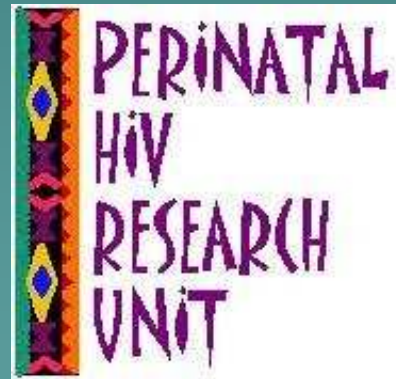


Jaco Botha

Biostatistician

Perinatal HIV Research Unit




# Overview

1. Study Designs
  - ◆ Cohort
  - ◆ Case-control
    - Nested case-control
    - Cross-sectional
2. Basic Concepts in Epidemiology
3. Randomization
4. Sample Size
5. Statistics

“Clinical research does not just happen, it has to be thought out.”  
(Cyril Maxwell)

“The importance of the planning stage cannot be over-emphasized, since no amount of clever analysis later will be able to compensate for major design flaws.”  
(Douglas G. Altman)

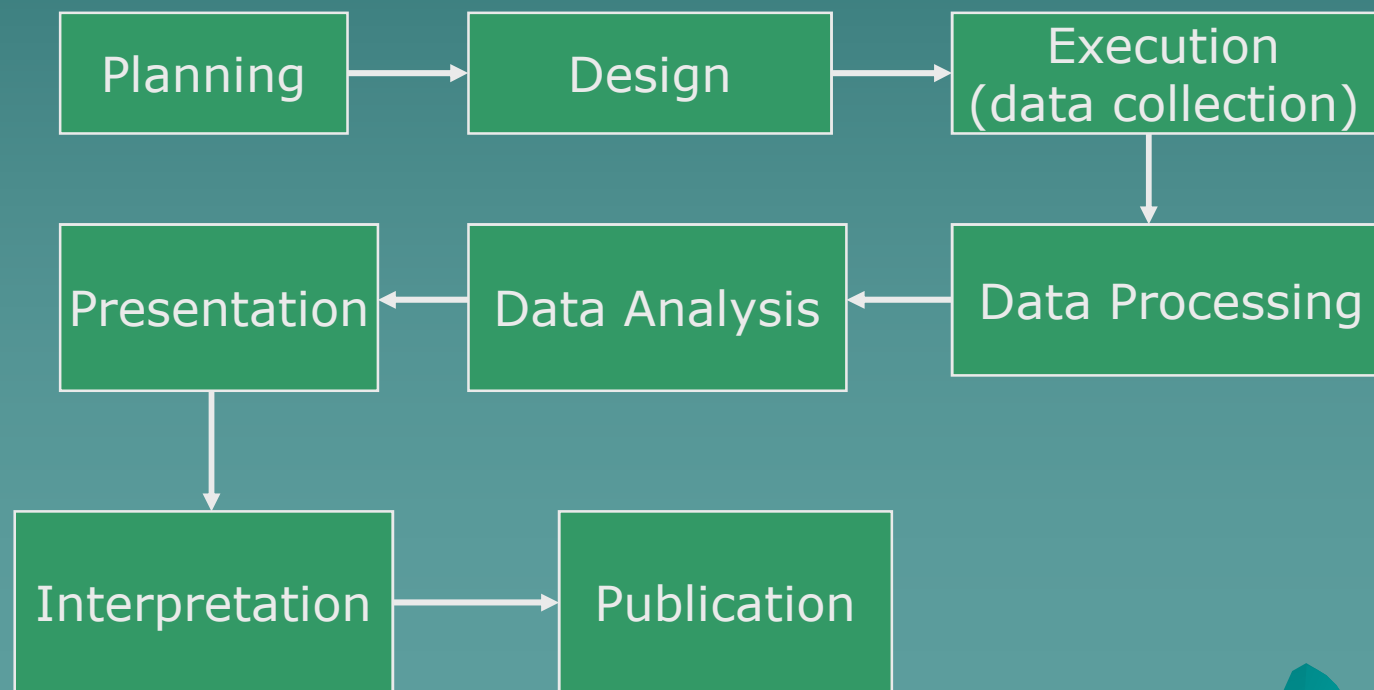


# Overview

1. **Study Designs**
  - ◆ Cohort
  - ◆ Case-control
    - Nested case-control
    - Cross-sectional
2. Basic Concepts in Epidemiology
3. Randomization
4. Sample Size
5. Statistics

# Study designs

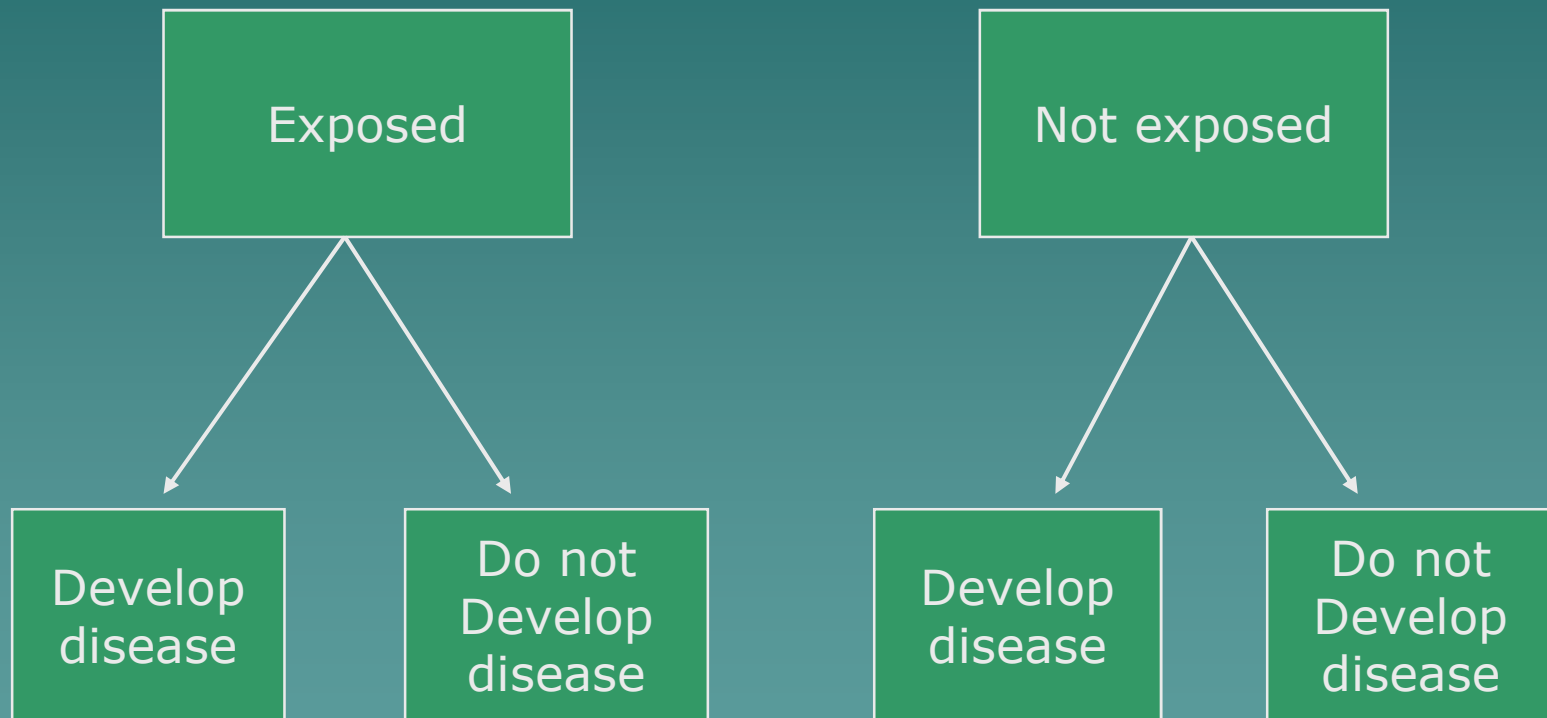
- ◆ General sequence of steps in a research project



# Study designs - Cohort

- ◆ Also called “prospective study”
- ◆ Investigator selects group of exposed individuals and group of non-exposed
- ◆ Both groups followed up – compare incidence of disease/rate of death
- ◆ Design may include more than 2 groups
- ◆ *Schematically...*

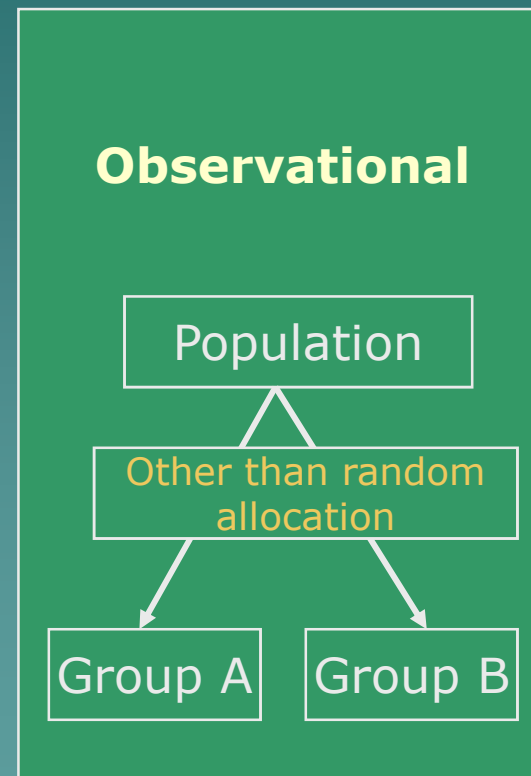
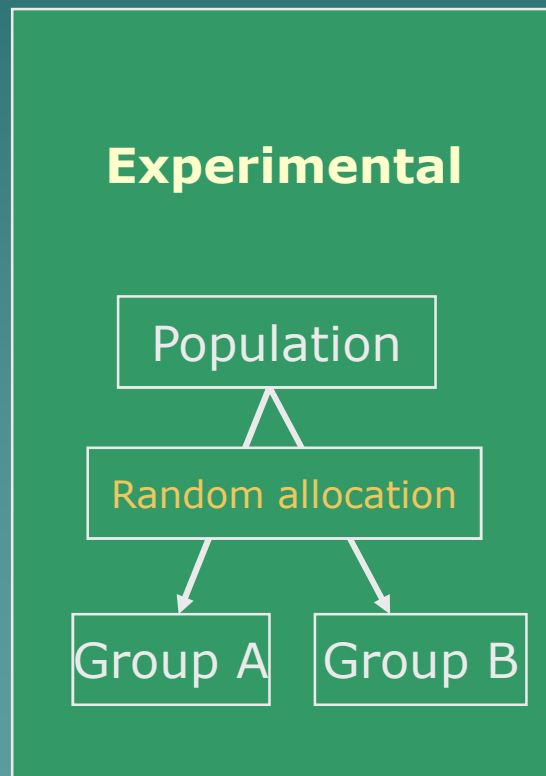
# ...Study design – Cohort



## ...Study design – Cohort

- ◆ Positive association between exposure and disease  $\Rightarrow$  proportion of exposed in whom disease develops  $>$  proportion of non-exposed in whom disease develops
- ◆ Types of cohort studies
  - Observational (described above)
  - Randomized trial (experimental cohort)
- ◆ *Schematically...*

# ...Study design – Cohort



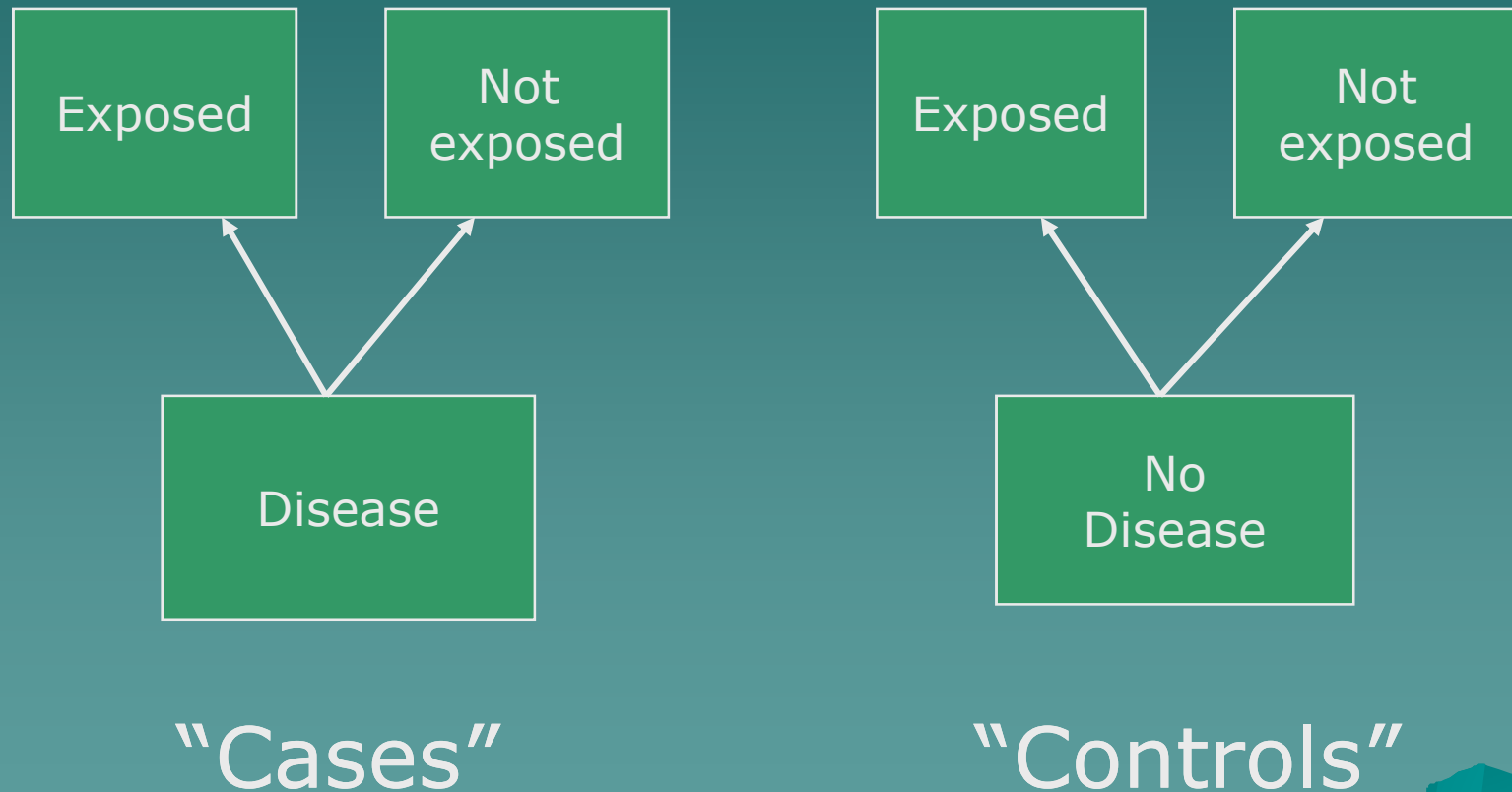
## ...Study design – Cohort

- ◆ Both types of studies compare exposed with non-exposed
- ◆ Difference = presence/absence of randomization
- ◆ Problem with cohort is that study population often must be followed up for long period – *may lead to bias*

# Study design – Case-control

- ◆ “Reverse” of cohort study
- ◆ Examine possible relation of an exposure to certain disease
- ◆ Identify individuals with disease (CASES) and group without disease (CONTROLS)
- ◆ Proportion cases exposed and proportion cases not exposed?

# ...Study design – Case-control



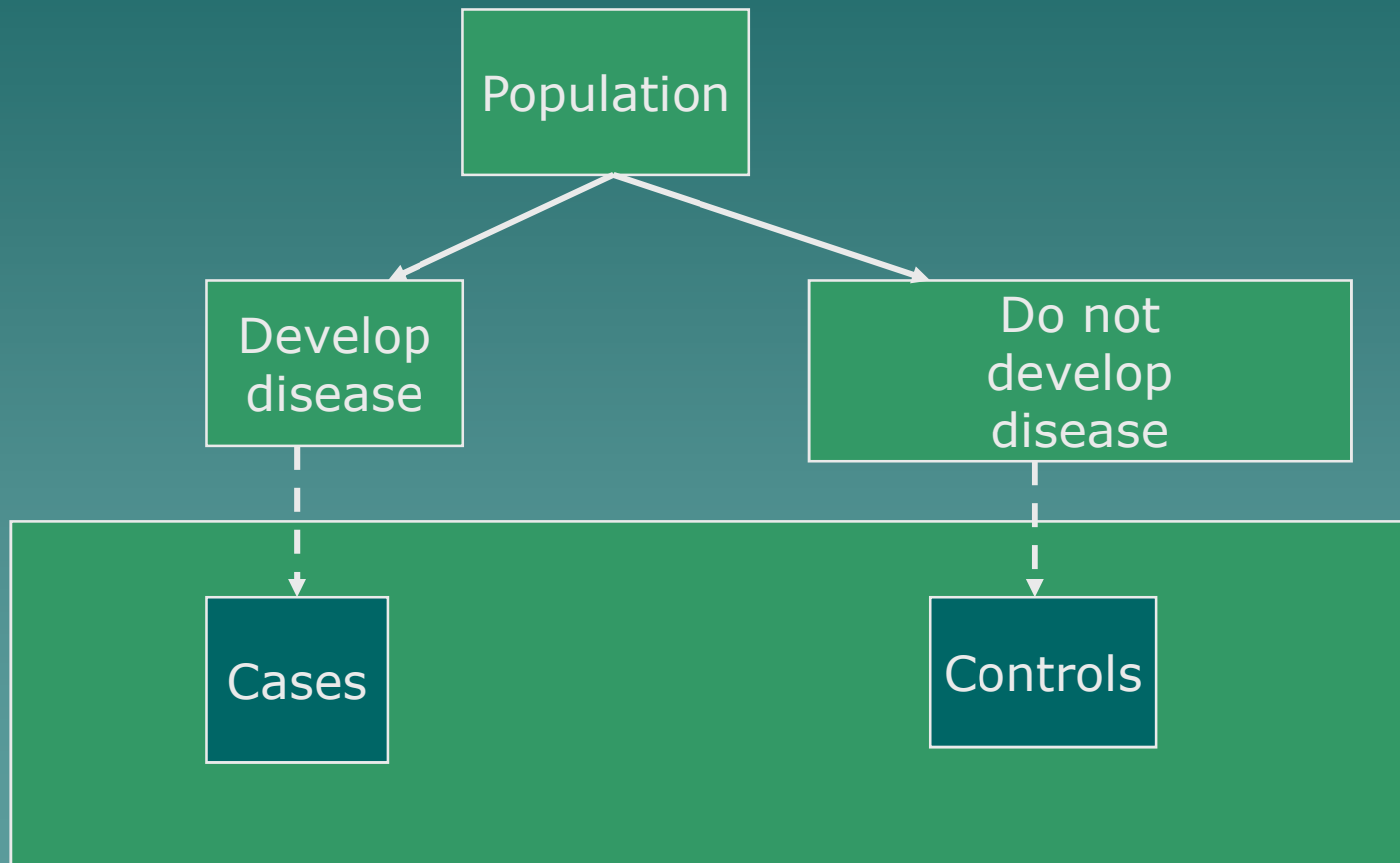
## ...Study design – Case-control

- ◆ Case-control – starts with people with disease (cases) and compares to people without disease (controls)
- ◆ Cohort – starts with group of exposed and compares to non-exposed
- ◆ Study design used increasingly = nested case-control study

## ...Study design – Case-control

- ◆ **Nested case-control** – hybrid design of case-control nested in cohort
- ◆ Population defined and followed over time
- ◆ Time when identified – baseline data
- ◆ *Schematically...*

# ...Study design – Case-control



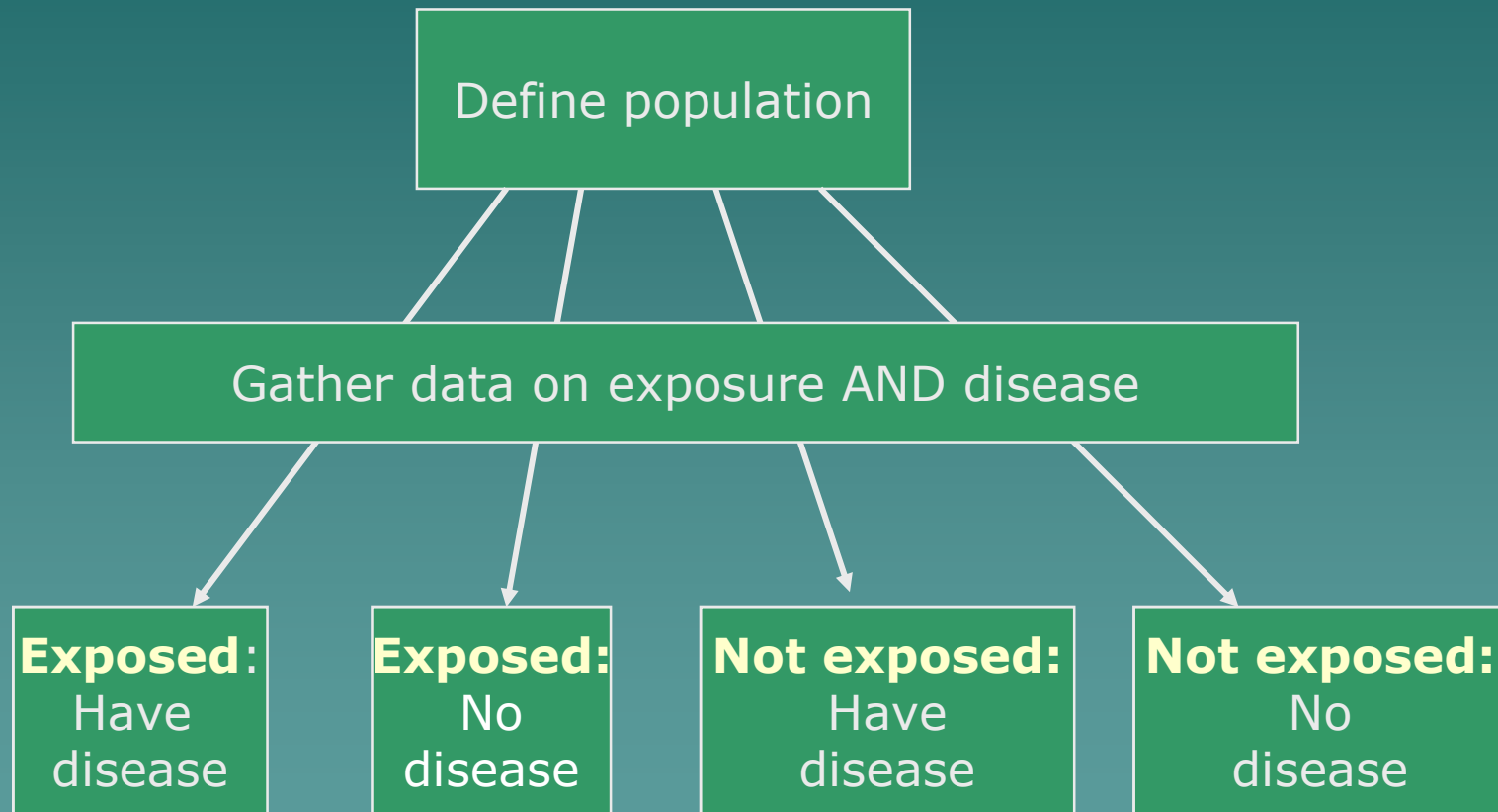
# ...Study design – Case-control

- ◆ Advantages nested case-control
  - ◆ Interviews at beginning of study (baseline) - data obtained before any disease ⇒ possibility of *recall bias eliminated*
  - ◆ Often more economical
- ◆ Another case-control = **cross-sectional**

# ...Study design – Case-control

- ◆ Both exposure and disease outcome determined simultaneously (observed only once)
  - ◆ Example: Possible relationship of ↑serum cholesterol (**exposure**) to ECG evidence of CHD (**disease**) ⇒ for each subject determine cholesterol level and perform ECG
- ◆ It is like “slicing” through population capturing cholesterol levels AND evidence CHD at same time
- ◆ *Schematically...*

# ...Study design – Case-control



# Study design – Summary

Case-control

=

Retro-spective

---

Cross-sectional

=

Prevalence study

---

Cohort

=

Longitudinal

=

Pro-spective

---

Randomized

=

Experimental study



# Overview

## 1. Study Designs

- ◆ Cohort
- ◆ Case-control
  - Nested case-control
  - Cross-sectional

## 2. **Basic Concepts in Epidemiology**

## 3. Randomization

## 4. Sample Size

## 5. Statistics

# Basic Concepts in Epidemiology

- ◆ **Incidence** = Number of NEW cases of disease that occur during specified period of time in population at risk
- ◆ Incidence is measure of events  $\Rightarrow$  measure of risk
- ◆ For incidence to be measure of risk you have to specify period of time
- ◆ *Example in cohort study...*

# ...Basic Concepts in Epi

	<b>Develop CHD</b>	<b>Do not develop CHD</b>	<b>Total</b>
<b>Smoke</b>	84	2916	<b>3000</b>
<b>No Smoke</b>	87	4913	<b>5000</b>
<b>Total</b>	<b>171</b>	<b>7829</b>	<b>8000</b>

- ◆ Incidence =  $84/3000 = 2.8\%$
- ◆ Or Incidence = 28 per 1000

## ...Basic Concepts in Epi

- ◆ **Prevalence** = Number of affected persons in population at specific time
- ◆ Difference between Incidence and Prevalence = Prevalence can be viewed as slice through population at point in time
- ◆ Cure and death ↓ prevalence
- ◆ Incidence ↑ prevalence

# ...Basic Concepts in Epi

- ◆ Case-control and Cohort designed to determine if there is an association between exposure and development disease
- ◆ In cohort studies  $\Rightarrow$  Relative risk (RR)
- ◆  $RR = \text{Risk in exp} / \text{Risk unexp}$
- ◆ *Example...*

# ...Basic Concepts in Epi

	<b>Develop CHD</b>	<b>Do not develop CHD</b>	<b>Total</b>
<b>Smoke</b>	84	2916	<b>3000</b>
<b>No Smoke</b>	87	4913	<b>5000</b>
<b>Total</b>	<b>171</b>	<b>7829</b>	<b>8000</b>

- ◆ Incidence Exp =  $84/3000 = 2.8\%$
- ◆ Incidence Unexp =  $87/5000 = 1.74\%$
- ◆ RR =  $2.8/1.74 = 1.61$

# ...Basic Concepts Epi

- ◆ In case-control – do not know incidence  $\Rightarrow$  no RR
- ◆ Thus, in case-control another measure of association = odds ratio (OR)
- ◆ *Example...*

# ...Basic Concepts in Epi

	<b>CHD (Cases)</b>	<b>No CHD (Controls)</b>	<b>Total</b>
<b>Smoke</b>	200	9800	<b>10000</b>
<b>No Smoke</b>	100	9900	<b>10000</b>
<b>Total</b>	<b>300</b>	<b>19700</b>	<b>20000</b>

◆  $OR = (200/9800) / (100/9900)$   
 $= (200 \times 9900) / (100 \times 9800) = 2.02$

## ...Basic Concepts Epi

- ◆  $OR = 1 \Rightarrow$  Exposure not related to disease
- ◆  $OR > 1 \Rightarrow$  Exposure + related to disease
- ◆  $OR < 1 \Rightarrow$  Exposure – related to disease or exposure protective

# Overview

1. Study designs
  - ◆ Cohort
  - ◆ Case-control
    - Nested case-control
    - Cross-sectional
2. Basic Concepts in Epidemiology
3. **Randomization**
4. Sample Size
5. Statistics

# Randomization

- ◆ Randomization one of fundamental principles of experimental design
- ◆ Two main reasons randomization
  - ◆ Prevent bias
  - ◆ Statistical theory based on random sampling
- ◆ Random  $\neq$  haphazard
- ◆ Random = each patient has known chance (usually equal) being given each treatment

# ...Randomization

- ◆ BUT treatment to be given cannot be predicted
- ◆ Methods
  - ◆ Tossing a coin
  - ◆ Table of random numbers
  - ◆ Random number generator
- ◆ Block randomization to keep numbers of subjects in different groups closely balanced

# ...Randomization

- ◆ E.g., subjects in blocks of four, with two treatments  $\Rightarrow$  six ways
  - ◆ AABB
  - ◆ ABAB
  - ◆ ABBA
  - ◆ BBAA
  - ◆ BABA
  - ◆ BAAB
- ◆ Also stratified randomization

## ...Randomization

- ◆ Approximate balance of NB characteristics without sacrificing advantages of randomization
- ◆ Produce separate block randomization for each stratum

# Overview

1. Study designs
  - ◆ Cohort
  - ◆ Case-control
    - Nested case-control
    - Cross-sectional
2. Basic Concepts in Epidemiology
3. Randomization
4. **Sample size**
5. Statistics

# Sample Size

“Of all the errors that occur in medical research, the vast majority are related to the sample on which the work was done. Never underestimate the importance of your sample and never hesitate to invite criticism of your intended method of drawing it; but do this always before you start; afterwards is too late.”  
(Cyril Maxwell)

## ...Sample size

- ◆ Why is estimation of sample size NB?
- ◆ Number of subjects in trial always be large enough to provide reliable answer to questions addressed
- ◆ # of subjects usually determined by primary objective of trial
- ◆ In order to justify generalization of results – sample representative of population

## ...Sample size

- ◆ Main idea behind sample size calculation – high chance of detecting a worthwhile effect (if exists) as statistically significant
- ◆ The more sure we can be the larger the sample size becomes = power (usually 80% =  $1 - \beta$ )
- ◆ Sampling small proportion of large population – errors occur - minimize

# ...Sample size

	Treatments are not different	Treatments are different
Conclude treats are not different	Correct	Type II (Prob = $\beta$ )
Conclude treatments differ	Type I (Prob = $\alpha$ )	Power (Prob = $1 - \beta$ )

- ◆ Power = Probability decide basis of results there is a difference, if in reality there is difference – tells how good study is

## ...Sample size

- ◆ Type I = significant result when it does not exist – probability  $\alpha$
- ◆ Type II = no significant result when it exists – probability  $\beta$
- ◆ Make provision drop-outs – usually taken as 20%, if not specified
- ◆ Many formulas and software packages available
- ◆ NB to quote reference

# ...Sample size

- ◆ Following should be specified:
  - Primary objective of trial
  - Primary variable
  - Estimates of the quantities used in calculations – e.g.,  $p_1$ ,  $p_2$
  - Errors allowed
    - ◆  $\alpha$  - usually 5%
    - ◆ Power - usually 80% =  $1 - \beta$
  - Drop-out rate

# Overview

1. Study designs
  - ◆ Cohort
  - ◆ Case-control
    - Nested case-control
    - Cross-sectional
2. Basic Concepts in Epidemiology
3. Randomization
4. Sample size
5. **Statistics**

# Statistics

- ◆ If there are statistical errors – conclusions may be incorrect
- ◆ If clinical trial not designed to show specific finding (X) – cannot make conclusions regarding X
- ◆ Use **INFO** from individuals to make **INFERENCES** about **POPULATION**

# ... Statistics

- ◆ 2 types of data
- ◆ Categorical data (also binary)
  - ◆ Smoker, non-smoker
- ◆ Continuous data (some form of measurement) – calculate descriptive statistics
  - ◆ Height, weight, age
  - ◆ mean, median, mode, percentiles/quartiles, range

## ...Statistics

- ◆ **Median** = value halfway when data are ranked
- ◆ Useful when extreme data values
- ◆ **Mode** = most common value
- ◆ **Percentiles** = divide data into 100 equal parts
- ◆ E.g., 1% of data fall below  $p_1$

## ...Statistics

- ◆ **Quartiles** = divide data into 4 equal parts
- ◆ 25<sup>th</sup> (Q1), 50<sup>th</sup> (median = Q2) and 75<sup>th</sup> (Q3) quartile
- ◆ **IQR** = numerical difference between Q3 and Q1
- ◆ **SD** = measure of the spread of data
- ◆ Way of quantifying variability

## ... Statistics

- ◆ Variance =  $(SD)^2$
- ◆ Parametric data = assume the data follow a normal distribution  $\Rightarrow$  parametric tests
- ◆ Non-parametric data = no assumption regarding distribution of data  $\Rightarrow$  non-parametric tests
- ◆ Can make transformation to data to follow normal distribution

## ...Statistics

- ◆ E.g., log-transformation
- ◆ Geometric mean = log-transform data, calculate mean, anti-log
- ◆ Geom Mean very close to median

## ...Statistics

- ◆ **Sensitivity** = ability of test to identify correctly those who have disease
- ◆ **Specificity** = ability of test to identify correctly those who do not have disease
- ◆ *Example...*

# ... Statistics

	Disease	No disease	Total
Positive	80	100	180
Negative	20	800	820
Total	100	900	1000

- ◆ Sensitivity =  $80/100 = 80\%$
- ◆ Specificity =  $800/900 = 89\%$

# ...Statistics

	Disease	No disease
Positive	True Pos. ( <b>TP</b> ) = disease and + test	False Pos. ( <b>FP</b> ) = no disease but have + test
Negative	False Neg. ( <b>FN</b> ) = disease and - test	True Neg. ( <b>TN</b> ) = No disease and - test

- ◆ Sensitivity =  $TP / (TP + FN)$
- ◆ Specificity =  $TN / (TN + FP)$

## ...Statistics

- ◆ Another important question for physician = if test results are + in patient, what is probability that the patient has disease
- ◆ What proportion of patients testing + actually have disease
- ◆ Positive predictive value (PPV)

# ...Statistics

- ◆ If test results are -, what is probability that patient does not have disease
- ◆ Negative predictive value (NPV)
- ◆ *Example...*

# ... Statistics

	Disease	No disease	Total
Positive	80	100	<b>180</b>
Negative	20	800	<b>820</b>
Total	<b>100</b>	<b>900</b>	<b>1000</b>

- ◆  $PPV = 80/180 = 44\%$
- ◆  $NPV = 800/820 = 98\%$

*The End*

