched pairs** they had **90% power** to detect a small, yet important **22% increase** in this primary outcome (retaining 100 participants per cluster as before)

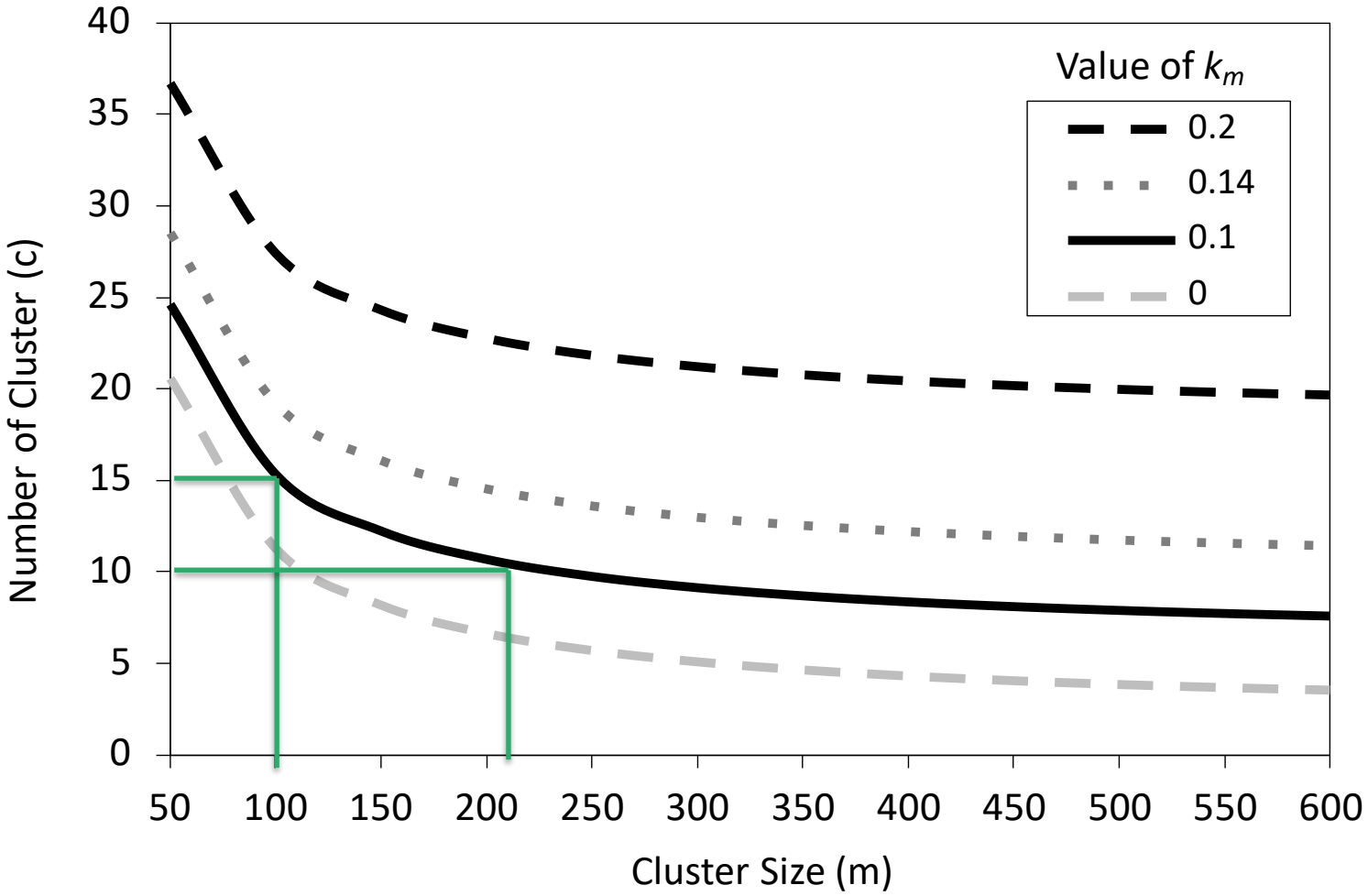
Cluster Size versus number of clusters

- In some situations, may have choice over both size of each cluster as well as the number of clusters.
- If the outcome is measured in a random sub-sample of each cluster, we may alter the size of this sub-sample.
- For example, in the Well London trial the average size of each cluster was 1,000 -1,500 residents, hence potential for sub-sampling larger number in each cluster. Would this reduce the number of clusters required?
- Statistical efficiency of a CRT is increased if we employ a larger number of smaller clusters rather than a few larger clusters

Example: The Well London Program

- **In initial sample size calculations**, assumed that **outcome** (the proportion eating 5 portions of fruit and veg a day) would be **27% in the control arm versus 33% in the intervention** arm.
- Assuming **100 members of each cluster would be surveyed (*m*) and a k_m of 0.10, it was estimated that 16** matched pairs would be required
- Alternatively, **number of clusters could be decreased to 11 matched pairs if the sample surveyed in each cluster increased to approx. 210**
- For **larger values of k**, there is **decreasing scope for reducing the numbers of clusters** by increasing the sample size within each cluster

Number of clusters required for a CRT



Note: for matched pair designs (as in example on previous slide), add one to number of clusters to account for loss of degree of freedom

Analysis of CRTs -
Basic principles and cluster-level analysis

Content

1. Basic principles of analysis
2. Baseline data and analysis
3. Estimates of effect
4. Cluster level analysis
5. Adjusting for covariates
6. Non-parametric tests

Basic principles of analysis

- Want to make valid inferences about *effect measures* when outcomes are compared between study arms
- Need to use methods that take account of *within-cluster correlation*, otherwise SE will be too small and CI too narrow
- Two main analytical approaches
  - Cluster-level analysis
  - Individual-level analysis
- This session focuses on *cluster-level analysis* and presents a method which has proven *robust* over a wide range of situations

Experimental and observational units

- In a CRT the experimental unit is the *cluster*
- Observational units usually include *individuals*, but observations are sometimes made at several levels
  - E.g. health district, GP practice, GP, patient

Basic principles of analysis

- Need to define *observational units*, *primary outcome* and *effect measure* of interest

Statistical model for parameter of interest

$$\theta_{ij} = \alpha + \beta_i + u_{ij}$$

Where θ_{ij} = expected outcome in observational units in j^{th} cluster in i^{th} study arm

(Intervention: i = 1; Control: i = 0)

α = expected outcome in control arm

β_i = effect of intervention ($\beta_0 = 0$)

u_{ij} = cluster-level *random effect* for j^{th} cluster in i^{th} study arm

The u_{ij} are assumed to have mean zero and variance σ_B^2

Basic principles of analysis

Example: Model for rates

Number of events d_{ij} in j^{th} cluster in i^{th} study arm is Poisson with mean

$$E(d_{ij}) = \lambda_{ij} \times y_{ij}$$

where λ_{ij} is the true rate of the outcome event in that cluster, and y_{ij} is the person-years of observation in that cluster

Model for rate difference

$$\lambda_{ij} = \alpha + \beta_i + u_{ij}$$

Mean rate in intervention arm = $\alpha + \beta_1$
Mean rate in control arm = α
so β_1 = rate difference

Model for rate ratio

$$\log(\lambda_{ij}) = \alpha + \beta_i + u_{ij} \qquad \text{and } \beta_1 = \log(\text{rate ratio})$$

Note: See lecture notes for equivalent models for means and proportions

Baseline data and analysis

Reasons for collecting and analysing baseline data

- To characterise the study population (to assess generalisability)
- To inform study conduct
- To present baseline data by study arm and identify any important *imbalances* between study arms
- To allow an *adjusted analysis* to be carried out in order to:
  - Adjust for any imbalances between study arms
  - Reduce between-cluster variability and hence improve precision and power
- Decisions on which baseline variables to adjust for should depend on the degree of imbalance between arms and their likely correlation with outcome – *not on p-values for the difference between arms!*

Statistical analysis plans (SAPs)

- Should be finalised either at start of study or *before*

study unblinding and should identify any adjusted

analysis to be performed

Estimates of effect

Point estimates of effect measure

- Can be obtained using either cluster summaries *or* individual level data
- Will illustrate using data from *Smoke free generation* trial

Example

	Intervention		Control	
	0 / 42	0	5 / 103	0.049
	1 / 84	0.012	3 / 174	0.017
	9 / 149	0.060	6 / 83	0.072
	11 / 136	0.081	6 / 75	0.080
	4 / 58	0.069	2 / 152	0.013
	1 / 55	0.018	7 / 102	0.069
	10 / 219	0.046	7 / 104	0.067
	4 / 160	0.025	3 / 74	0.041
	2 / 63	0.032	1 / 55	0.018
	5 / 85	0.059	23 / 255	0.102
	1 / 96	0.010	16 / 125	0.128
	10 / 194	0.052	12 / 207	0.058
Overall:	58 / 1341	0.043	91 / 1479	0.062
Mean:		0.039		0.060
SD:		0.026		0.035

$RR_O = 0.043 / 0.062 = 0.70$ $RR_M = 0.039 / 0.060 = 0.65$

Estimates of effect

- RR_O and RR_M are usually similar
- Which should be used?

Overall response (RR_O)

Individual level analyse: depends on whether you took betwee/within clustering into account.

- Ease of estimation
- Equal weight given to each *individual*
- Consistent estimator of population rates if a random cluster sample

Mean of cluster responses (RR_M)

- Better tie-in with tests (eg. t-test)
- Equal weight given to each *cluster*
- Do not always have a random cluster sample

Cluster level analysis

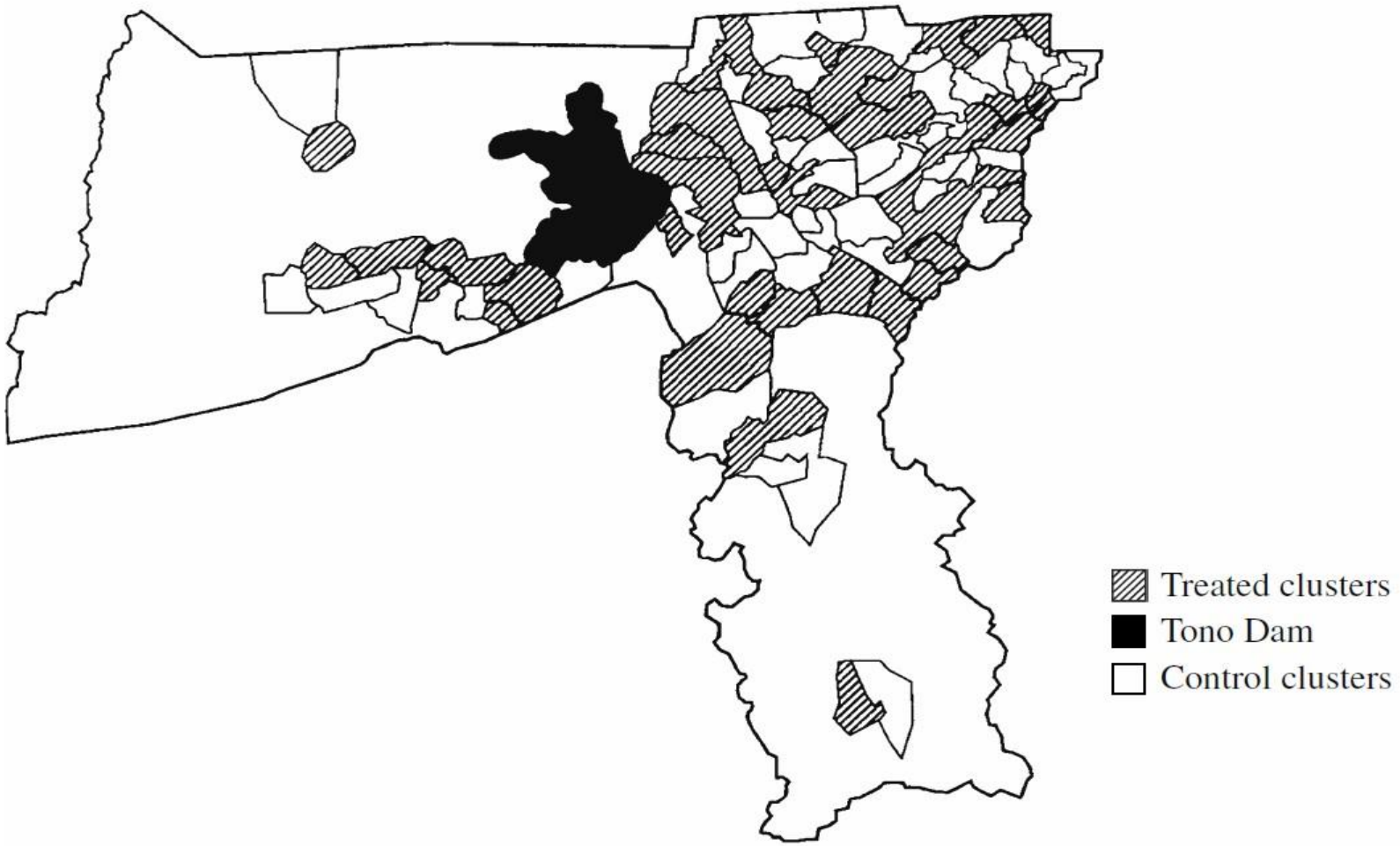
Cluster level analysis: the basic approach

- Calculate *summary measure* of outcome of interest for each cluster
- Compare these summary measures between study arms using the *unpaired t-test*
- Robustness – provides valid estimates even when few clusters

Example

- Will illustrate using data from the Ghana bednet trial
- 96 clusters in Northern Ghana
- Outcome: all-cause mortality in children aged 6-59 months

Some case studies



Cluster level analysis

Estimated Child Mortality Rates by Treatment Arm and Estimated Intervention Effects in the Ghana Bednet Trial

	Intervention Arm	Control Arm	Effect Estimates
Number of clusters	48	48	
Total deaths	396	461	
Total person-years	16841.1	16494.8	
Analyses Based on Individual-Level Data			
Overall rate (/1000 person-years)	23.51	27.95	
Rate difference (/1000 person-years)			−4.44
Rate ratio			0.841
Analyses Based on Cluster Summaries			
Mean of cluster rates	23.97	27.92	
SD of cluster rates	9.73	12.75	
Rate difference (/1000 person-years)			−3.95
Rate ratio			0.859

Cluster level analysis

t-test on cluster-level rates

$$t = \frac{\bar{r}_1 - \bar{r}_0}{s\sqrt{\frac{1}{c_1} + \frac{1}{c_0}}}$$

where $\bar{r}_1$ is the mean rate in the *Intervention* arm and $\bar{r}_0$ is the mean rate in the *Control* arm, c_1 *and* c_0 are the numbers of clusters in the two arms, and the pooled estimate of the standard deviation is

$$s^2 = \frac{\sum_{ij}(r_{ij} - \bar{r}_i)^2}{c_1 + c_0 - 2}$$

- The computed value of *t* is referred to percentage points of the *t* distribution with ($c_1 + c_0 - 2$) degrees of freedom
- This gives an estimate of the *rate difference* with 95% confidence interval

Cluster level analysis

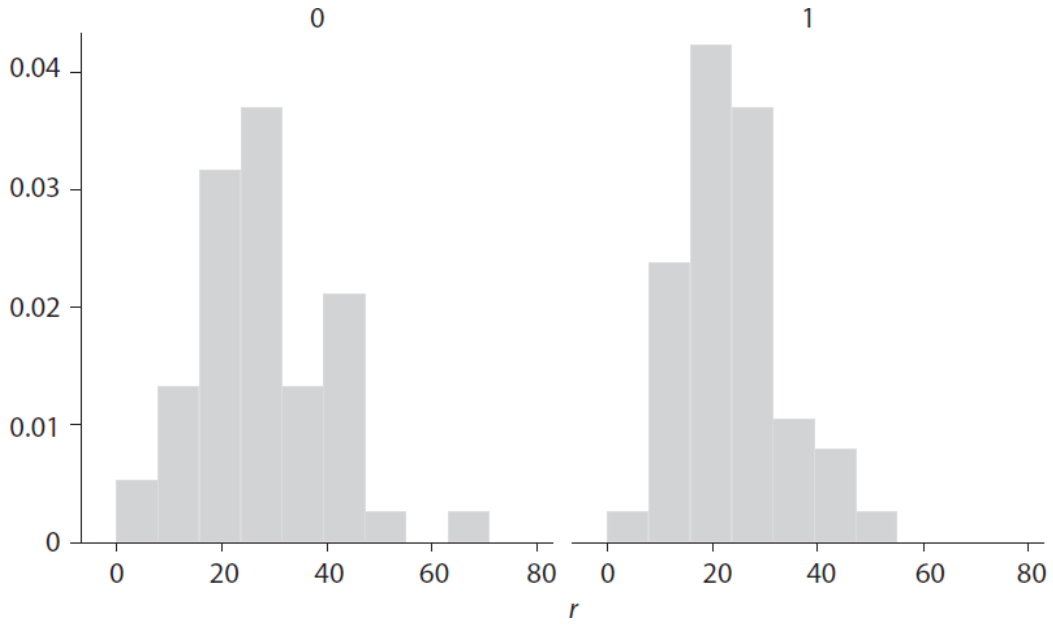
```
. use ghana_bednet.dta

. collapse (mean) bednet (sum)outcome follyr, b

. gen rate=(outcome/follyr)*1000

. ttest rate, by(bednet)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      48    27.92242    1.840102     12.7486    24.22061    31.62423
       1 |      48    23.97155    1.404359    9.729682    21.14635    26.79676
---------+--------------------------------------------------------------------
combined |      96    25.94699    1.168985    11.45367    23.62626    28.26772
---------+--------------------------------------------------------------------
    diff |            3.950866    2.314778               -.6451805    8.546912
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   1.7068
Ho: diff = 0                     degrees of freedom =       94

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9544         Pr(|T| > |t|) = 0.0912          Pr(T > t) = 0.0456
```

Nonhlanhla.yende
2023-06-27 15:15:55
---
bnc is cluster

Cluster level analysis

Using the log transformation

- Rates and risks are often positively skewed



- Taking logs helps to reduce skewness and also allows us to make inferences about the *rate ratio* or *risk ratio* which is often the effect measure of interest

Cluster level analysis

Using the log transformation

Instead of the r_{ij} we work with

$$l_{ij} = log(r_{ij})$$

When we take the mean of the l_{ij} over all the clusters in each study arm, we obtain

$$\bar{l}_i = \frac{\sum l_{ij}}{c_i} = \log(\bar{r}_{Gi})$$

Where $\bar{r}_{Gi}$ is the *geometric mean of the rates in the* ith study arm

It follows that

log RR

$$\bar{l}_1 - \bar{l}_0 = \log(\bar{r}_{G1}) - \log(\bar{r}_{G0}) = log\left(\frac{\bar{r}_{G1}}{\bar{r}_{G0}}\right)$$

Cluster level analysis

```
. gen l=log(r)
(1 missing value generated)

. ttest l, by(bednet)
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      47   -3.658178    .0713969    .4894724   -3.801893   -3.514464
       1 |      48   -3.817606    .0638355    .4422655   -3.946026   -3.689185
         +
combined |      95   -3.738731    .0482824     .470599   -3.834597   -3.642865
---------+--------------------------------------------------------------------
    diff |            .1594277    .0956702               -.0305544    .3494097
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   1.6664
Ho: diff = 0                     degrees of freedom =       93

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9505         Pr(|T| > |t|) = 0.0990          Pr(T > t) = 0.0495
```

Cluster level analysis

From the *Stata* output we obtain an estimate of the *log RR* as (-0.1594) with a 95% CI of (-0.3494, .0306)

Taking exponentials we obtain an estimate of the RR as:

exp(-0.1594) = 0.85

with a 95%CI of (0.71, 1.03)

Note: In this example there is one cluster with zero deaths. One option in this case is to add 0.5 to the events in each cluster before dividing by the person-years to obtain the rate

An alternative formula to obtain a CI for the RR is given in the lecture notes

Adjusting for covariates

Basic ideas

- We may wish to adjust for covariates
  - To adjust for imbalances in baseline variables between arms
  - To reduce between-cluster variation by adjusting for variables that are important risk factors for the outcome
- If we just want to adjust for cluster-level variables, we can do this by carrying out *linear regression* on the outcome of interest, with study arm and the adjustment variable(s) as factors in the model
- If we want to adjust for individual-level variables, we can use a *two stage* procedure
  - **Stage 1: Obtain *covariate-adjusted residuals***
  - **State 2: Use these residuals to obtain effect estimates**

    and/or to carry out a *t-test*

Adjusting for covariates

Stage 1: Obtaining covariate-adjusted residuals

Rates

- Fit a *Poisson regression* model to the data including the adjustment variables but *not* including the study arm. (Do not adjust for clustering – this will be taken care of at *Stage 2*)

$$\log(\lambda_{ijk}) = \alpha + \sum_{l} \gamma_l z_{ijkl}$$

- Use the *fitted model* to determine the *expected* number of events e_{ij} in each cluster, assuming no intervention effect

$$e_{ij} = \sum_{k} y_{ijk} \hat{\lambda}_{ijk}$$

- We then go on to compute *residuals* by comparing the observed and expected numbers of events

Adjusting for covariates

Computing the residuals

- If we want to estimate a *rate difference*, we obtain the *difference-residual* for each cluster

$$R_{dij} = \frac{d_{ij} - e_{ij}}{y_{ij}}$$

- If we want to estimate a *rate ratio*, we obtain the *ratio-residual* for each cluster

$$R_{rij} = \frac{d_{ij}}{e_{ij}}$$

Note: for outcomes that are proportions or means, we use logistic regression or linear regression instead of Poisson regression

See lecture notes for full details and definitions of

difference-residuals and *ratio-residuals*

Adjusting for covariates

Stage 2: Using the covariate-adjusted residuals

- If the intervention has no effect, the residuals should be similar in the two study arms
- So Stage 2 involves comparison of the residuals between study arms as follows, depending on what effect measure we want (ratio or difference) ratio

Rate ratio or risk ratio

Adjusted rate ratio or risk ratio $= \frac{\bar{R}_{r1}}{\bar{R}_{r0}}$

Rate difference, risk difference, mean difference

Adjusted effect $= \bar{R}_{d1} - \bar{R}_{d0}$

- The *residuals* can also be compared using the *t-test* as before

Adjusting for covariates

```
. poisson outcome i.agegp i.sex, exp(follyr) irr

Poisson regression                                Number of obs     =     26,342
                                                  LR chi2(5)        =     364.61
                                                  Prob > chi2       =     0.0000
Log likelihood = -4709.4989                       Pseudo R2         =     0.0373

------------------------------------------------------------------------------
     outcome |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       agegp |
          1  |   .3958551   .0382054    -9.60   0.000     .3276301    .4782872
          2  |   .2988595    .030804   -11.72   0.000     .2441924    .3657649
          3  |   .1713254   .0255614   -11.82   0.000     .1278863    .2295195
          4  |   .1823218   .0415171    -7.47   0.000     .1166833    .2848843
             |
       1.sex |   .9417309   .0643878    -0.88   0.380     .8236234    1.076775
       _cons |   .0511386   .0027729   -54.83   0.000     .0459826    .0568727
  ln(follyr) |          1  (exposure)
------------------------------------------------------------------------------
```

```
.predict fv
.collapse (mean) bednet (sum) fv outcome follyr, by(bnc)
```

Adjusting for covariates

```
. gen residr=outcome/fv

. gen residd=(outcome-fv)/follyr

. ttest residr, by(bednet)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      48    1.090595    .0710939    .4925527    .9475728    1.233618
       1 |      48    .9203973    .0502305    .3480069    .8193467    1.021448
---------+--------------------------------------------------------------------
combined |      96    1.005496    .0441661    .4327381    .9178154    1.093177
---------+--------------------------------------------------------------------
    diff |            .1701979    .0870485               -.0026389    .3430346
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   1.9552
Ho: diff = 0                     degrees of freedom =       94

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9732         Pr(|T| > |t|) = 0.0535          Pr(T > t) = 0.0268
```

Adjusting for covariates

Summary of results of Ghana bednet trial

	Intervention arm	Control arm
Clusters	48	48
Total deaths	396	461
Total person-years	16841.1	16494.8
Mean of cluster rates /1000py	23.97	27.92
Unadjusted RR	0.859 (0.721, 1.023)	1 p = 0.091
Adjusted RR	0.844 (0.713, 0.999)	1 p = 0.054

Non-parametric tests

- The t-test for comparison of rates, proportions or means between study arms has been shown to be robust to departures from assumptions – e.g. that the distribution of the cluster rates, proportions or means is normal
- However when the number of clusters is very small or where there are particular concerns about the normality assumption, it may be appropriate to also carry out a non-parametric test that does not make such distributional assumptions
- The Wilcoxon rank sum test or permutation test can be used for this purpose

Non-parametric tests

Wilcoxon rank sum test

- Cluster summaries are collectively ranked from 1 (smallest) to $c_1 + c_0$ (largest)
- Mean ranks are compared between study arms. For Ghana bednet data:

```
. ranksum r, by(bednet)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

      bednet |      obs    rank sum    expected
-------------+---------------------------------
           0 |       48        2580        2328
           1 |       48        2076        2328
-------------+---------------------------------
    combined |       96        4656        4656

unadjusted variance    18624.00
adjustment for ties        0.00
                     ----------
adjusted variance      18624.00

Ho: r(bednet==0) = r(bednet==1)
               z =   1.847
    Prob > |z| =   0.0648
```

- The rank sum test can also be applied to *residuals* in a two stage analysis

Non-parametric tests

Permutation test

- A more powerful alternative to the *rank sum test*
- If intervention has no effect, *randomly permuting* the allocation of clusters between study arms will have no effect on the observed effect measure
- Procedure:
  - The effect measure is computed for every possible permutation
  - The *p value* is calculated as the proportion of all permutations that give an effect measure at least as extreme as the observed one
  - If there is a very large number of permutations, a random selection of them can be used to generate the *p value*

Note: Need at least 4 clusters per study arm to obtain

$p < 0.05$ in a non-parametric test

Summary

1. Basic principles of analysis
   - experimental vs observational units
   - account of *within-cluster correlation (*otherwise SE will be too small)

2. Baseline data and analysis
   - to characterise the study population (to assess generalisability)
   - check for baseline imbalances, decisions on adjustment

3. Estimates of effect
   - either cluster summaries *or* individual level data

Summary

4. Cluster level analysis
   - carry out *t*-test on cluster level summaries
   - simple, but very useful method
   - robust; works well in many situations where more complicated methods fail

5. Adjusting for covariates
   - to account for imbalance between clusters
   - adjusting for individual level covariates can be done with 2 step procedures

6. Non-parametric tests
   - useful if t-test inappropriate due to violations of assumptions (but less powerful)

Further methods of analysis of CRTs

Aims and learning outcomes

By the end of this lecture you will:

- Know how to use generalised estimating equations to analyse CTRs
- Understand when to use different analysis methods
- Know how to analyse stratified trials
- Know how to analyse matched trials
- Understand the issues regarding controlling for baseline values

Generalised estimating equations

- GEEs are an individual level regression method that can account for clustering
  - they provide an alternative to random effects models

- They model the correlation between observations from the same cluster
  - as opposed to random effects models which model the between cluster variation

- They do not include additional terms in the model and produce *population average* measures of effect

- Like random effect models, they only work well if there are a sufficient number of clusters

Generalised estimating equations

- We must specify a correlation matrix. For cluster randomised trials the most appropriate is an *exchangeable* matrix.
- This means that individuals in the same cluster have the same correlation coefficient ρ but individuals in different clusters are uncorrelated
- ρ is estimated from the data and is used in the estimation of regression coefficients and standard errors
- If ρ is large less weight is given to individuals in large clusters
- GEEs provide valid inference even if our assumed correlation structure is wrong
  - Although getting it right improves power

Generalised estimating equations

- Unlike random effects models, GEEs do not provide a full probability model for the data.

- This means we cannot get *maximum likelihood estimates* or use *likelihood ratio* tests

- Inference is by the Wald test, i.e. we calculate

$$z = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

and refer this to the normal distribution. 95% confidence intervals can be obtained as:

$$\hat{\beta} \pm 1.96 \times SE(\hat{\beta}).$$

- We use the *sandwich variance estimator* to obtain standard errors

GEEs: binary data

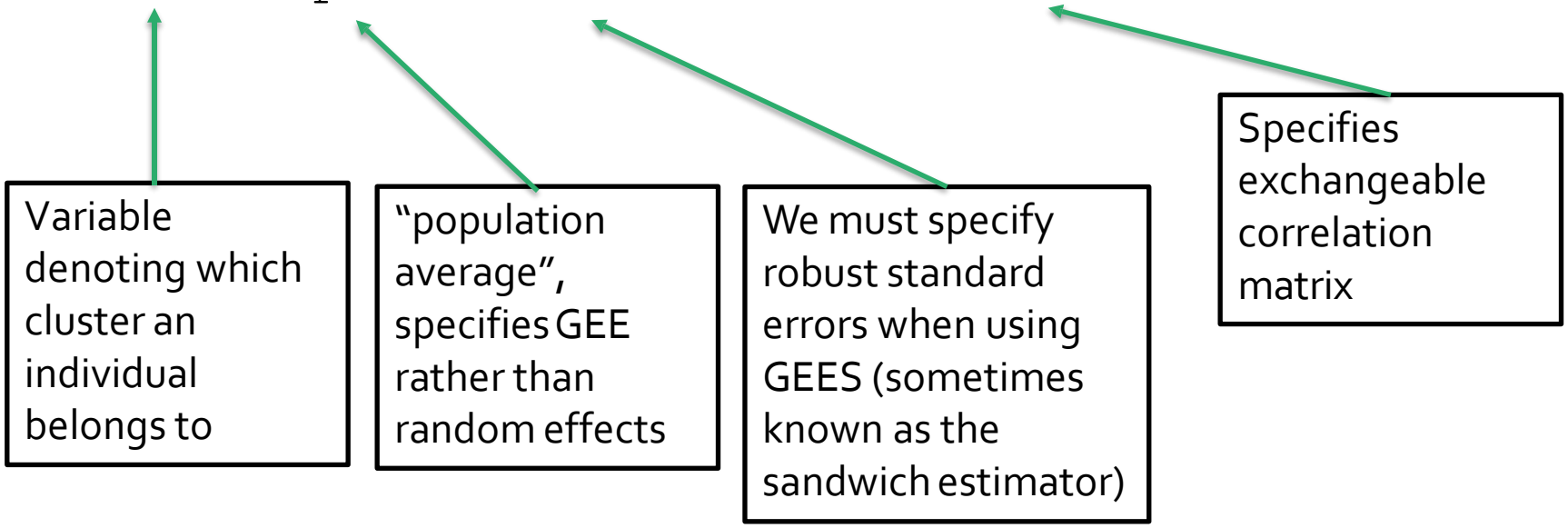
- Unlike random effects models, there are no additional terms in the model
- The regression model for GEEs is identical to a model that ignores clustering
- So for binary data the model is:

$$\text{logit}(\pi_{ijk}) = \log\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \alpha + \beta_i + \sum_l \gamma_l z_{ijkl}$$

GEEs: binary data

Stata command for a GEE model:

```
xtlogit outcome i.intervention covariates, or ///
i(cluster) pa vce(robust) corr(exch)
```



GEEs: binary data, example

- The THPP trial measured the effect of a simplified cognitive behavioural therapy delivered by peers on a number of outcomes in Pakistan.

- The trial had 40 clusters

- Here we look at the secondary outcome of recovery from depression, which is a binary outcome.

GEEs: binary data

```
xtlogit recovery i.arm, ///
i(cluster) pa or robust corr(exch)
```

```
GEE population-averaged model                   Number of obs     =        445
Group variable:                      cluster    Number of groups  =         40
Link:                                  logit    Obs per group:
Family:                             binomial                 min =          7
Correlation:                   exchangeable                 avg =       11.1
                                                            max =         15
                                                Wald chi2(1)      =       5.97
Scale parameter:                           1    Prob > chi2       =     0.0146

                                 (Std. Err. adjusted for clustering on cluster)
------------------------------------------------------------------------------
             |               Robust
    recovery | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         arm |
intervention |    1.58598   .2993676     2.44   0.015     1.095534    2.295987
       _cons |    .312936   .0510095    -7.13   0.000     .2273564    .4307287
------------------------------------------------------------------------------
```

GEEs: binary data

- The estimated odds ratio of the intervention is 1.59 (95% CI 1.10-2.30). The output provides a z-statistic of 2.44 and corresponding p-value of 0.015
- A random effects model gives an odds ratio of 1.60 (95% CI 1.05-2.42), p = 0.028
- The inference from the two models is similar
- The OR from the random effects model is very slightly further from the null.
- They are so similar is because there is low between cluster variation, ($\rho = 1.6 \times 10^{-7}$, where ρ is the ICC).

- GEE models may also be fitted for event-rate data and quantitative outcomes.

- As with binary data, an **exchangeable** correlation matrix and **robust** standard errors are used.

- We also use the **Wald** test to carry out significance tests.

Analytical method for unstratified trials

- If there are a small number of clusters per treatment arm (fewer than around 15) cluster-level analyses are recommended because with few clusters GEEs and random effects models underestimate standard errors

- For trials with larger numbers of clusters, either approach can be used
  - individual-level regression methods can be more efficient
  - and more convenient when analysing the effects of individual-level covariates or in the presence of effect modification

- New research is being done on methods that allow GEE and random effects to be used when the number of clusters is small
  - We will look at one such method in the practical
  - They may only have more power than cluster level analyses in specific circumstances (e.g. large variation in cluster sizes + large coefficient of variation of the outcome)

Analysis of stratified trials

- Stratified trials can be analysed using cluster level summaries or individual level regression methods
- We begin with analysis using cluster level summaries
- (For the sake of illustration we will work with risk differences, applications to other measures of effect are straightforward)

Analysis of stratified trials

- Instead of a *t*-test on cluster level summaries, we perform a *linear regression* on cluster level summaries *with adjustment for stratum as a factor variable*

- The *t*-test is a special case of linear regression with a single binary independent variable

- If stratum is predictive of the outcome, then it will reduce residual variance and increase the precision of the estimate of the effect size

Analysis of stratified trials: example

- The Manas trial looked at an intervention led by lay health counsellors for depressive and anxiety disorders in primary care in Goa, India.
- The 24 clusters (primary health care facilities) were in 2 strata: 12 public facilities and 12 private facilities.
- In each stratum, 50% of the clusters were randomised to the intervention arm.
- The primary endpoint was recovery from common mental disorders 6 months after initiation of treatment (binary)

Analysis of stratified trials: example

- We must first calculate the log proportion recovery for each cluster.

- There are many way of doing this in Stata, here is an example of how it might be done:

```
. gen n=1

. collapse (sum) cmd_recover n (median) strata arm,
by(clinic)

. gen precover= cmd_recover/n

. gen lprecover =log(precover)
```

Analysis of stratified trials: example

- We then perform a linear regression on these cluster level summaries, adjusting for stratum

```
. regress lprecover i.arm i.strata

      Source |       SS           df       MS      Number of obs   =        24
-------------+----------------------------------   F(2, 21)        =      3.97
       Model |  .502749385         2  .251374692   Prob > F        =    0.0346
    Residual |  1.33127056        21  .063393836   R-squared       =    0.2741
-------------+----------------------------------   Adj R-squared   =    0.2050
       Total |  1.83401995        23  .079739998   Root MSE        =    .25178

------------------------------------------------------------------------------
   lprecover | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     1.arm   |   .2050838   .1027893    -2.00   0.059    -.0086783    .4188458
    2.strata |   .2042846   .1027893     1.99   0.060    -.0094774    .4180467
       _cons |  -.5331364   .0890181    -5.99   0.000    -.7182597    -.348013
------------------------------------------------------------------------------
```

- Since this is on the log scale, we must exponentiate to get the risk ratio:

```
. disp exp(.2050838), exp(-.0086783), exp(.4188458)
1.2276279 .99135925 1.5202059
```

- So the risk ratio is 1.23, the 95% confidence interval is (0.99 – 1.52) and the p-value is 0.059, indicating weak evidence that the intervention is effective.

Analytical methods for stratified trials

- Analysing stratified trials using individual level analysis is relatively straight forward.

- We fit either a GEE or random effects regression model, as described previously, with the addition of a **fixed effect** for each **stratum**. Assuming a binary outcome, we may fit the following random effects logistic regression model:

$$\text{logit}(\pi_{sijk}) = \log\left(\frac{\pi_{sijk}}{1-\pi_{sijk}}\right) = \alpha_s + \beta_i + \sum_l \gamma_l z_{sijjkl} + u_{ij}$$

where α_s is the stratum effect and all other parameters are as before.

- As a rough guideline, individual-level regression is not recommended for stratified trials with less than 20 clusters per arm, due to the extra parameters.

Analysis of pair-matched CRTs

- Cluster level analysis of pair-matched CRTs is very straightforward.
- We simply perform a paired *t*-test on the cluster level summaries.
- i.e. we take the risk (or whatever) difference in each pair:

$$h_j = r_{1j} - r_{0j}$$

- The overall rate difference is the average of these:

$$h = \frac{1}{c}\sum_j h_j = \frac{1}{c}\sum_j \left(r_{1j} - r_{oj}\right) = \bar{r_1} - \bar{r_0}$$

Analysis of pair-matched CRTs

- The confidence interval is

$$\bar{h} \pm t_{\nu,0.025} \times \frac{s_m}{\sqrt{c}}$$

where c is the number of pairs,

$$s_m^2 = \frac{1}{c-1}\sum_j \left(h_j - \bar{h}\right)^2$$

and

$$\upsilon = (c-1).$$

Analysis of pair-matched CRTs

- A hypothesis test is performed by calculating

$$t_m = \bar{h} \Big/ \sqrt{\frac{s_m^2}{c}}$$

and referring to the *t*-distribution with $\upsilon = (c - 1)$ degrees of freedom.

Analysis of pair-matched CRTs

- Individual level analyses of pair matched trials are difficult.
- It may seem natural to fit a model analogous to that for stratified trials:

$$\text{logit}(\pi_{ijk}) = \log\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \alpha_j + \beta_i + \sum_l \gamma_l z_{ijkl} + u_{ij}$$

where α_j is the pair effect.

- However, estimation and inference

to be

lihood

- If either the outcome is quantitative, or the number of clusters is large, it is possible to fit a model with a random effect for cluster nested within a random effect for pair.

Analysis of pair-matched CRTs

- One option in the analysis of pair-matched studies is to break the matching for the analysis.
- This could achieve balance without loss of degrees of freedom in the analysis.
- However this method may be less effective at balancing on a range of factors than restricted randomisation (as described in lecture 2)
- If breaking is to be used, the investigator must specify this in, resulting in a loss of power if it turns out that the matching was highly effective.
- Calculating the matching correlation once the trial is complete and using this to decide whether to use unmatched or matched analysis increases the *type I error.*

Controlling for baseline values

- In some CRTs, data are collected on baseline values of the main endpoints.

- We may wish to adjust estimated intervention effects for baseline values, for two reasons:
  - To control for any baseline imbalances in the endpoint of interest between treatment arms.
  - To reduce between-cluster variation in the endpoint and thus to increase the power and precision of the study.

- Repeated cross-sectional design give baseline and follow-up data on **different individuals**. In this situation we can only adjust for the baseline value of the endpoint at the **cluster level**.

- A cohort design gives baseline and follow-up data on the **same individuals**, allowing adjustment at **individual level.**

Controlling for baseline values

- Two alternative analytical approaches can be used to adjust for baseline values
  - analyse the **change** in the endpoint of interest
  - adjust for the baseline value as a **covariate** in the regression

- Because of random variation in measured endpoints those with low observed values at baseline are expected to show an increase in observed value at follow-up (and vice versa) even in the absence of any true change. This phenomenon is called **regression to the mean**.

- Therefore adjustment for baseline as a **covariate** is generally the method of choice.

Summary

- GEEs are an individual level analysis method that provide an alternative to random effects models

- When there are a small number of clusters (fewer than 15 per arm), cluster level methods are generally recommended. With more clusters, individual level regression models may be more efficient.

- Stratification or matching should be taken into account in the analysis

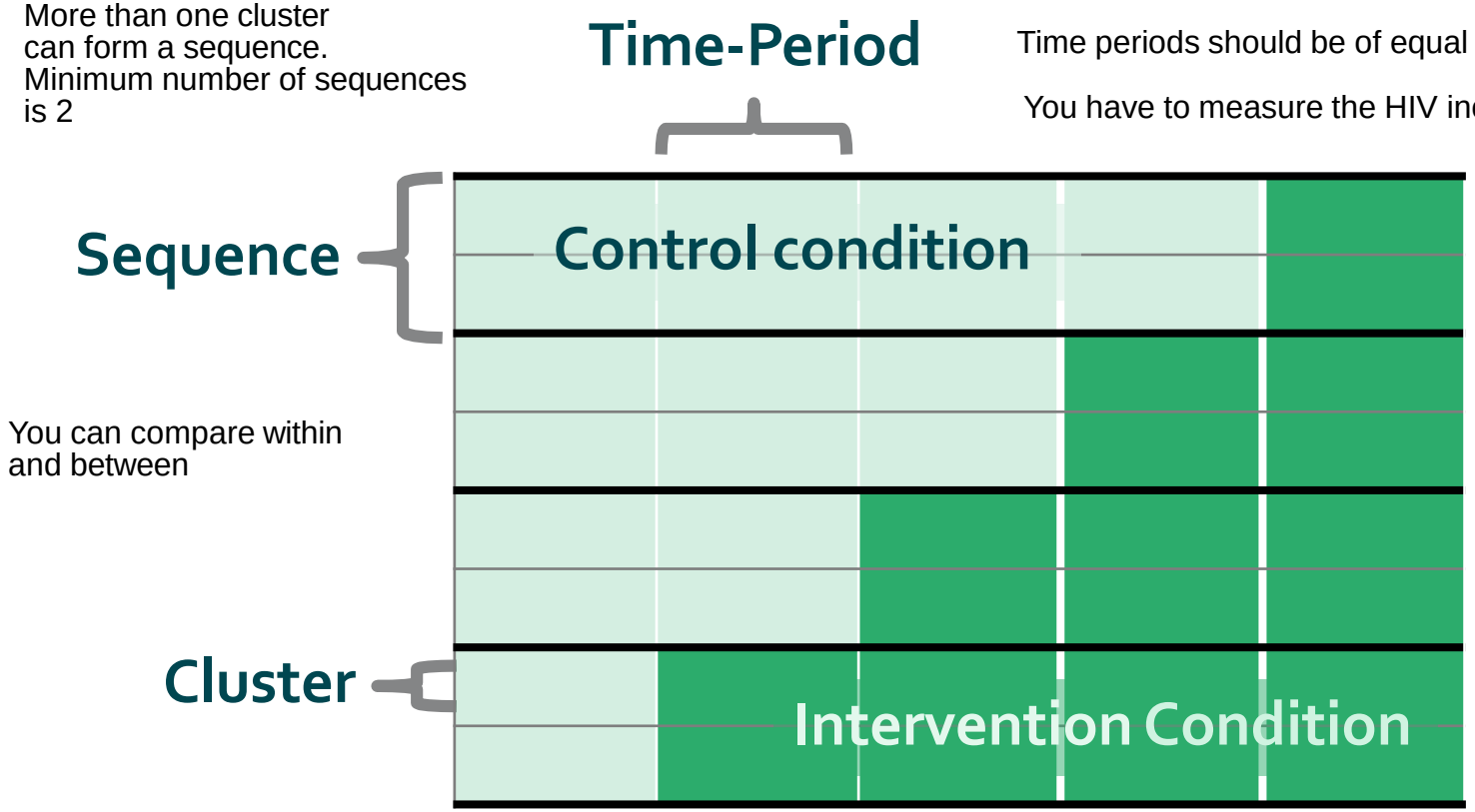
- Linear regression of cluster level summaries, with adjustment for stratum can be used to analyse stratified trials with a small number of clusters; random effects/GEEs with adjustment for stratum can be used for trials with a large number of clusters

- A paired t-test can be used to analyse matched trials

- When taking baseline variables into account, it's generally better to adjust for the baseline variables, rather than analyse difference from baseline

What is a stepped wedge trial?

More than one cluster can form a sequence. Minimum number of sequences is 2



Time periods should be of equal distance

You have to measure the HIV incidence at each time-period

You can compare within and between

Sample size calculation incorporates the sequence.

Recruit all the clusters at baseline so that you can collect control data

You randomie clusters to the time period when they'll switch to the intervention.

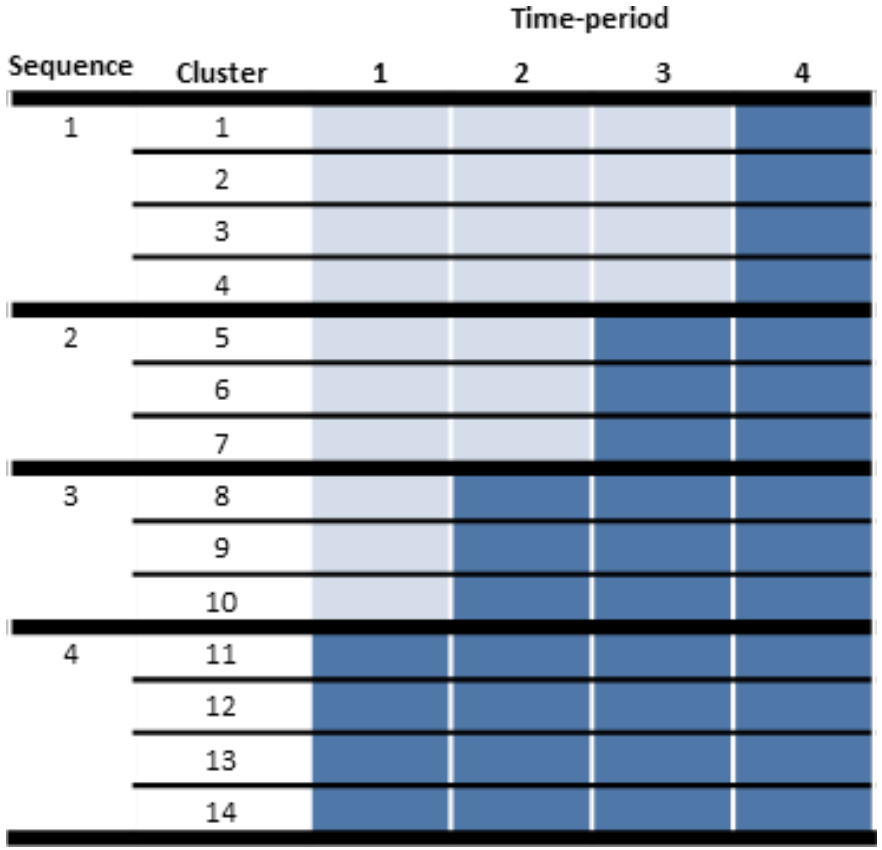
You can have unequal number of clusters at each time-period

You should blind clusters about when they'll start to avoid them behaving differently

Example: School breakfast trial

This example is better for HIV trials

- Trial in New Zealand
- Impact of school breakfast on school attendance
- Each term, free breakfasts were introduced to schools in another sequence
- Children's attendance was measured throughout the year.
- Most children in the study were measured in all four periods



No need to have all clusters starting in a control state. But that means there will be higher intervention period than control period.

Mhurchu CN, *et al* Effects of a free school breakfast programme on children's attendance, academic achievement and short-term hunger: results from a stepped-wedge, cluster randomised controlled trial. *JECH* 2013

Rationale

Why might we consider using a stepped wedge trial design?

- Fewer resources are required for a phased rollout
  - Staggered rollout means the intervention doesn't need to be introduced to many clusters at once
- Improved recruitment and retention of clusters
  - Increase the acceptability of the control condition, as the intervention will be given to all clusters before the end of the trial.
  - This is thought to improve recruitment of clusters as well.
- Evaluation during program rollout
  - The only way to incorporate randomisation
- Power might be higher than a parallel CRT with the same sample size

Rationale: Example (1)

- Example: School breakfast trial
- Possible reasons:
- The breakfast program only needed to be introduced to 3 or 4 schools each term
- There may have been a country wide initiative to introduce free breakfasts to all schools, so all schools had to be given free breakfasts
- Schools might not have been as keen to join a trial where they might have been randomised to the control: They just carry on as normal but have researchers coming asking them for data about attendance

Rationale: Example (2)

- Example: School breakfast trial
- Possible reasons:
- The breakfast program only needed to be introduced to 3 or 4 schools each term
- There may have been a country wide initiative to introduce free breakfasts to all schools, so all schools had to be given free breakfasts
- Schools might not have been as keen to join a trial where they might have been randomised to the control: They just carry on as normal but have researchers coming asking them for data about attendance

Disadvantages

What are the downsides to the stepped wedge design?

- Complex to implement
  - Increasing workload
  - Higher burden of observation collection than a parallel CRT
  - Timing of implementing intervention is important
- Complex to analyse
  - Intervention effect is confounded with other changes in the outcome over time
  - More on this later…
- Power might be lower than a parallel CRT with the same sample size

Disadvantages (2)

Example: School breakfast trial

- All schools in the trial have free breakfast programs by the end of the trial: this could be complex to maintain
- We need to know attendance of children in each term, rather than say just in the last term
- Some schools might have trouble starting up the free breakfasts at the start of the term that they are randomised to
- Complex analysis to be covered later

Summary

- In a stepped wedge trial, clusters are randomised to when they will switch from a control condition to an intervention condition
- There are many benefits to using this design but also many drawbacks

In the next part of this lecture, we will explore what these trials look like in more detail focusing on how individuals are exposed to the intervention and how their outcomes are measured in different types of stepped wedge trials.

Methodological issues in evaluating real-world vaccine effectiveness during the global rollout of COVID-19 vaccines

Sisonke phase 3B implementation study

THE LANCET

Effectiveness of the Ad26.COV2.S vaccine in health-care workers in South Africa (the Sisonke study): results from a single-arm, open-label, phase 3B, implementation study

Linda-Gail Bekker, Nigel Garrett, Ameena Goga, Lara Fairall, Tarylee Reddy, Nonhlanhla Yende-Zuma, Reshma Kassanjee, Shirley Collie, Ian Sanne, Andrew Boulle, Ishen Seocharan, Imke Engelbrecht, Mary-Ann Davies, Jared Champion, Tommy Chen, Sarah Bennett, Selaelo Mametja, Mabatlo Semenya, Harry Moultrie, Tulio de Oliveira, Richard John Lessells, Cheryl Cohen, Waasila Jassat, Michelle Groome, Anne Von Gottberg, Engelbert Le Roux, Kentse Khuto, Dan Barouch, Hassan Mahomed, Milani Wolmarans, Petro Rousseau, Debbie Bradshaw, Michelle Mulder, Jessica Opie, Vernon Louw, Barry Jacobson, Pradeep Rowji, Jonny G Peter , Azwi Takalani, Jackline Odhiambo, Fatima Mayat, Simbarashe Takuva, Lawrence Corey, Glenda E Gray, and the Sisonke Protocol Team, on behalf of the Sisonke Study Team

- **P**opulation: Health-care workers and essential workers in South Africa (aged ≥18 years)
- **I**ntervention: Vaccination with single-dose Johnson & Johnson Ad26.COV2.S
- **C**omparison group: unvaccinated essential workers at high risk of contracting COVID-19
- **O**utcome: Occurrence of COVID-19 related (i) admissions, (ii) admissions requiring ICU or CCU, (iii) deaths at ≥ 28 days post vaccination
- **T**ime: Vaccination (Feb 17- May 17, 2021); follow-up to July 17, 2021 (slightly covers huge spike in cases country-wide)
- **D**esign: Cohort study (1:1 matched on age, sex, number of comorbidities, geographical location (health district), socioeconomic status
- **Sample size:** 215 813 in each group (intervention and comparator)

Sisonke phase 3B implementation study

THE LANCET

Effectiveness of the Ad26.COV2.S vaccine in health-care workers in South Africa (the Sisonke study): results from a single-arm, open-label, phase 3B, implementation study

Linda-Gail Bekker, Nigel Garrett, Ameena Goga, Lara Fairall, Tarylee Reddy, Nonhlanhla Yende-Zuma, Reshma Kassanjee, Shirley Collie, Ian Sanne, Andrew Boulle, Ishen Seocharan, Imke Engelbrecht, Mary-Ann Davies, Jared Champion, Tommy Chen, Sarah Bennett, Selaelo Mametja, Mabatlo Semenya, Harry Moultrie, Tulio de Oliveira, Richard John Lessells, Cheryl Cohen, Waasila Jassat, Michelle Groome, Anne Von Gottberg, Engelbert Le Roux, Kentse Khuto, Dan Barouch, Hassan Mahomed, Milani Wolmarans, Petro Rousseau, Debbie Bradshaw, Michelle Mulder, Jessica Opie, Vernon Louw, Barry Jacobson, Pradeep Rowji, Jonny G Peter , Azwi Takalani, Jackline Odhiambo, Fatima Mayat, Simbarashe Takuva, Lawrence Corey, Glenda E Gray, and the Sisonke Protocol Team, on behalf of the Sisonke Study Team

- Vaccine effectiveness (VE) estimates:
  - 67% (95% CI: 62-71) against admissions
  - 75% (95% CI: 69-82) against ICU& CCU admissions
  - 83% (95% CI: 75-89) against death

Data sources: Medical aid schemes databases (insured population), Electronic Vaccination Data System (EVDS) , National Population Register, COVID-19 Notifiable Medical Conditions Sentinel Surveillance (NMCSS) master list, Western Cape Provincial Health Data Centre.

National vaccine roll-out strategy

Phase	Priority group	Start date
1a	Front-line and health care workers (Sisonke Protocol)[11][12]	17 Feb 2021
1b	Remaining health care workers[14]	17 May 2021
	Age Based Prioritisation	
2ai	People over 60[14]	17 May 2021
2aii	People 50-59[16]	5 July 2021
2aii	People 35-49[17]	15 July 2021
	Employment Based Prioritisation[18]	
2bi	Teachers and support staff[19]	23 June 2021
2bii	Police Force, SANDF, inmates & prison staff[20]	5 July 2021
3	People 18-34[23]	20 Aug 2021
4a	Kids and teenagers 12-17 1st Dose[24]	20 Oct 2021
4b	Boosters for Sisonke Trial participants [25]	10 Nov 2021[26]
4c	Additional doses for immunocompromised over 18[27]	1 Dec 2021
4d	Booster doses (For those with single dose of the Johnson and Johnson vaccine at least two months ago, or two-dose Pfizer series at least six months ago	24 December 2021
4e	Boosting interval reduced to 90 days for Pfizer and heterologous boosting permitted	23 February 2022

Sisonke: limitations and data issues

- Despite matching, SARS-CoV-2 exposure might be lower in the comparator group than in the health-care worker population
- Comparator group likely to introduce bias
- Not an RCT design, matching could not have eliminated residual confounders or bias
- Ascertainment and linking of various databases was a challenge: confounders or effect modifiers, vaccination status for the comparator group (censoring purposes), endpoints, previous SARS-CoV-2 infection
- Best option was to use medical insurance data: out of the 477 102 vaccinated HCW, 215 813 (45%) analysed
- Possibility of selection bias due to linking of data via medical schemes (generalizability of Sisonke results to the general population)

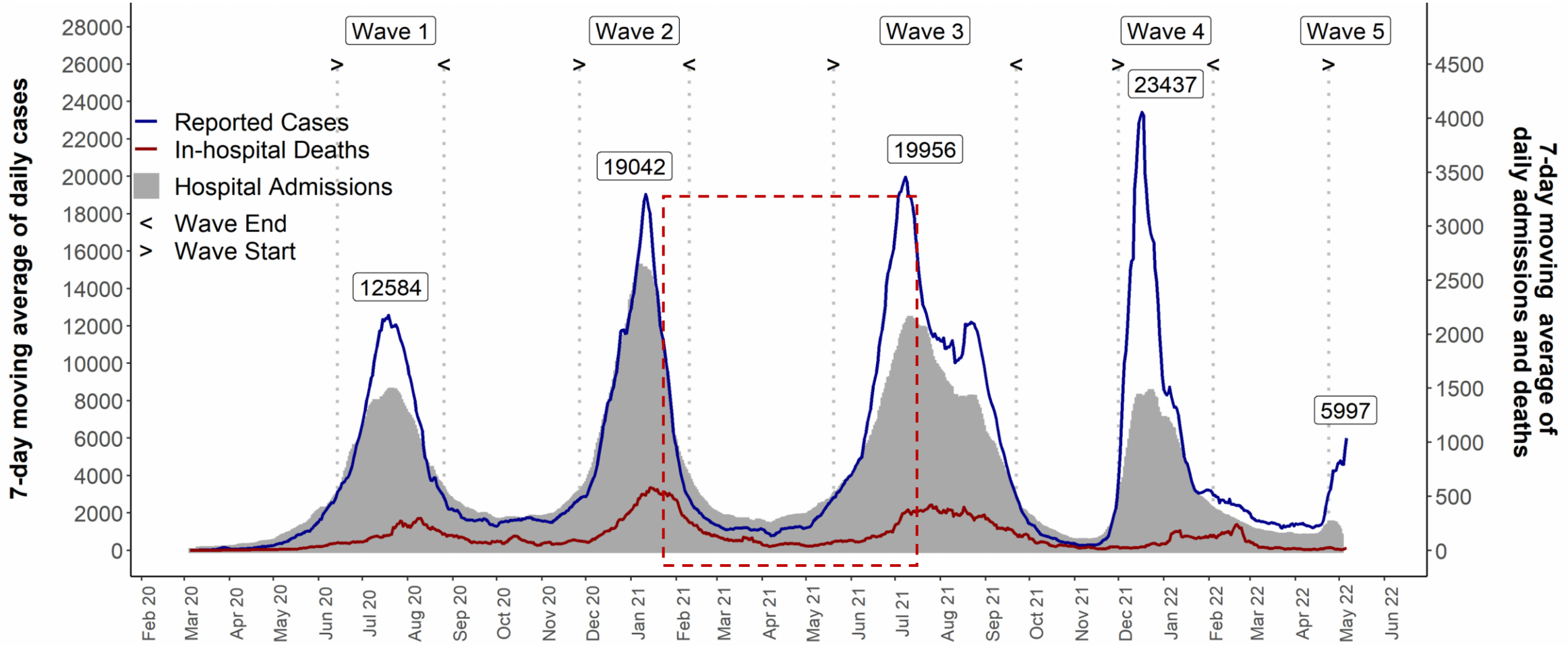
Sisonke: limitations and data issues

- Medical insurance data:
  - Allowed for complete adjustment for effect modifiers/confounding and stratification (sub-group) analyses. HIV burden in our setting.
  - In Sisonke , self-reported HIV prevalence was 8.3% vs. 16.3% obtained from medial insurance data (2-fold higher). Misclassification of participants due to underreporting.

- High prevalence of comorbidities, including a high HIV prevalence, could have reduced vaccine effectiveness compared with other studies.
- To assess robustness of Sisonke results, three analyses conducted using datasets from two medical insurance schemes and a provincial public sector database of health-care workers (strong comparator group).
- Where same endpoint analysed on all three datasets, VE estimates were similar

Covid-19 in South Africa

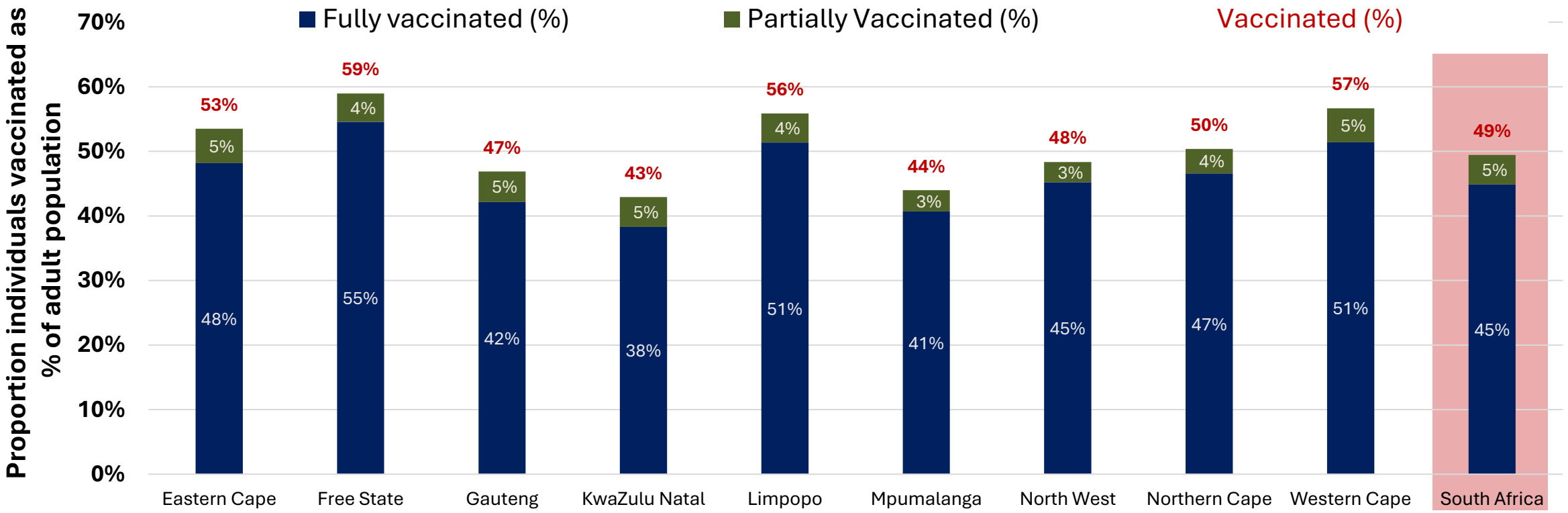
up to - 06 May 2022



Source of hospital admissions data: Lucille Blumberg, Richard Welch and Waasila Jassat – DATCOV, NICD
Diagram from Salim Abdool Karim, CAPRISA

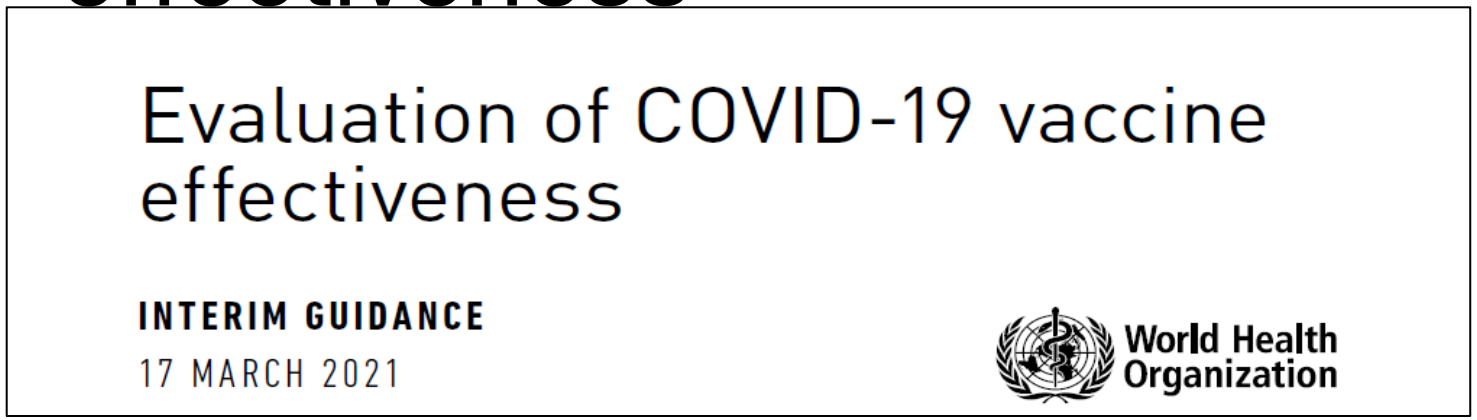
Proportion of individuals vaccinated as % of adult population

up to – 06 May 2022



Source: *Department of Health.* https://sacoronavirus.co.za/latest-vaccine-statistics/ *Diagram from Salim Abdool Karim, CAPRISA*

Study designs to assess vaccine effectiveness



- Screening method (case-population method)
- Regression discontinuity (RDD)

TND design

- Widely used to estimate influenza vaccine effectiveness (VE) in non-randomised studies
- Patients who seek care for a defined set of symptoms/signs are enrolled
- Valid estimates: The same selective forces that lead individuals to be tested will operate on both those who test positive and those who test negative
- Design easiest to implement; it only requires collecting information about potential risk factors for the disease of interest from symptomatic people who were tested
- VE calculated by comparing odds of vaccination among cases compared with controls, after adjusting for potential confounding factors

Sensitivity analysis

- Period of non-effectiveness (should be guided by immunological response and formal guidelines)
- Different inclusion criteria and occupational exposure
- Different methodological approaches
- Sensitivity to background prior SARS-CoV2 infection rate
- Sensitivity to misclassification of exposure

TND design

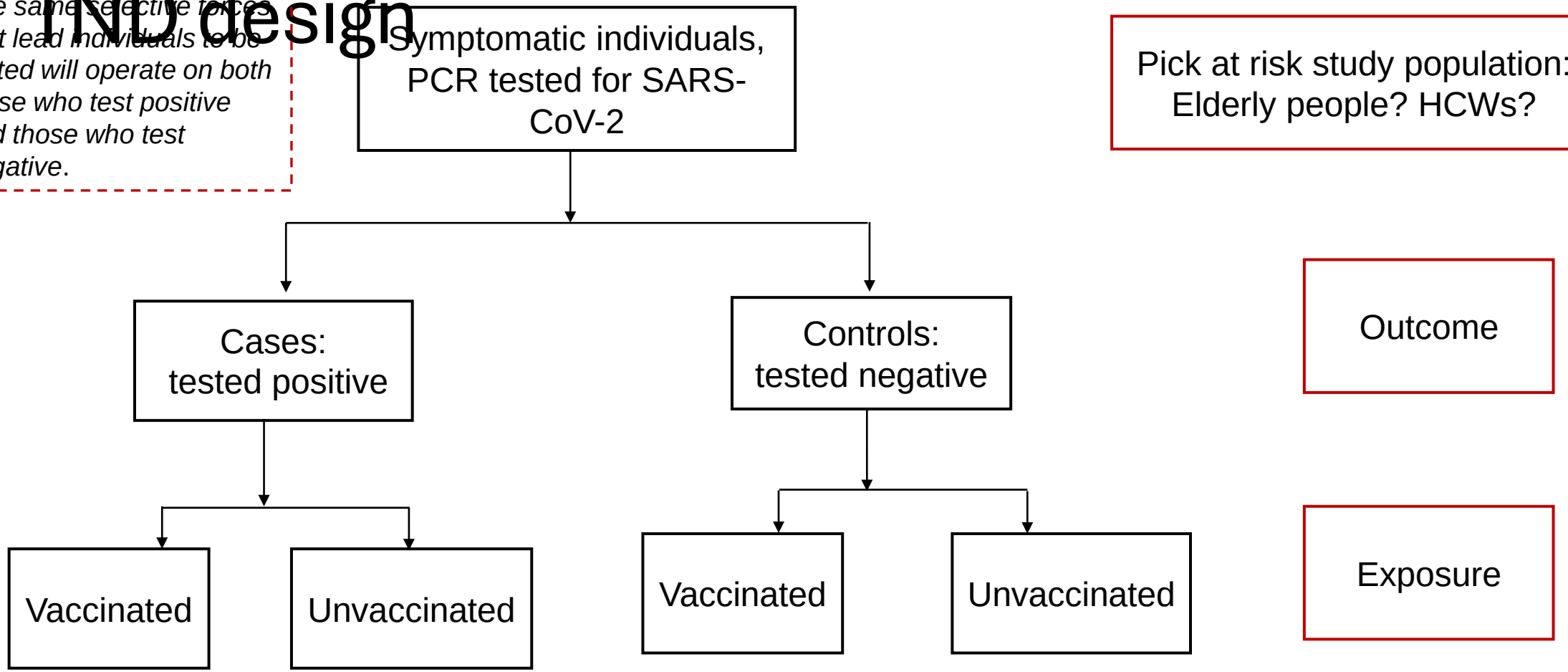
- However, health seeking behaviour might differ between vaccinated and unvaccinated people, especially if they differ in age, underlying comorbidities, etc (end up with more unvaccinated people in the study)
  - Careful when choosing study population
- Selection bias usually introduced by people who order diagnostic testing (concentrate on unvaccinated people)
- Control selection should be selected independent of vaccination status
- Ideally, controls should be tested more than once to confirm their status before the final analyses (to avoid misclassification)

Cohort study

- Include all population members (not restricted to symptomatic people like TND)
- Able to study different outcomes: COVID-19 cases (irrespective of symptoms?), hospitalizations, deaths
- BUT ascertainment of reliable endpoints linked with vaccination data in our setting is a challenge
- Compare outcomes for vaccinated vs. unvaccinated
  - Matching by potential confounders is encouraged because vaccinated and unvaccinated are different
    - Likelihood of infection differ
    - Likelihood of seeking medical care differ
    - Prognostic factors for severe illness and death
- Matching forces vaccinated and unvaccinated groups to look similar in many measured characteristics
- Vaccination date important to censor people when they get vaccinated (reliable data sources!)
- Complete data on potential confounders essential to avoid matching only selected group of people

TND design

The same selective forces that lead individuals to be tested will operate on both those who test positive and those who test negative.



Pick at risk study population: Elderly people? HCWs?

Outcome

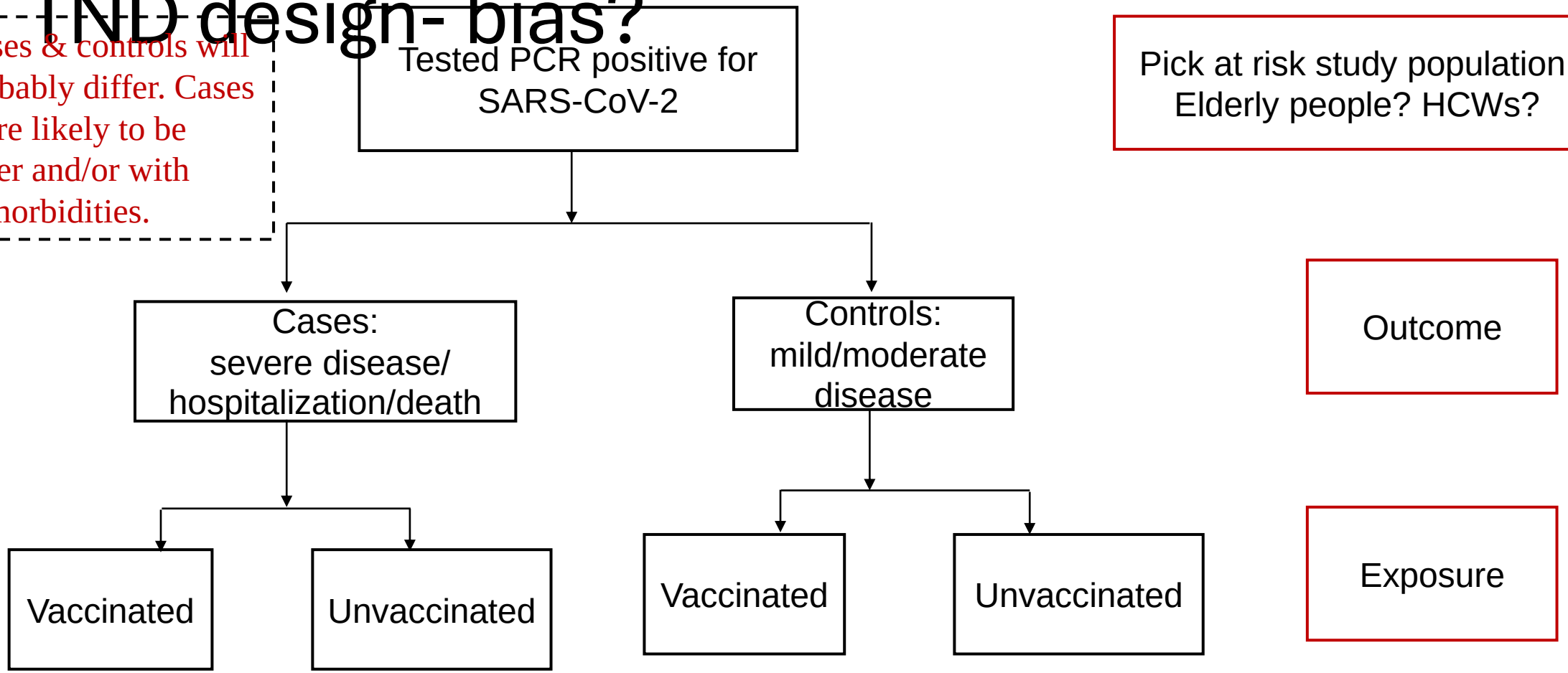
Exposure

Potential confounders: age, gender, calendar time, geographical location, race, socio-economic status and comorbidities, etc.

VE: 1- adjusted odds ratios

TND design- bias?

Cases & controls will probably differ. Cases more likely to be older and/or with comorbidities.



Pick at risk study population: Elderly people? HCWs?

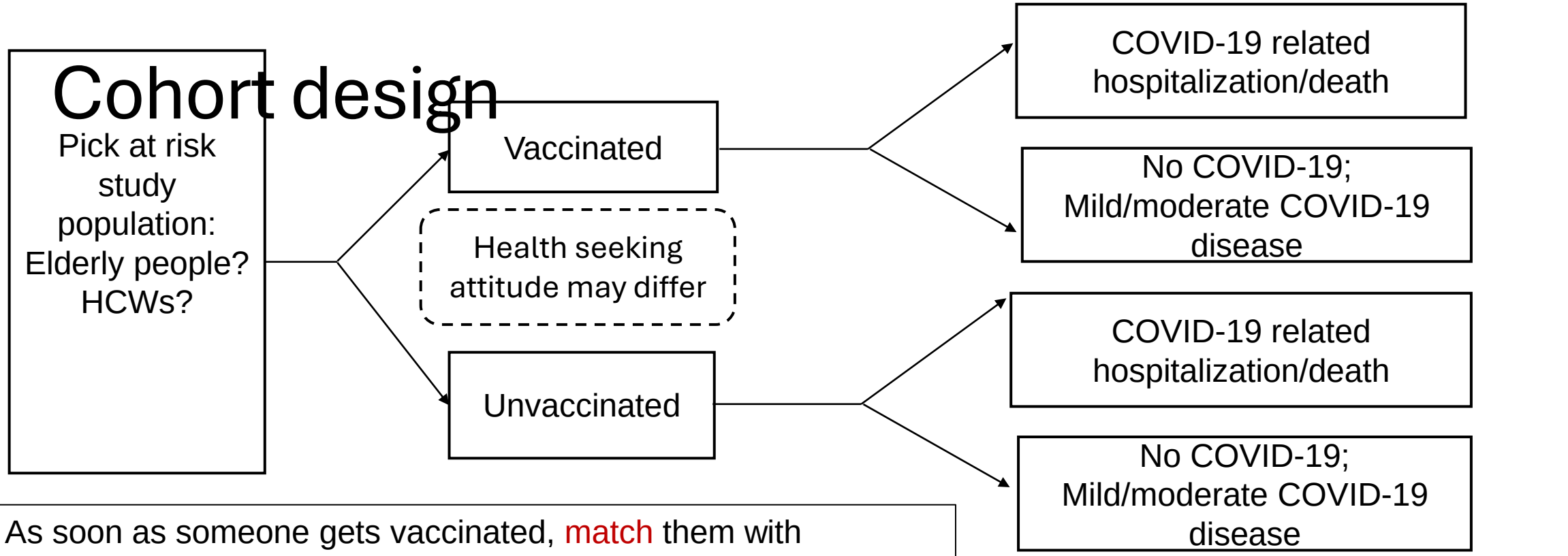
Outcome

Exposure

Potential confounders: age, gender, calendar time, geographical location, race, socio-economic status and comorbidities, etc.

VE: 1- adjusted odds ratios (likely biased)

Cohort design



1. As soon as someone gets vaccinated, match them with someone who is unvaccinated (**database should record the pairing**)
2. Alternatively, use propensity score (PS) to match vaccinated and unvaccinated people

- Matching or PS variables: age, gender, calendar time, geographical location, race, socio-economic status and health status, etc.
- BUT missing variables disrupt this process

Note:

- Outcome should be clearly defined and easily accessible
- Matching variables carefully chosen and easily accessible

VE: 1- adjusted risk or hazard ratios

Biases, weaknesses, strengths of COVID-19 vaccine effectiveness studies

- Each design has strengths and weaknesses
- Although not without biases, WHO recommend conducting the TND as an efficient and accurate method in L/MICs to assess VE against severe and symptomatic COVID-19.

Strengths and weakness of VE designs

Type of study	Strengths	Weaknesses
Test negative case control	Reduces bias of differences in health care seeking behaviour and access by vaccine status	False-negative misclassification more likely as both cases and controls have COVID-19-like illness
	Vaccination status often obtained before results of laboratory tests available, minimizing diagnostic bias	Test-negative controls more likely to be tested for exacerbation of an underlying illness (e.g. COPD), that is an indication for COVID-19 vaccination leading to increased VE
	All cases and controls seek care at same facilities, potentially decreasing differences in access to vaccines and community-level confounders	Cases and controls need to be matched or the analysis needs to be adjusted by time

Strengths and weakness of VE designs

Type of study	Strengths	Weaknesses
Cohort	Results easily communicated to policy-makers and stakeholders	Vaccination status difficult to determine in retrospective cohorts without good vaccination records
	Can potentially be used to study asymptomatic or mildly symptomatic infections	If prospective, possible ethical dilemma in following unvaccinated persons who are recommended for vaccination
		Requires large sample size, especially if outcome of interest is uncommon such as severe COVID-19

Potential biases of COVID-19 vaccine effectiveness studies

Bias	Description	Designs affected	Typical magnitude	Direction on VE estimate	Methods to minimize bias
Care seeking behaviour/ access to care	Those more likely to get vaccine seek care more, thus more likely to be cases	TND, cohort	Large	Decrease	Use TND; enrol only severe patients
Care seeking based on vaccine status	Vaccinated persons less likely to seek care/testing due to COVID-19-like illness due to perception of protection	TND, cohort	Small-moderate	Increase in cohort. Decrease in TND, if vaccine confers some protection	Smaller magnitude in TND

Potential biases of COVID-19 vaccine effectiveness studies

Bias	Description	Designs affected	Typical magnitude	Direction on VE estimate	Methods to minimize bias
Diagnostic bias	Health workers more likely to test unvaccinated persons for COVID-19	TND, cohort	Varies on setting	Increases	Test all persons or random sample
Misclassification of the outcome	False negatives (persons with COVID-19 disease who test negative)	TND, cohort	Small	Decrease	Use highly sensitive test; limit to illness onset ≤ 10 days; exclude TND controls with COVID-19-specific symptoms

Potential biases of COVID-19 vaccine effectiveness studies

Bias	Description	Designs affected	Typical magnitude	Direction on VE estimate	Methods to minimize bias
Exposure misclassification	Vaccine effect may start before/after specified cut-off for considering individual vaccinated	TND, cohort	Large but can be nearly eliminated by design	Decrease	Exclude outcomes occurring in periods of ambiguous vaccine effect, e.g. 2 weeks after first dose
Prior infection	If known prior SARS-CoV-2 infection, less likely to get vaccinated	TND, cohort	Small-moderate (depends on seroprevalence / past incidence of infection)	Decrease	Sensitivity analysis excluding those with prior SARS-CoV-2 infection by history or lab

Evaluating durability of vaccine effectiveness against clinical outcomes

- Several biases can affect VE evaluations of duration of protection
- Differential rates of infection
- Different variants
- Depletion of susceptible persons between vaccinated and unvaccinated cohorts
- Bias of the VE over time can also occur if either health seeking or diagnostic testing changes over time, and is related to vaccination status.

Evaluating durability of vaccine effectiveness against clinical outcomes

- Durability of vaccine effectiveness has been assessed using the following designs

- Following a cohort of only vaccinated individuals and compare rates of disease in strata defined by duration since date of vaccination.
- Test Negative Case Control (Substudy for each time-since-vaccination stratum)
- Adjusted cohort approach (Time varying Cox model, calendar time as underlying timescale)

Evaluating effectiveness of booster doses

- Misclassification of exposure
- Matched cohort design is not feasible to assess multiple boosting regimens
- Accounting for time since primary vaccination is crucial
- Statistical power (Effect size? Non-inferiority comparisons in the analysis? Endpoints? Booster uptake)

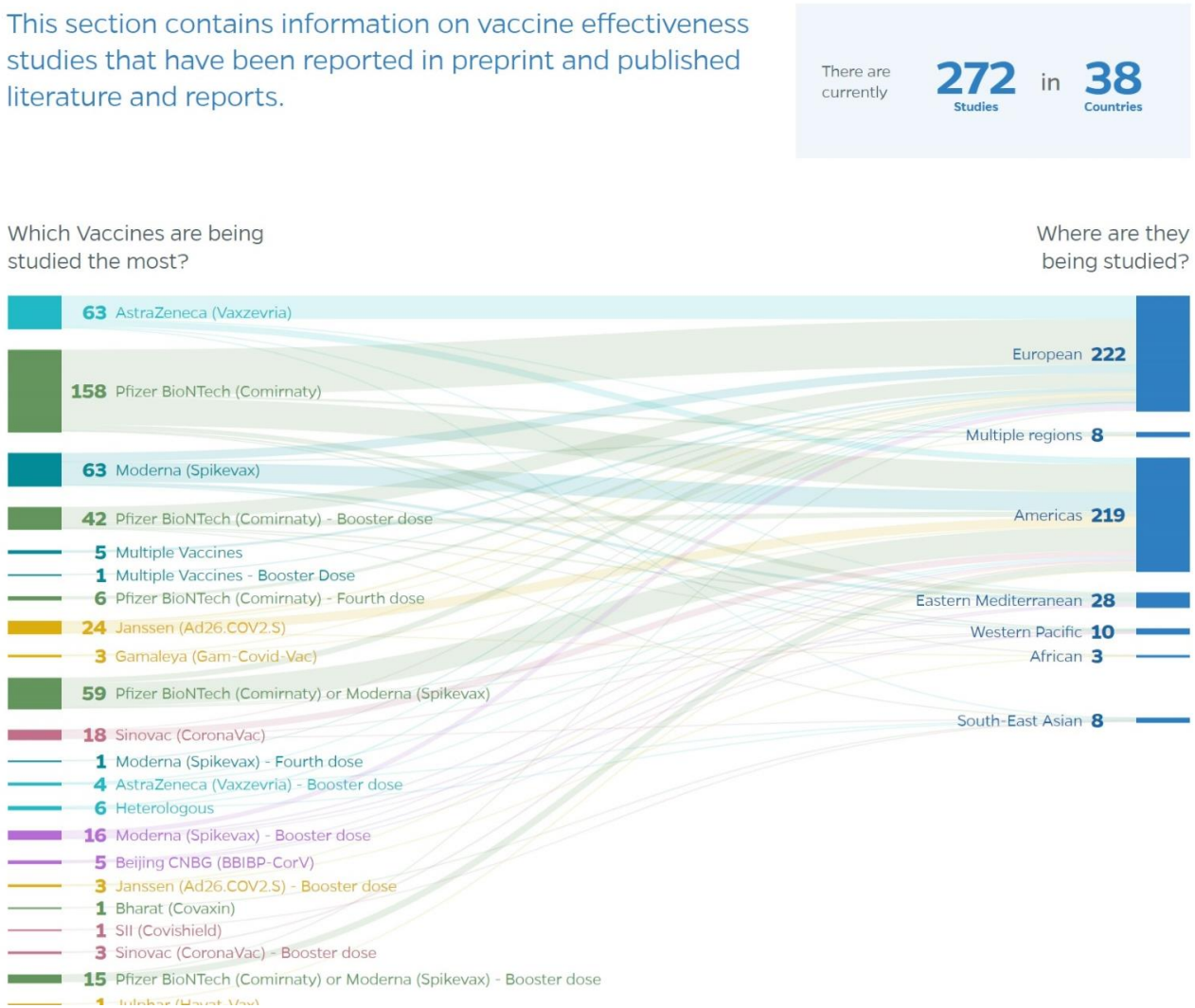
The impact of (known) prior SARS-CoV2 infection

- Failure to account for prior infection can induce non-comparability between vaccinated and unvaccinated

- How well documented is this?
- Testing rates and policies differ substantially in different settings
- NMC data (Access)
- In LMICs, health insurance data may prove useful (or more problems for HCWs in the public sector)
- The baseline seroprevalence in the population, if known, can help to quantify the expected bias on VE estimates.

Access, capacity and infrastructure to handle large data?

- [Effectiveness Studies | ViewHub (view-h](#)



Data sources

Source	Variables collected	Variables need to be collected	Notes
Electronic Vaccination Data System (EVDS)	vaccination status, date of vaccination, age, gender, occupation, geographical location	race, comorbidities, socio-economic status?	Primary source for confounders: Variables should be collected for everyone registered irrespective of vaccination status.
Notifiable Medical Conditions Sentinel Surveillance (NMCSS) of SARS-CoV-2 cases	Sample collection date, SARS-CoV-2 results (PCR)	symptoms data	Only SARS-CoV-2 positive results.
NICD testing data	sample collection date, SARS-CoV-2 result (PCR), age, geographical location	symptoms data	Both SARS-CoV-2 positive and negative results.
DATCOV (sentinel hospital surveillance for COVID-19 admissions and deaths)	date of hospitalization, date of death, age, gender, race, comorbidities, occupation, geographical location	socio-economic status?	Does not include deaths occurring outside hospitals. Timing of hospitalisation with respect to SARS-CoV-2 diagnosis date. Do all hospitals submit data?
Department of Home Affairs and SAMRC death registry	date of death, cause of death, age, gender, geographical location		Cause of death and timing of death with respect to SARS-CoV-2 diagnosis date.

Access, capacity and infrastructure to handle large data

- An effectiveness study using HCWs or select health insurance schemes may not provide valid VE estimates for the general population.
- A National Health Data Centre which consolidates all person level health data would allow estimation of VE against rare but severe outcomes
- Real world vaccine effectiveness evaluations involve large datasets
- Computationally intensive analysis
- Need servers with high capacity (massive storage, ultra-fast recovery, and high-end analytical capability)
- Data availability issues: Individual level data on vaccination in the national rollout
- Statistical capacity: Some with strong RCT background may lack exposure to real world effectiveness studies- collaboration is key

Guidelines on evaluating effectiveness of boosters

- There have been no updates to the WHO interim guidance (July 2021) to incorporate booster doses
- Formal guidance is crucial to standardize analytical approaches and improve robustness of findings

Pooling estimates

- Variability needs to be carefully considered
- Heterogeneity in terms of :
- Vaccine programmes/policies
- Health systems
- Care seeking behaviours
- Infection risk overall.