UNIVERSITY OF ™
KWAZULU-NATAL

INYUVESI
YAKWAZULU-NATALI

Bias and Confounding

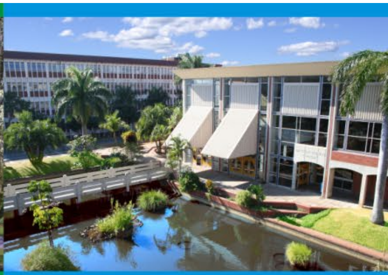
Department of Public Health Medicine

EDGEWOOD CAMPUS

HOWARD COLLEGE CAMPUS

NELSON R MANDELA SCHOOL OF MEDICINE

PIETERMARITZBURG CAMPUS

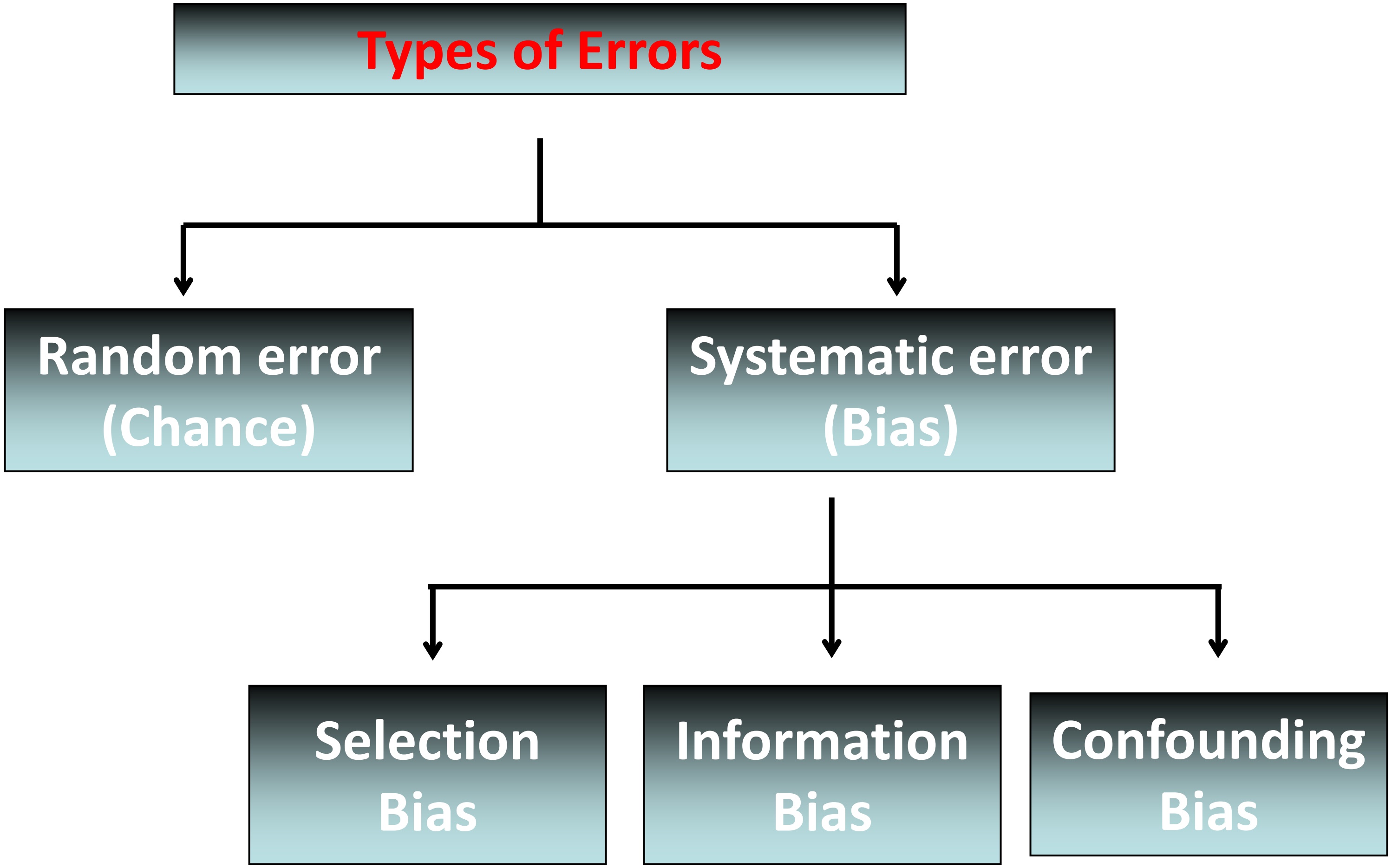

WESTVILLE CAMPUS

UKZN INSPIRING GREATNESS

Goal of epidemiologic investigation

- To determine precise and valid estimate of effect for an exposure on an outcome,
- The clear and present danger to achieve this goal is error.
  - Random error and systematic error (bias)
- When planning/designing a study, preventing and minimizing error embedded in the design/planning phase
- Some of the bias if not prevented (during the design and conduct of the study) CANNOT never be removed by statistical analysis.

Sources of Error in Epidemiologic Research



Types of Errors

Random error (Chance)

Systematic error (Bias)

Selection Bias

Information Bias

Confounding Bias

Random error

- Collect information from a sample to estimate a statistic (measure of occurrence) or association (measure of effect) for the whole population
- Random samples from the same population will give different results – due to chance–
- Results may not be same as true population result

What is confounding

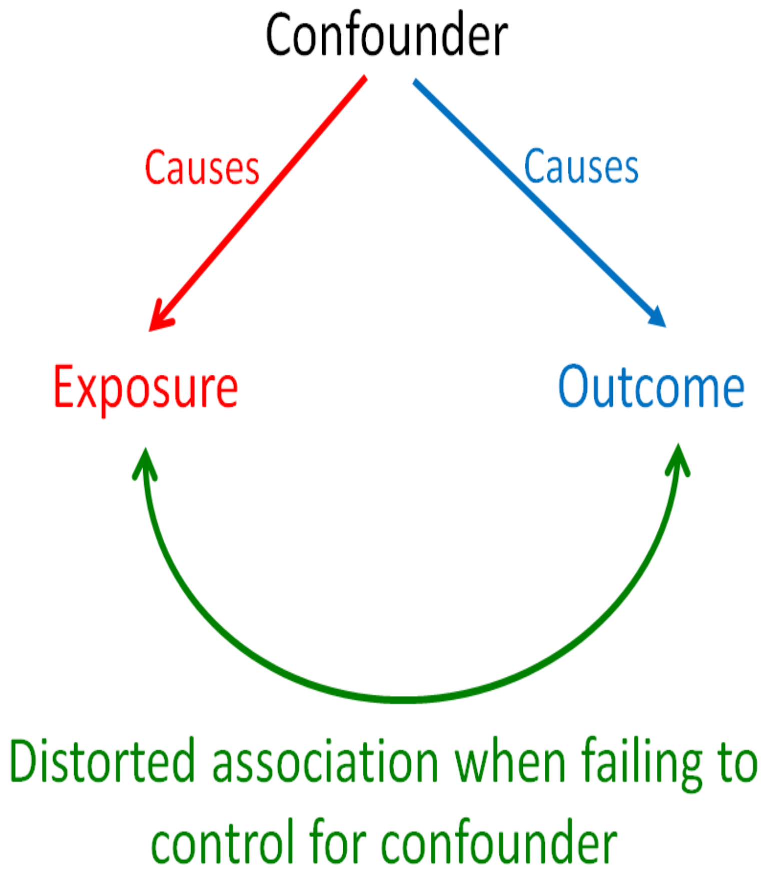
- Confounding occurs when an association (or lack of) is distorted by another factor
- Is there an alternative explanation for the observed exposure-disease association?
- In observational studies, people who are exposed to a particular risk factor may also have other characteristics in common that influence their risk of the disease

Classical definition of Confounder

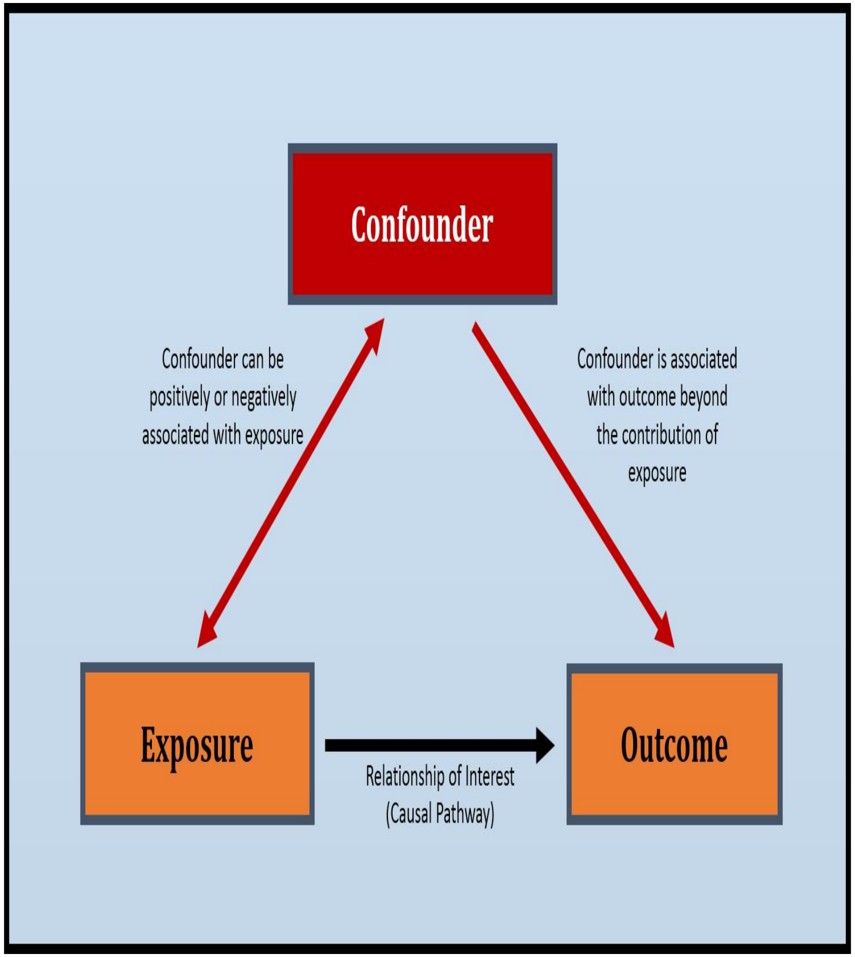
- A confounding factor therefore must have an effect, and it must be imbalanced between the exposure groups."
- "A confounder must be associated with the disease
  - (either as a cause or as a proxy for a cause, but not as an effect of the disease)."
- "A confounder must be associated with the exposure."
- "A confounder must not be on causal pathway."

Rothman (2012)

Confounder



Distorted association when failing to control for confounder


Confounder
Confounder can be positively or negatively associated with exposure
Confounder is associated with outcome beyond the contribution of exposure
Exposure
Outcome
Relationship of Interest (Causal Pathway)

Limitations to classical definition

- It applies only to the classical condition in which there is just one variable to consider.
- Reality, several variables are considered at once while keeping them distinct, as when some have been measured and others have not.
  - the status of a variable as a confounder, as well as the degree and direction of confounding, can change drastically according to which variables are controlled.
- One consequence is that control of a variable that meets the above definition can at times introduce more confounding than it removes

Maldonado and Greenland, 2002

Confounders

- A confounder is a variable, which, if not “controlled” or “adjusted” for during the design phase of the study and/or in the analysis phase, would result in a biased estimate of the causal parameter of interest.
    - Generally, known and suspected risk factors for the outcome of interest are potential confounders
- Variables that when stratified on or adjusted for will eliminate (or diminish) the spurious component of the association between exposure and disease.

Spiegelman and Zhou, 2018

How Do I Know What Might Be a Potential Confounder for my Specific Research Question?

1. Know your subject area.
   - "Causal knowledge as a prerequisite for confounding evaluation."-Kleinbaum et al. 1982
   - "Knowledge of the causal structure is a prerequisite to accurately label a variable as a confounder."-Hernán, 2012

2. Complete a comprehensive literature review and read, read, read.
   - Previous research will help you to identify "known" confounders.

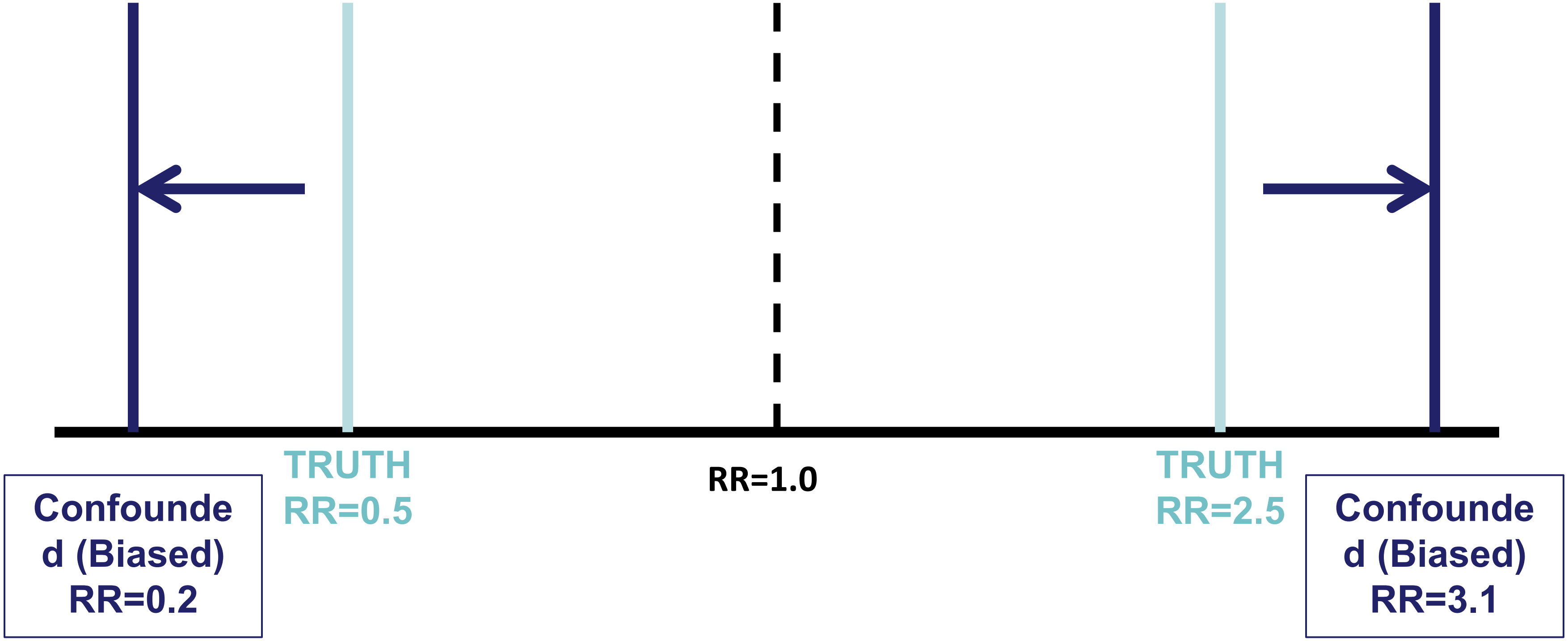
3. "Historical" confounders
   - Some variables are always considered potential confounders (age, sex, race/ethnicity).

Impact of confounding

- Observed relationship can be attributable, totally or in part to effect of confounder

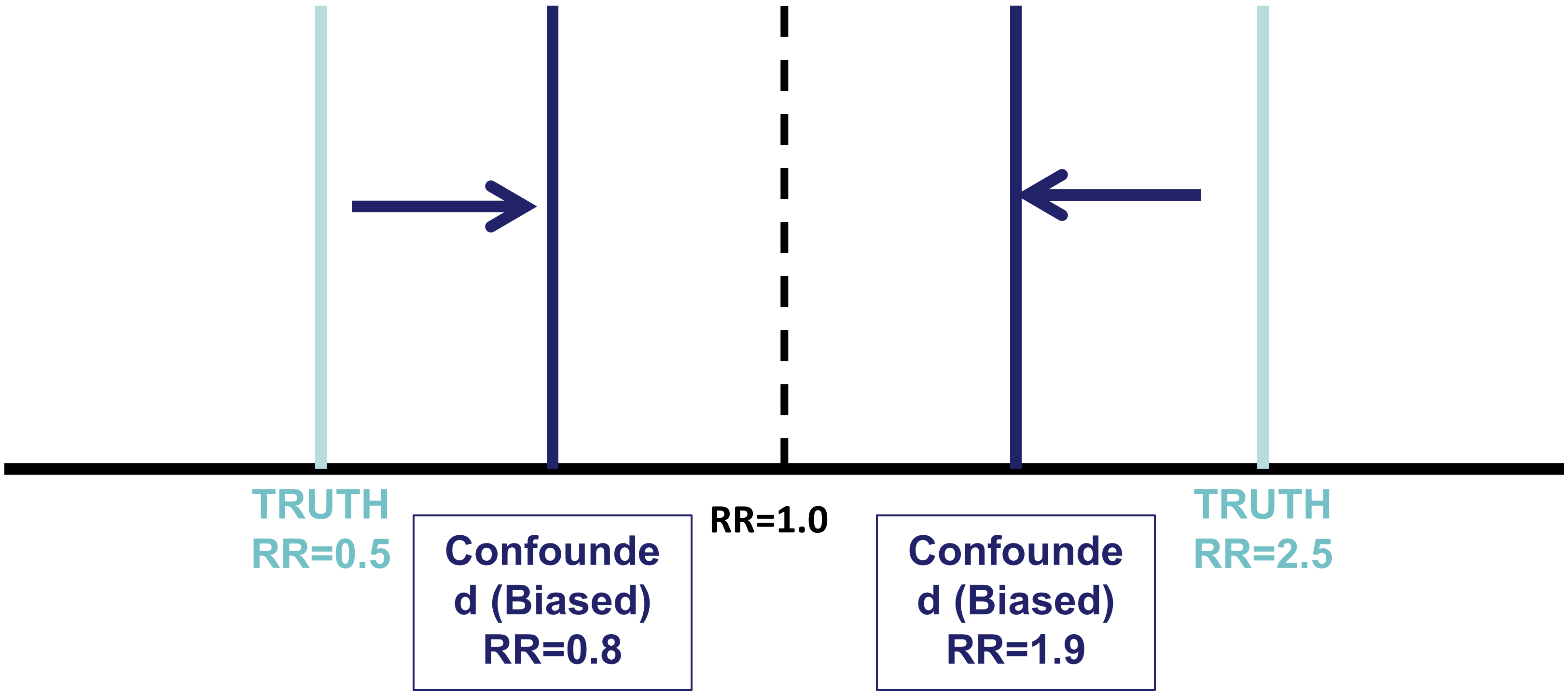
- Lead to overestimation or underestimation of true association/effect
  - Confounding can bias results away from the null
  - Confounding can bias results towards the null
- Change the direction of true association/effect

Confounding can bias results away from the null



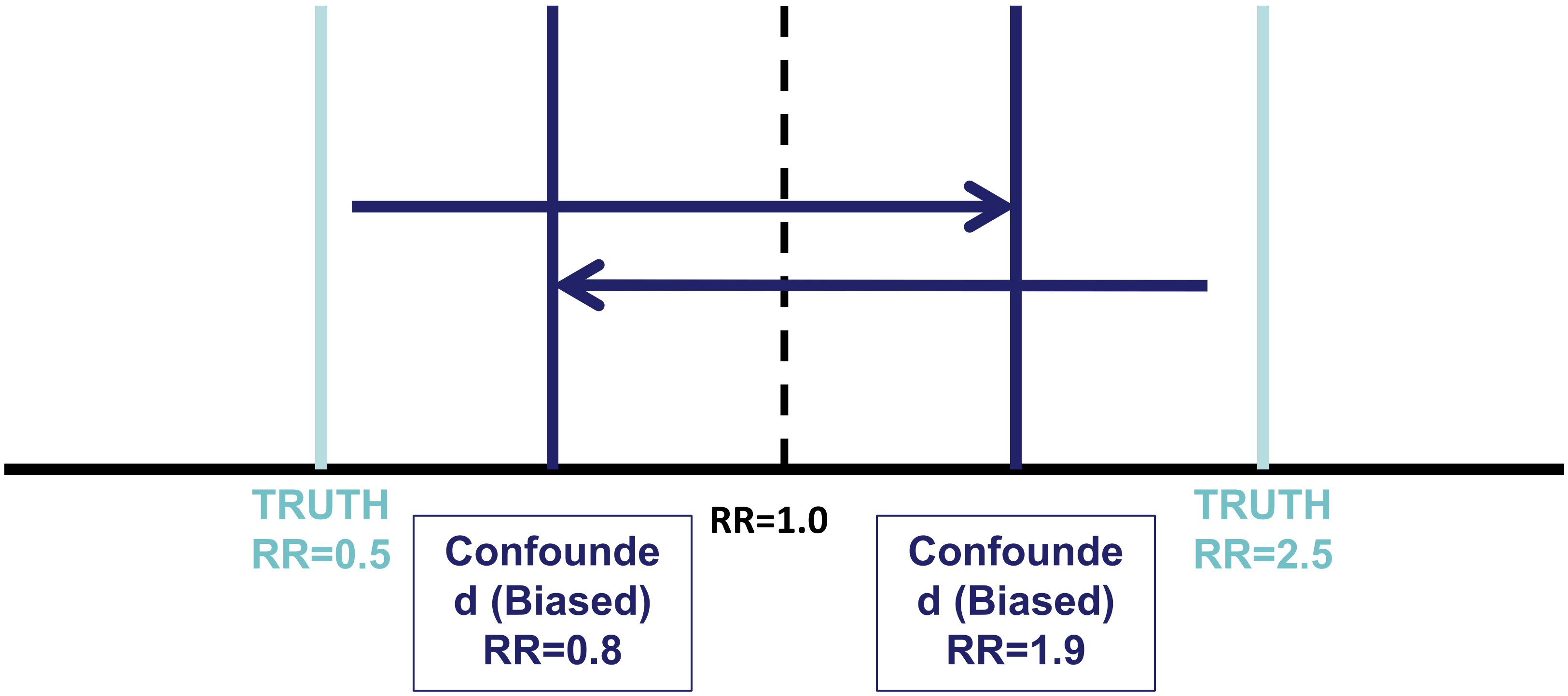
Positive Confounding – exaggerate (overestimate) the true association/effect

Confounding can bias results towards the null



Negative Confounding – hide/minimize (underestimate) the true association/effect

Confounding can change the direction of the true effect



Inverse association/effect

Confounding can be “Controlled” or “Adjusted”

Design Phase

- Randomization
- Restriction
- Matching

Analysis Phase

- Stratified Analysis
- Multivariate Analysis

The goal of these methods is to “break” the association between the confounder and the exposure.

Randomization

Randomly allocate study subjects to treatment groups so each subject has an equal chance of being assigned to the treatment or comparison group

- With an adequate number of subjects, randomization ensures baseline comparability of exposed and unexposed groups in terms of both known and unknown confounders.

- This only works when:
  - Study is large enough
  - Treatment assignment is not influenced by investigator

- When doesn't work, confounding must be addressed in the analysis

Randomization

Strengths

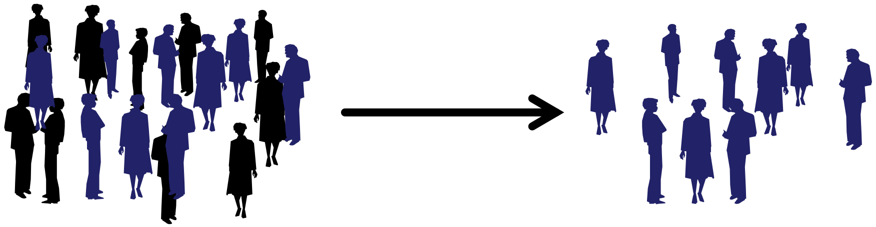
- No limit on the number of confounders that can be controlled for
- Do not need information about unknown confounders
  - Do not need to know what they are
  - Do not need to measure them

Limitations

- Limited to experimental studies
- Less effective with smaller sample size

Restriction

Limit study to people who are within one category of the confounder.



Examples:

- If sex is a confounder, limit your study groups to only men or only women.
- If age is a confounder, limit your study population to only persons , say >65 or <65.
- If smoking is a confounder, limit your study population to only non-smokers or only smokers.

Restriction

Strengths

- Simple – conceptually and practically
- Effective control of characteristics being restricted

Limitations

- Only possible for known, measured confounders
- Incomplete control for confounding (residual confounding) if restriction is not narrow enough
- Cannot evaluate restricted variable
- Nonetheless, restriction on many factors can;
    - Limits sample size
    - Limits generalizability of results

Matching

Select study subjects so that confounders are distributed identically among the exposed and unexposed (cohort study) or cases and controls (case-control study).

Matching is a form of stratification
  a matched set = a stratum, e.g., twins

Example: You are interested in evaluating the association between bacterial and viral microbiota and LRTIs, but you know that age and sex are confounders.

Matching

Strengths

- Simple and effective control of characteristics being matched
- Useful for variables that are complex or difficult to capture (e.g., neighborhood)

Limitations

- Only possible for known, measured confounders
- Can be difficult, expensive, and time-consuming to find appropriate matches
- Cannot evaluate matched variable
- Matches may become difficult or impossible to find if one attempts to match on more than a few factors.

Confounding can be "Controlled" or "Adjusted"

Design Phase

- **Randomization**
- **Restriction**
- **Matching**

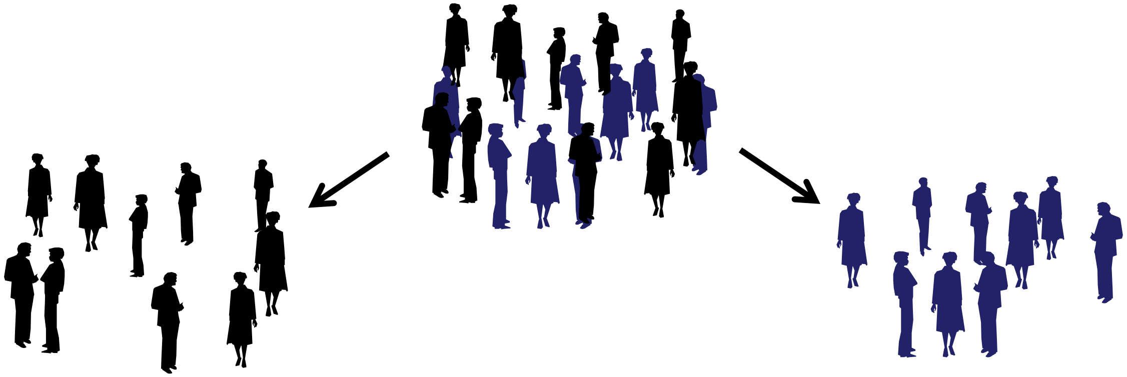
Analysis Phase

- **Stratified Analysis**
- **Multivariate Analysis**

Design-based methods are often infeasible or insufficient to produce exchangeability.

Stratification

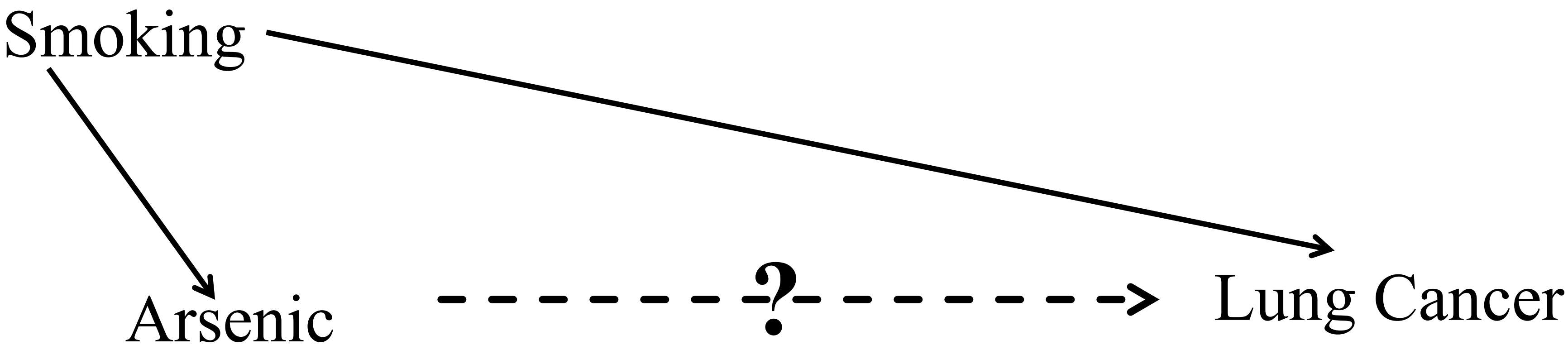
Stratify (separate) your study population into subgroups where one group has the confounder characteristic and one group does not. Then calculate a measure of association for each subgroup.



Example:

- Research Question: Does arsenic exposure increase the risk of lung cancer?
- Smoking is a confounder of the relation between arsenic and lung cancer.
- If differences between smokers and non-smokers confuse the relationship between arsenic and lung cancer, then you can examine the relationship <u>*separately*</u> in smokers and non-smokers.

Example of Stratification



Smoking is confounder so we will assess the relationship between arsenic and lung cancer separately among smokers and non smokers.

Smokers
Non Smokers

The stratum-specific results are not confounded by smoking.

Stratified Analysis:

Cohort study of the association between arsenic and lung cancer

Calculate Crude Measure of Association

	Arsenic (+)	Arsenic (-)
Lung cancer	**500**	**50**
Total	**1400**	**600**

Crude RR = 4.3

Concern: Smoking may be a confounder of the association

Stratified Analysis:

Divide Subjects into Strata of Smokers and Non-Smokers

	Arsenic (+)	Arsenic (-)
Lung cancer	500	50
Total	1400	600

Non-Smoker

	Arsenic (+)	Arsenic (-)
Lung cancer	110	
Total		

Smoker

	Arsenic (+)	Arsenic (-)
Lung cancer	390	
Total		

Stratified Analysis:

Calculate Stratum-Specific Measures

	Arsenic (+)	Arsenic (-)
Lung cancer	500	50
Total	1400	600

Crude
RR = 4.3

Non-Smoker

	Arsenic (+)	Arsenic (-)
Lung cancer	110	35
Total	200	340

Stratum-specific
RR = 5.3

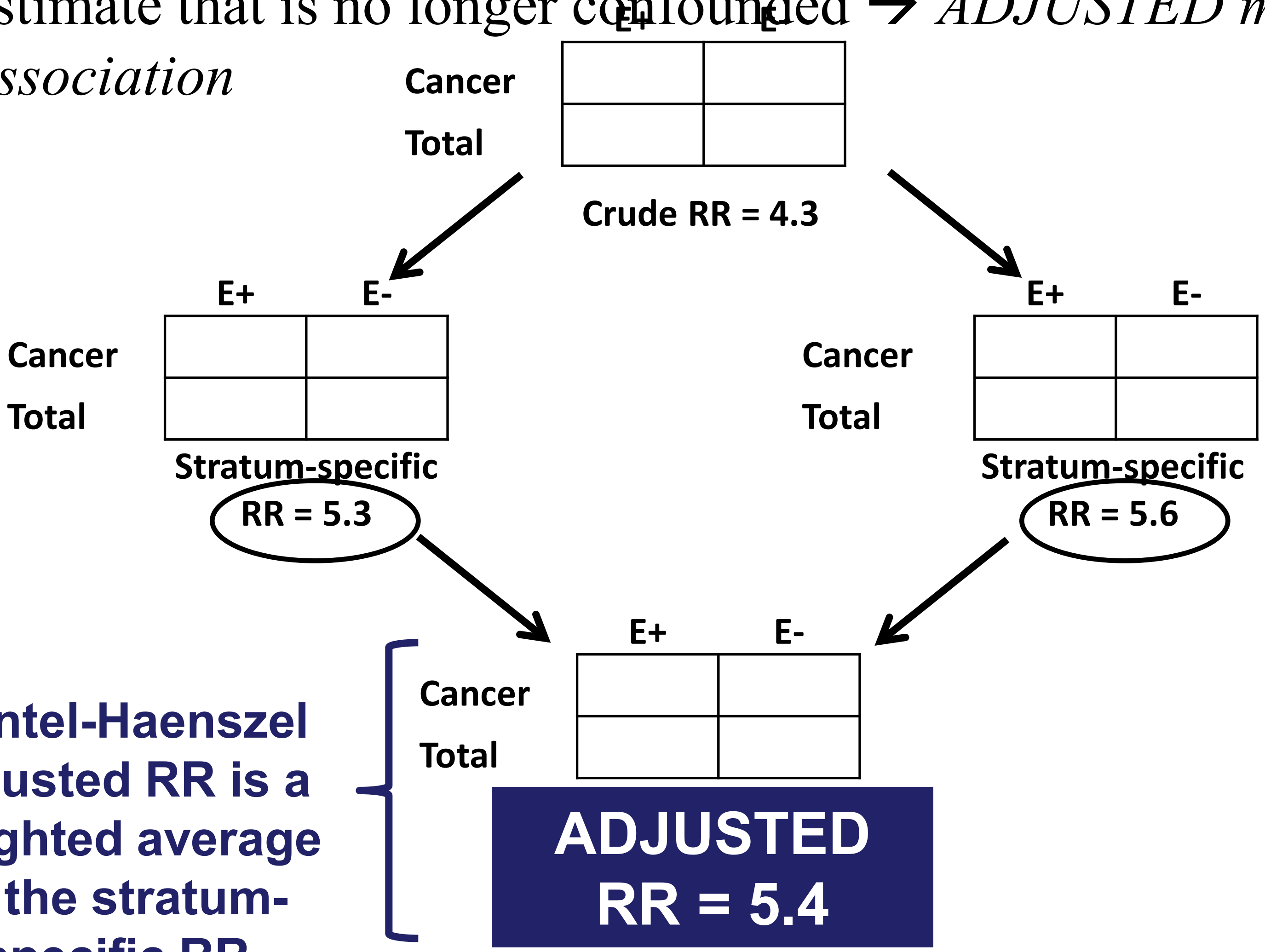
Smoker

	Arsenic (+)	Arsenic (-)
Lung cancer	390	15
Total	1200	260

Stratum-specific
RR = 5.6

Stratified Analysis:

Stratum-specific estimates are combined into a single summary estimate that is no longer confounded → *ADJUSTED measure of association*



E+ E-

Cancer

Total

Crude RR = 4.3

E+ E-

Cancer

Total

Stratum-specific RR = 5.3

E+ E-

Cancer

Total

Stratum-specific RR = 5.6

E+ E-

Cancer

Total

ADJUSTED RR = 5.4

Mantel-Haenszel Adjusted RR is a weighted average of the stratum-specific RR estimates

Stratified Analysis: Is Crude =Adjusted?

To determine whether confounding has occurred, compare a crude (unadjusted) measure of association to a measure that has been adjusted for confounding.

- If crude RR = adjusted RR→ no confounding is present
- If crude RR ≠ adjusted RR → confounding is present

Adjusted RR (4.3) ≠ Crude RR (5.4) → confounding was present

Stratification

Strengths

- Straightforward and easy to perform
- Effective control of characteristics being stratified

Limitations

- Difficult to control for many confounders simultaneously due to sparse data problems
- Difficult presentation, esp. if many confounders
- Continuous variables not easily stratified

Modeling/Regression

- Involves construction of statistical model (requires computer) that describes the association between exposure, disease, and confounder

- Advantage: Simultaneously adjusts for several variables

- Disadvantage: Difficult to conceptualize, data need to fit into an available statistical model (assumptions needed)

- Examples:
  - Multiple linear regression for continuous outcomes
  - Multivariate logistic regression for dichotomous outcomes
  - Cox proportional hazards model for longitudinal data

Limitations: Control of Confounding

- Most methods can only control known potential confounders
- Very hard to control unmeasured confounders
- Randomization: can control for unknown potential confounders

Residual Confounding

- Confounding that persists despite efforts to control or adjust for confounding

- Sources:
  - Confounders for which no data were collected
  - Inaccurate data on a confounder
  - Use of broad categories of a confounder in your analysis
    - Wide age groups, ever smokers vs. never smokers

- Residual confounding should be acknowledged and addressed in the discussion section of a published paper.

Summary

- Confounding occurs when the effect of the exposure is mixed together with the effect of another variable
- A systematic bias that results in an incorrect estimate
- Causal knowledge as a prerequisite for confounding evaluation
- Attempts to control confounding can be undertaken in the study design and/or analysis phases.

BIAS

- Can affect all types of studies but observational studies especially those based on routinely collected data are particularly vulnerable

- Lead to overestimation or underestimation of true association/effect
- bias results away from the null
- bias results towards the null

- The direction depends on a number of factors,
- measurement error on exposure, disease, or other study variables

TYPES OF BIAS

- Selection bias:
- Results from the selection and retention of the study population
- Information bias:
- Results from poor measurement of study variables – exposure, outcome,
- Measurement error or misclassification – Differential or non-differential

INFORMATION BIAS

- Occurs due to poor measurement (classification) of study variables (exposure, outcome, confounders)
- Particularly problematic when using secondary data Primary data: collected for research purposes.

  Secondary data: collected for clinical/administrative

- Distinguish two basic types of information bias:

➢ Non-differential - Misclassification between groups is approximately equal

➢ Differential - Amount of misclassification differs between groups

Information bias

- Imperfect tools for data collection;
  - Diagnostic
  - Self reports
- Data collected by people from people
- Measurement errors occur when we collect data;
  - Outcome
  - Exposure
  - Other study variables.

Sources of information bias

- Cultural differences
- Poorly worded questions
- Faulty recall
- Observer bias
- Multiple observers

One person’s diarrhoea is another’s upset stomach

- What survey question meant:
  - Three or more loose stools
  - In a 24-hour period

- What participants may have thought survey question meant:
  - Watery stools over several days
  - Dehydration
  - Fever

Information Bias

Information can be misclassified in two ways:

- Differential, when information collected from one group is accurate but information collected from the other group is inaccurate
- Non- differential, when information collected from both groups is inaccurate.

Diffrential Misclassification

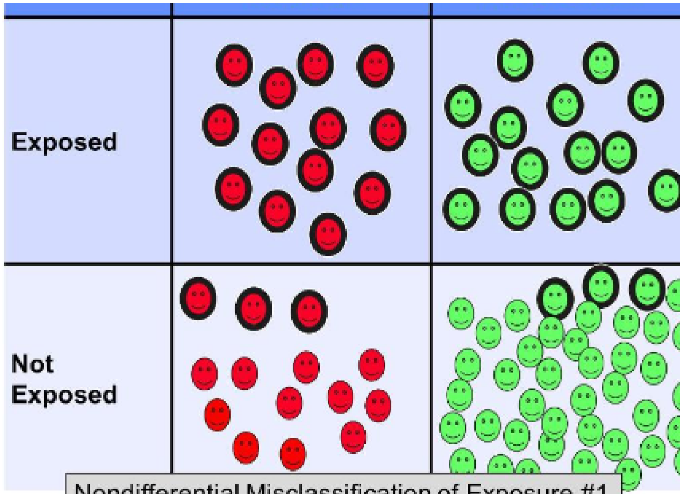
- *Emphysema is diagnosed more frequently in smokers than in non-smokers. However, smokers may visit the doctor more often for other conditions (e.g. bronchitis) than non-smokers, which means that a reason smokers could be diagnosed with emphysema more often is simply because they go to the doctor more often — not because they actually have higher odds of getting the disease. Unless steps are taken to control for this possibility, emphysema will be under-diagnosed in non-smokers, which is a classification error because the diagnosis is related to the variable "how often smokers visit the doctor, versus non-smokers".*
- https://www.statisticshowto.com/non-differential-misclassification/

Non-differential Misclasification

- Case-control study was conducted to examine the association between a high sugar diet and Type 2 DM. Subjects with T2DM and controls without T2DM might be recruited and asked to complete questionnaires about their dietary habits.
- It is difficult to assess dietary sugar content accurately from questionnaires, so it would not be surprising if there were errors in classification of exposure.
- However, it is likely that in this scenario the misclassification would occur with more or less equal frequency regardless of the eventual disease status. Nondifferential misclassification of a dichotomous exposure always biases toward the null. In other words, if there is an association, it tends to minimize it regardless of whether it is a positive or a negative association.

Non-differential Misclassifiction

- Disease status is correctly classified,
- Exposed subjects are incorrectly classified as non-exposed.
- result in bias toward the null. True odds ratio for the association between high sugar diet and T2DM is 5.0, but if about 20% of the exposed subjects were misclassified as 'not exposed' in both disease groups, the biased estimate might give an odds ratio of, say, 2.4. In other words, it resulted in bias toward the null.



Nondifferential Misclassification of Exposure #1

Preventing information bias

- Cannot undo it
- Design phase: work hard to prevent information bias
  - Carefully designing the study questionnaire
    - Writing understandable questions-both an art and a science (psychometrics)
    - Avoid open-ended, ambiguous questions
    - Blinding (interviewers and/or participants)
    - interviewer administered vs self-administered
  - use biological measurements and preexisting records for the necessary study data
  - Pilot test – then do more pilots
- Analysis phase: Assess effects of information bias using sensitivity analysis

Selection bias

- Selection bias is introduced when the study population does not represent the target population.

- Selection bias can be controlled
  - when the variables influencing selection are measured on all study subjects and either;

(a) they are antecedents of both exposure and outcome or

(b) the joint distribution of these variables (plus exposure and outcome) is known in the whole target population, or

(c) the selection probabilities for each level of these variables are known.

Selection Bias

- Type of systematic error which results from– procedures used to select study participants. E.g., TB prevalence survey: convenience sampling at hospital so that only those at highest risk of having TB participate à higher (biased) estimate of TB burden for city.
- Factors that influence participation/retention in the study E.g., TB prevalence survey: waived written consent (thumbprint) as undocumented migrant workers would not participate
- LTFU important in longitudinal studies and can also bias RCTs!
- Bias of the estimated effect of an exposure on outcome due to conditioning on a common effect of the exposure and the outcome (collider bias)

Selection bias

It can be introduced at any stage of a research study:

- Design stage
  - –bad definition of the eligible population,
- Sampling stage
  - –lack of accuracy of sampling frame,
- Implementation stage.
  - –Non-response/refusal/loss to follow up
  - –Missing data

Types of Selection bias

Refusal bias

- non-responders or those declining study participation differ from respondents with respect to exposure or outcome

Healthy worker survivor effect

- Occupational exposure; general population for comparison
- People who can work are healthier than the general population

Types of Selection bias

Detection bias

- Those initially recruited into screening programs do not represent the general population; screening program may show benefit initially but not long-term
- asymptomatic or mild disease of interest are more likely to be detected in persons with frequent medical surveillance
  - women taking estrogens likely to have endometrial cancer detected
  - hypertension and airline pilots
- more thorough examinations on exposed individuals can create detection bias
  - Health care providers may be influenced topic of study

Types of Selection bias

Berkson’s bias

- Hospital-based case control studies
- Exposure increases the risk of hospitalization
- More so among the cases than the noncases
- The odds of exposure in the hospital sample won’t reflect the odds in the population
- characterized by selective factors that lead hospital cases and controls to be systematically different from one another.
- Occurs when the combination of exposure and disease increases the probability of hospitalization

Avoiding selection bias

Reducing Selection Bias:

- develop explicit (objective) case definition
- enroll all cases in a defined time and region
- strive for high participation rates

Cases:

- are all medical facilities thoroughly canvassed?
- is there an effective system for case ascertainment?
- do all cases require medical attention

Controls:

- is the prevalence of the exposure credible?
- draw controls from a variety of sources?

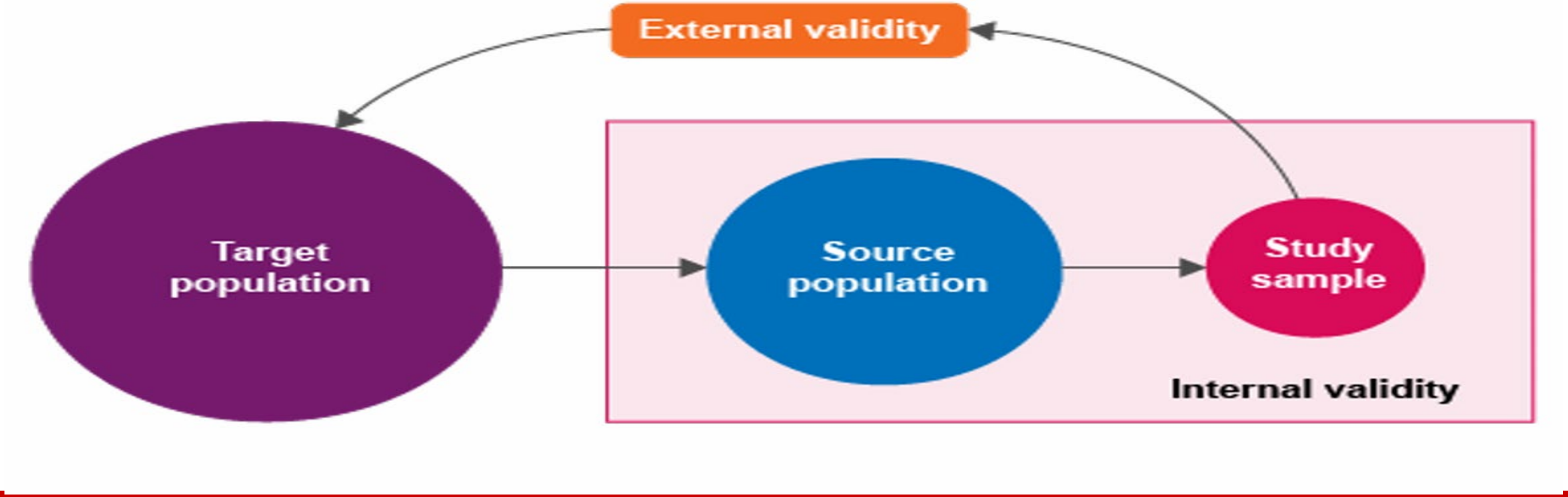
Exploring selection bias

- Analysis Phase: explore potential selection effects using simple sensitivity analyses or bias models
- "worst case scenario" - assume nonrespondent cases were unexposed and all nonrespondent controls were exposed.

Note: different rates of response between case and control does not itself introduce bias. No bias if exposure prevalence is the same for refusing and participating subjects.

External validity

- A result is externally valid if the true effect in the study sample is unbiased for the true effect in the target population
- Versus a result that is internally valid when the effect estimated in the study sample is unbiased for the true effect in that sample



External Validity

- If study sample is not the same as the target population we cannot assume that the true causal effect in a study sample will be the same as the true effect in the population
- It is near-universally overlooked that estimates of causal effects obtained from a study sample are only well-defined if they include specific reference to a target population in which they are said to apply GENERA

External Validity

- Target validity = Internal validity + external validity
- RCT conducted in a white MSM under-30 (low external validity; high internal validity)
- Observational study conducted in a sample of population where all groups are represented (high external validity; low internal validity)
- If you are producing evidence more relevant to public health policy change then need to consider internal validity AND external validity

- Daniel Westreich and others, Target Validity and the Hierarchy of Study Designs, American Journal of Epidemiology, Volume 188, Issue 2, February 2019, Pages 438–443, https://doi.org/10.1093/aje/kwy228 Why is external validity a problem?

THANK YOU

- **ACKNOWLEDMENTS:**

Dr Jabulani Ncayiyana:

UKZN Intermediate epidemiology course coordinator