Introduction to study design

Nicki Tiffin
SANBI, UWC
ntiffin@sanbi.ac.za

Aims of this overview

- A refresher in study design

Subtext:

- Use the correct language to describe the research that has been done
- Learn the right way to interact with a statistician for effective data analysis
- Learn how to describe the data in a way that makes sense

What is epidemiology?

Epidemiology is the study of the patterns, causes, and effects of health and disease conditions in defined populations

Outcomes and exposures

- An **OUTCOME** is the main health outcome for which we are trying to understand causes and patterns. In many cases, this is a disease.

- An **EXPOSURE** is a factor that may have an effect on, or association with the OUTCOME.

  ≈ “risk factors”, “causes” or “determinants”.

Note about CAUSALITY

- **CAUSALITY** is when an exposure **causes** an outcome.
- In general, causality is very difficult to prove.
- In most cases, we can determine whether there is an **ASSOCIATION** between an exposure and an outcome

e.g. When there is greater exposure there is more disease…

But this could be the result of an external factor that is causing both the exposure and the outcome!

STUDY DESIGN

Five main types of study design:

- Case series
- Cross-sectional
- Case control
- Cohort
- Randomised control trial

We will discuss for each:

- Characteristics of study design
- Benefits and limitations of study design
- What can we calculate from each study design

Then we will discuss:

- What is bias?
- What is confounding?

1. CASE SERIES

- This is a group of clinical cases normally assembled by a clinician/s.

  (a case study or case report is for ONE patient)

- Examples include:

  A register of patients with a serious disease

  A group of patients with the same illness

- There is no control set, or comparison set, without the disease

1. CASE SERIES

BENEFITS

By exploring the nature of the disease and the medical/environmental histories of patients, clinicians gain insights into the natural history of a disease

=> Better understanding of the disease and its processes

1. CASE SERIES

LIMITATIONS

- No inferences can be made about the underlying causes of the disease. The study is:
- **OBSERVATIONAL** –observe and record "what is"
- **DESCRIPTIVE** – no analyses to understand relationships between **exposures** and **outcomes**.

2. CROSS-SECTIONAL STUDY

Data collected at a single time point from a population; a 'snapshot'

Measures in the defined **POPULATION**:

(i) the **PREVALENCE** of an OUTCOME (disease)
(ii) the **PREVALENCE** of an EXPOSURE

OBSERVATIONAL study design

– *assess "what is"*

2. CROSS-SECTIONAL STUDY

PREVALENCE = How much of the outcome or exposure is present in the population under study AT ONE POINT IN TIME.

Assess whether there is an **ASSOCIATION** (relationship) between the outcome and the exposure

= **ANALYTIC** study

2. CROSS-SECTIONAL STUDY

Study population

Disease status

D+, D-

Exposure status

E+, E-

+ Disease + Exposure	**a**
- Disease + Exposure	**b**
+ Disease - Exposure	**c**
- Disease - Exposure	**d**

2. CROSS-SECTIONAL STUDY

	D+	D-	*total*
E+	**a**	**b**	a+b
E-	**c**	**d**	c+d
total	a+c	b+d	a+b+c+d

WHAT CAN YOU CALCULATE?

- PREVALENCE of outcome

  = a+c/a+b+c+d

- PREVALENCE of outcome in exposed

  = a/a+b

- PREVALENCE of outcome in unexposed

  = c/c+d

you can do similar calculations for the exposure

2. CROSS-SECTIONAL STUDY

WHAT CAN YOU CALCULATE?

	D+	D-	*total*
E+	**a**	**b**	a+b
E-	**c**	**d**	c+d
total	a+c	b+d	a+b+c+d

PREVALENCE RATIO of the outcome in the exposed vs in the unexposed:

$$\frac{\text{Prevalence D+E+}}{\text{Prevalence D+E-}} = \frac{a/a+b}{c/c+d}$$

2. CROSS-SECTIONAL STUDY

WHAT CAN YOU CALCULATE?

Interpretation:

People who have the exposure are [*PREVALENCE RATIO*] times as likely as people who have not had the exposure, to have the disease.

2. CROSS-SECTIONAL STUDY

BENEFITS

- Very good for describing disease prevalence in a population
- Quick and inexpensive
- Can look at many exposures and diseases in one study

2. CROSS-SECTIONAL STUDY

LIMITATIONS

- More likely to sample prevalent cases for disease with long duration (length bias)

  e.g. HIV infection compared to flu

- Cannot assess whether disease or exposure came first

  e.g. high white blood cell count and leukaemia

2. CROSS-SECTIONAL STUDY

LIMITATIONS

- Does not include those who have died from the disease

  e.g. Ebola virus

- Only measures prevalent cases; cannot determine risk of developing disease

  e.g. asbestosis and asbestos exposure

3. CASE-CONTROL STUDY

- Compares a series of cases of a particular OUTCOME with a set of comparable controls WITHOUT THE OUTCOME
- Investigates exposures or characteristics of interest between the two groups

Other names: retrospective study, case-referent study, case-comparison study, case-compare study

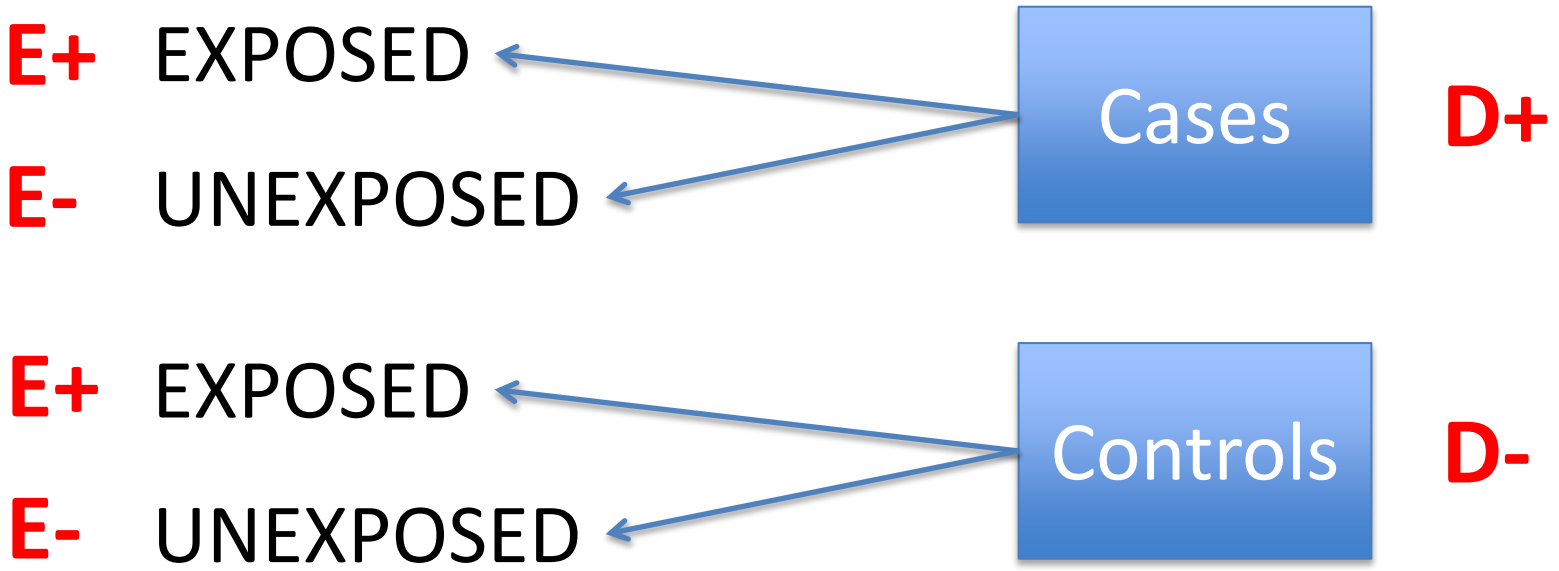
3. CASE-CONTROL STUDY

- **CONTROLS** should represent people who would have been selected as cases if they had the disease
  - i.e. from same population as cases
- Number of controls can vary: 4 controls per case is a rule of thumb but resources and power calculations determine actual number.

  "Matched" analyses are different: consult a statistician!!

3. CASE-CONTROL STUDY

CHARACTERISTICS



CASE-CONTROL STUDY

WHAT CAN YOU CALCULATE:

- You **CANNOT** calculate PREVALENCE i.e. how much disease there is in the population

This is because you have not sampled a population randomly, but rather selected those with the disease.

3. CASE-CONTROL STUDY

WHAT CAN YOU CALCULATE:

- You CAN look at what exposures are related to the disease = **OBSERVATIONAL** study

The distribution of exposure in cases and controls is compared = **ANALYTIC** study

3. CASE-CONTROL STUDY

WHAT CAN YOU CALCULATE:

	D+	D-	*total*
E+	**a**	**b**	a+b
E-	**c**	**d**	c+d
total	a+c	b+d	a+b+c+d

- Comparing exposures is calculated with an **ODDS RATIO**

3. CASE-CONTROL STUDY

WHAT CAN YOU CALCULATE:

	D+	D-	*total*
E+	**a**	**b**	a+b
E-	**c**	**d**	c+d
total	a+c	b+d	a+b+c+d

ODDS of exposure (E+) among cases (D+) = $\frac{a/a+c}{c/a+c}$ = **a/c**

ODDS of exposure (E+) among controls (D-) = $\frac{b/b+d}{d/b+d}$ = **b/d**

ODDS RATIO = $\frac{a/c}{b/d}$ = **ad/bc**

3. CASE-CONTROL STUDY

Interpretation:

People who have the outcome are *[ODDS RATIO]* times as likely as people who do not have the outcome to have the exposure.

3. CASE-CONTROL STUDY

BENEFITS

- Quick and not too expensive
- Good for rare outcomes – smaller numbers of participants required than for a cross-sectional study
- Can study multiple exposures

3. CASE-CONTROL STUDY

LIMITATIONS

- Controls can be difficult to select or find
- Selection bias

  e.g. clinicians select cases
- Recall bias

  e.g. those with disease more likely to remember exposures
- No temporality (what came first)
- Cannot estimate prevalence or risk of developing disease

Nicki Tiffin

4. COHORT STUDY

- A **cohort** is a group of people who share a common characteristic within a defined period (i.e. from the same population)

  e.g. a birth cohort is a group born in a certain time period

  Also called: follow-up study, longitudinal study, incidence study

4. COHORT STUDY

CHARACTERISTICS

- All participants at the start do not have the outcome.
- Some have the exposure, others do not.
- Follow up the cohort over time to see who develops the outcome = **OBSERVATIONAL** study

4. COHORT STUDY

CHARACTERISTICS

- Can be **PROSPECTIVE**

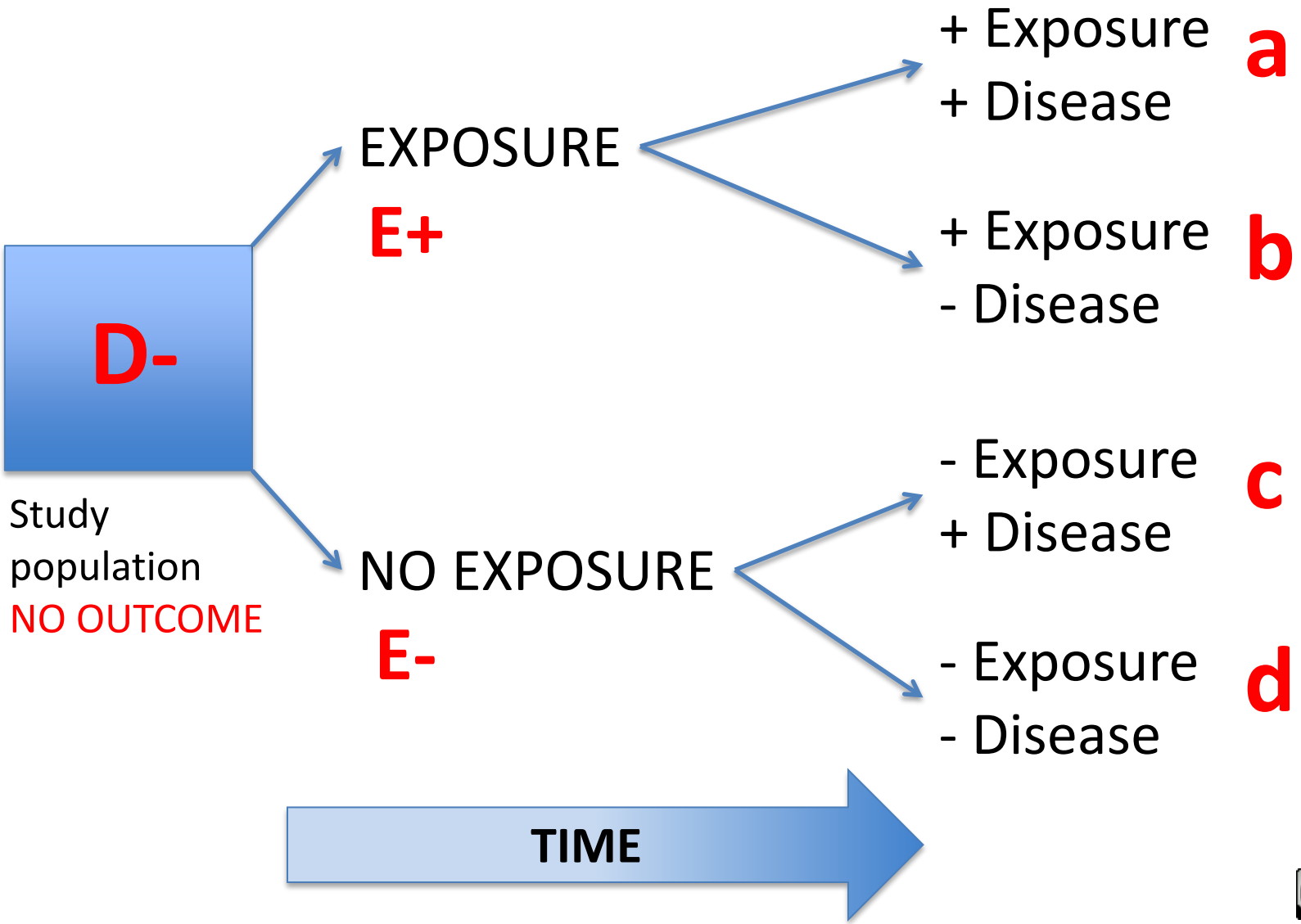
- Recruit participants and then follow forward

or **RETROSPECTIVE:**

- Investigate a cohort from some years back for a certain period of time.

e.g. Army personnel from 1990 to 2000: In 1991, 100 were transferred to a different base. Compared to 100 who remained at the original base, what was the risk of malaria infection for the two groups?

4. COHORT STUDY



4. COHORT STUDY

	D+	D-	*total*
E+	**a**	**b**	a+b
E-	**c**	**d**	c+d
total	a+c	b+d	a+b+c+d

WHAT CAN YOU CALCULATE?

- You <u>CANNOT</u> calculate prevalence of the outcome because all participants initially have no outcome.
- You can calculate the **INCIDENCE** of the outcome: the number of new cases with time

  e.g. 7 cases per 100 persons per year

  or: 7 cases per 100 person years

Nicki Tiffin

4. COHORT STUDY

	D+	D-	*total*
E+	**a**	**b**	a+b
E-	**c**	**d**	c+d
total	a+c	b+d	a+b+c+d

WHAT CAN YOU CALCULATE?

- Compare incidence in exposed and unexposed
  => **ANALYTIC** study

Incidence of outcome (D+) in exposed (E+), with time *(cases per person yrs)*

Incidence of outcome (D+) in unexposed (E-), with time *(cases per person yrs)*

- Rate ratio (risk ratio) *(units cancel out)*

$$= \frac{\text{incidence D+E+}}{\text{incidence D+E-}} = \frac{a/(a+b)}{c/(c+d)}$$

4. COHORT STUDY

Interpretation:

People with the exposure are *[rate ratio]* times as likely as people without the exposure to develop the outcome over *[the period of time]*

4. COHORT STUDY

LIMITATIONS

- Time consuming – sometimes many years
- Costly
- Not good for rare conditions (large sample size needed)
- Loss to follow-up
- Selection bias – exposures are not always random *e.g. Women with a high risk of breast cancer are unlikely to be prescribed HRT*

4. COHORT STUDY

BENEFITS

- **Temporality** (exposure precedes disease)
- Calculate risks and rates of incidence directly
- Can study multiple outcomes
- Useful for rare exposures

5. RANDOMISED CONTROL TRIAL

- In an RCT, the investigator intervenes = **EXPERIMENTAL** study design.
- Active effort to influence outcome through controlled exposures
  - *medication; vaccine, behaviour change…etc*
- Attempt to define relationship between exposure under control, and outcome = **ANALYTIC**

5. RANDOMISED CONTROL TRIAL

CHARACTERISTICS

- **PROSPECTIVE**: recruit participants, then apply intervention
- **RANDOM ASSIGNMENT** of participants to exposure to avoid bias
- **DOUBLE BLINDING**: where possible, neither investigator nor participant know whether they are receiving treatment or control

5. RANDOMISED CONTROL TRIAL

WHAT CAN YOU CALCULATE?

- The calculations are like those for the COHORT study; i.e.
  - All participants start out without the outcome
  - Some have the exposure
  - Some experience the outcome during the time period of the study

Nicki Tiffin

5. RANDOMISED CONTROL TRIAL

WHAT CAN YOU CALCULATE?

- You CANNOT calculate prevalence of the outcome because all participants initially have no outcome.
- You can calculate the **INCIDENCE** of the outcome: the number of new cases with time

  e.g. 7 cases per 100 persons per year

  or: 7 cases per 100 person years

Nicki Tiffin

5. RANDOMISED CONTROL TRIAL

WHAT CAN YOU CALCULATE?

- Compare incidence in exposed and unexposed
  => **ANALYTIC** study

Incidence of outcome (D+) in exposed (E+), with time *(cases per person yrs)*
Incidence of outcome (D+) in unexposed (E-), with time *(cases per person yrs)*

- Rate ratio *(units cancel out)*

$$= \frac{\text{incidence D+E+}}{\text{incidence D+E-}}$$

Nicki Tiffin

5. RANDOMISED CONTROL TRIAL

Interpretation:

People with the intervention are *[rate ratio]* times as likely as people without the intervention to develop the outcome over *[the period of time].*

Nicki Tiffin

5. RANDOMISED CONTROL TRIAL

BENEFITS

- Provide the strongest information about whether interventions are effective
- If randomisation is successful, can prevent confounding and selection bias
- Temporality is confirmed (exposure precedes outcome)

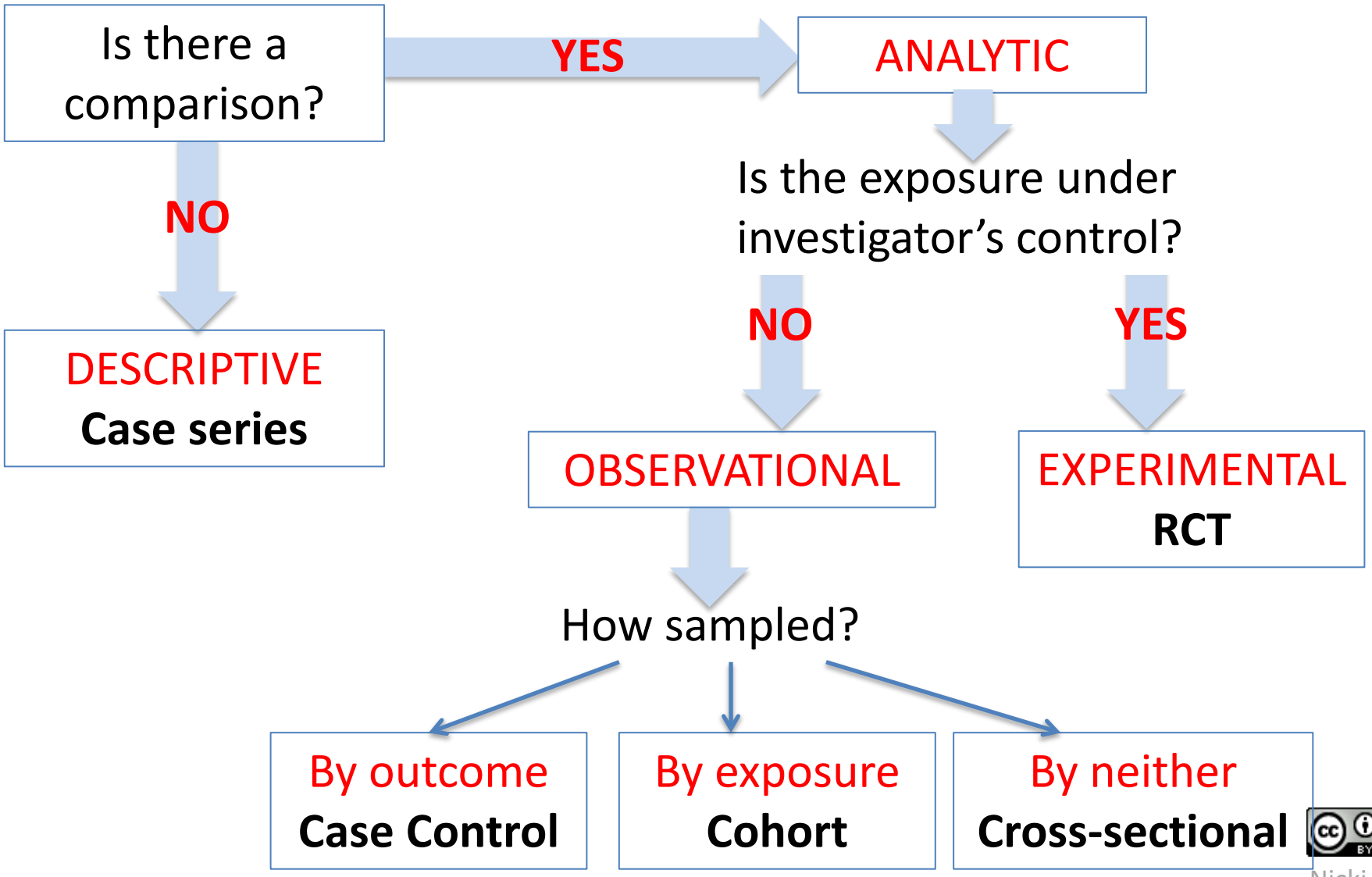
Nicki Tiffin

5. RANDOMISED CONTROL TRIAL

LIMITATIONS

- Complex and time-consuming
- Ethical considerations
  - *cannot apply an intervention or exposure that may be harmful*
  - *cannot withhold a treatment that may be beneficial*
- Bias or loss to follow-up

Summary of study design



What is BIAS?

VALIDITY is the 'truth' of the finding

Internal validity: what has been found for the sample is true for the study population

External validity: what has been found for the sample is true for other populations as well as the study population

Bias may compromise the validity of the study.

Nicki Tiffin

MEASUREMENT BIAS

Measurement bias arises out of misclassification or error in information obtained in the study, particularly if it affects one comparison group rather than another.

Examples:

recall bias

Participants remember differently depending on whether they have the outcome

ascertainment bias

Failure to represent all types of people in the population under study

calibration errors, faulty measuring instruments

laboratory error

Nicki Tiffin

SELECTION BIAS

Due to selection of participants who are different in some systematic way from the true underlying population of interest.

Examples:

volunteer bias

e.g. health conscious people more likely to join a health-based study

non-response bias

e.g. those indulging in risky sexual behaviour might be less likely to provide information

healthy worker bias

i.e. those who are ill are at home

loss to follow-up bias

e.g. Heavy drug users may be less likely to return for follow-up

How do we deal with bias?

- Bias results from BAD STUDY DESIGN

  "It is YOUR fault"
- Once the study has been conducted, you can't do anything about bias except acknowledge it in your discussion of the results
- Consider all the possible forms of bias BEFORE you start the study

Nicki Tiffin

What is CONFOUNDING?

Sometimes an unmeasured, independent exposure may be affecting the outcome and the exposure of interest

Example: "People who drink coffee are more likely to have lung cancer"

In fact: people who drink coffee are more likely to smoke; and people who smoke are more likely to have lung cancer

SMOKING is a confounder!

Nicki Tiffin

How do we deal with confounders?

- Confounders are not an error. They are real, and their effects are real
- There are statistical approaches to factor in confounders when analysing the relationship between exposure and outcome
- When designing a study, include data collection for possible confounders

Acknowledgement: Examples borrowed from Prof Erhlich at UCT

Study design examples

1. In a community study, women with an unemployed man in the house are more likely to report "family violence" (of whatever sort) than women in households where there is no unemployed man.

- STUDY DESIGN?

Acknowledgement: Examples borrowed from Prof Erhlich at UCT

Study design examples

1. In a community study, women with an unemployed man in the house are more likely to report "family violence" (of whatever sort) than women in households where there is no unemployed man.

- STUDY DESIGN? Cross-sectional

Acknowledgement: Examples borrowed from Prof Erhlich at UCT

Study design examples

1. In a community study, women with an unemployed man in the house are more likely to report "family violence" (of whatever sort) than women in households where there is no unemployed man.

- STUDY DESIGN? Cross-sectional
- SOURCES OF BIAS?

Acknowledgement: Examples borrowed from Prof Erhlich at UCT

Study design examples

1. In a community study, women with an unemployed man in the house are more likely to report "family violence" (of whatever sort) than women in households where there is no unemployed man.

- STUDY DESIGN? Cross-sectional
- SOURCES OF BIAS?

  Self-reporting of violence may not be accurate

  Definition of 'employment' may vary

Acknowledgement: Examples borrowed from Prof Erhlich at UCT

Study design examples

2. Women recently diagnosed with breast cancer are no more likely, when asked, to recall having previously used the injection form of contraception than women without breast cancer.

- STUDY DESIGN?

Acknowledgement: Examples borrowed from Prof Erhlich at UCT

Study design examples

2. Women recently diagnosed with breast cancer are no more likely, when asked, to recall having previously used the injection form of contraception than women without breast cancer.

- STUDY DESIGN? Case control study

Acknowledgement: Examples borrowed from Prof Erhlich at UCT

Study design examples

2. Women recently diagnosed with breast cancer are no more likely, when asked, to recall having previously used the injection form of contraception than women without breast cancer.

- STUDY DESIGN? Case control study
- SOURCES OF BIAS?

Acknowledgement: Examples borrowed from Prof Erhlich at UCT

Study design examples

2. Women recently diagnosed with breast cancer are no more likely, when asked, to recall having previously used the injection form of contraception than women without breast cancer.

- STUDY DESIGN? Case control study
- SOURCES OF BIAS?

  Recall bias – people who are ill are more likely to remember exposures

Acknowledgement: Examples borrowed from Prof Erhlich at UCT

Study design examples

3. Children who are admitted to hospital with pesticide poisoning are more likely to have poor scores on neurocognitive tests when tested two years later than children admitted to the hospital at the same time with other diagnoses.

- STUDY DESIGN?

Acknowledgement: Examples borrowed from Prof Erhlich at UCT

Study design examples

3. Children who are admitted to hospital with pesticide poisoning are more likely to have poor scores on neurocognitive tests when tested two years later than children admitted to the hospital at the same time with other diagnoses.

- STUDY DESIGN? Cohort study

Acknowledgement: Examples borrowed from Prof Erhlich at UCT

Study design examples

3. Children who are admitted to hospital with pesticide poisoning are more likely to have poor scores on neurocognitive tests when tested two years later than children admitted to the hospital at the same time with other diagnoses.

- STUDY DESIGN? Cohort study
- SOURCES OF BIAS?

Acknowledgement: Examples borrowed from Prof Erhlich at UCT

Study design examples

3. Children who are admitted to hospital with pesticide poisoning are more likely to have poor scores on neurocognitive tests when tested two years later than children admitted to the hospital at the same time with other diagnoses.

- STUDY DESIGN? Cohort study
- SOURCES OF BIAS?

  The type of tests can introduce bias: cross-cultural, language, observer bias if administered tests.

  Definition of 'pesticide poisoning' could vary

Thank you

ntiffin@uwc.ac.za

South African National Bioinformatics Institute
University of the Western Cape
South Africa

Acknowledging funding from:
Bill & Melinda Gates Foundation, Calestous Juma Fellowship
UKRI/MRC