Epidemiologic Study Designs

Jacky M Jennings, PhD, MPH

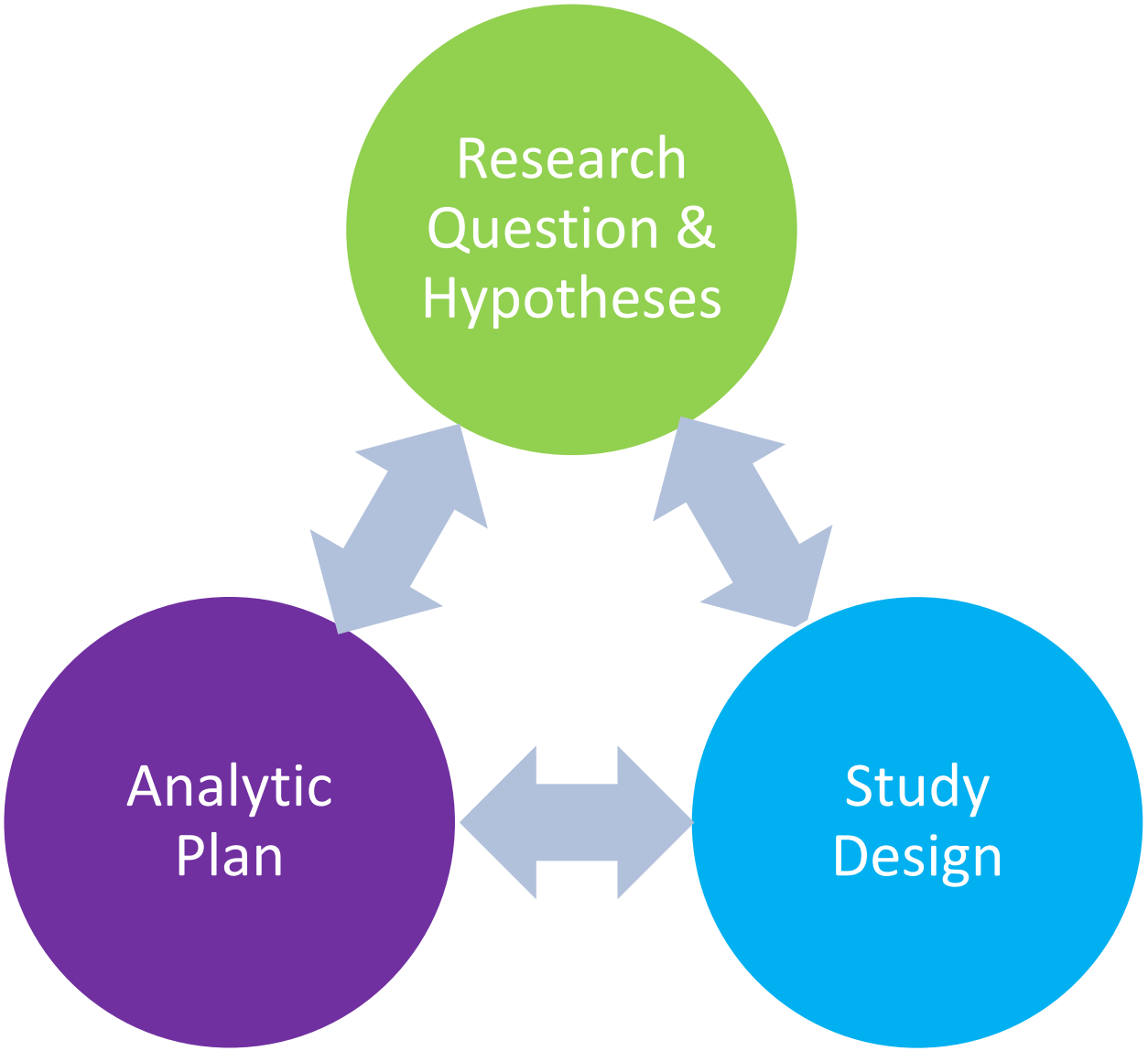
Associate Professor
Associate Director, General Pediatrics and Adolescent Medicine
Director, Center for Child & Community Health Research (CCHR)
Departments of Pediatrics & Epidemiology
Johns Hopkins University

The Center for Child and Community Health Research

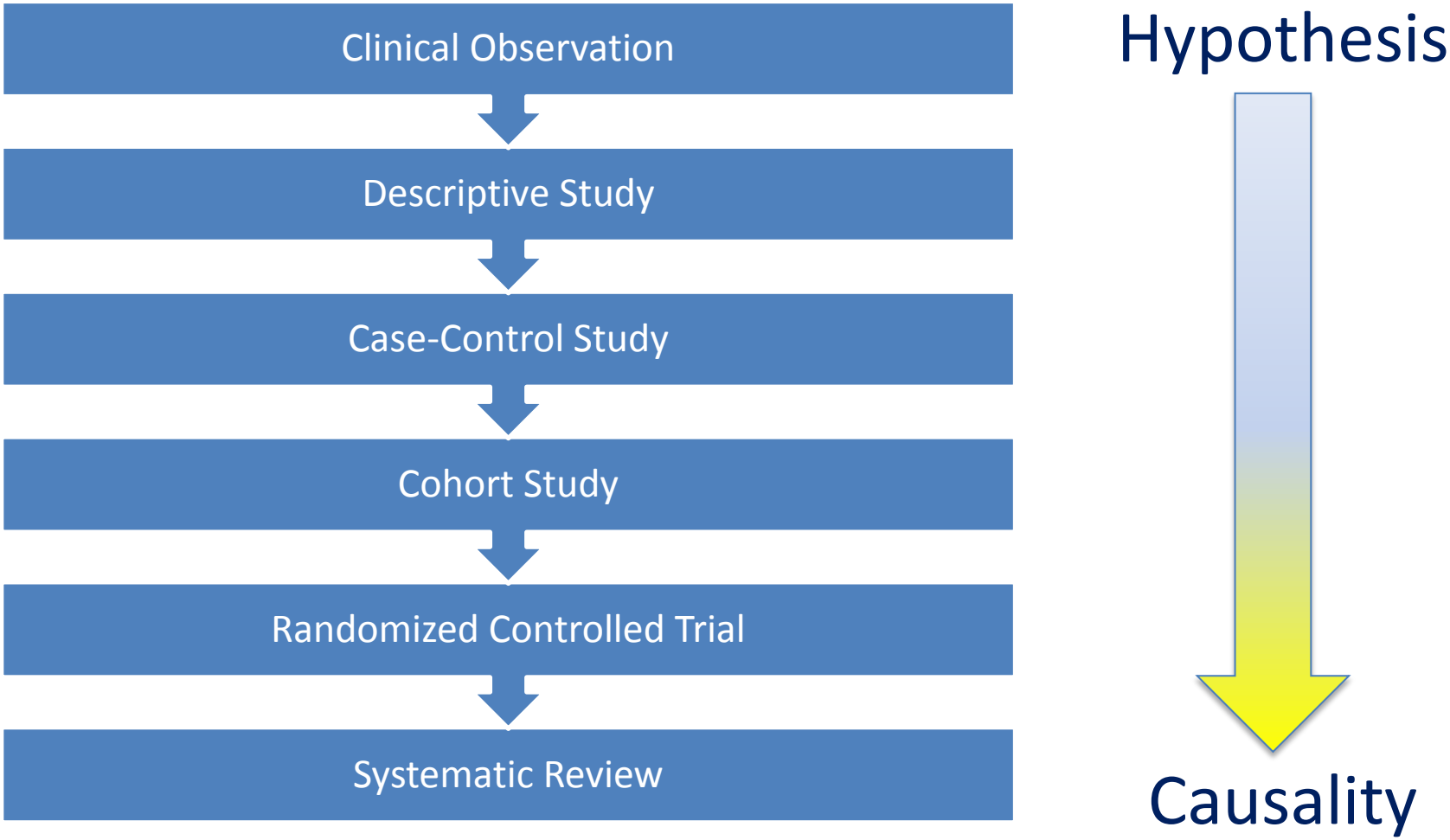
JOHNS HOPKINS CHILDREN'S CENTER

Learning Objectives

- Identify basic epidemiologic study designs and their frequent sequence of study
- Recognize the basic components
- Understand the advantages and disadvantages
- Appropriately select a study design


Research Question & Hypotheses
Analytic Plan
Study Design

Basic Study Designs and their Hierarchy



Adapted from Gordis, 1996

MMWR

1981 June 5;30:250–2

Pneumocystis Pneumonia — Los Angeles

In the period October 1980-May 1981, 5 young men, all active homosexuals, were treated for biopsy-confirmed *Pneumocystis carinii* pneumonia at 3 different hospitals in Los Angeles, California. Two of the patients died. All 5 patients had laboratory-confirmed previous or current cytomegalovirus (CMV) infection and candidal mucosal infection. Case reports of these patients follow.

Study Design in Epidemiology

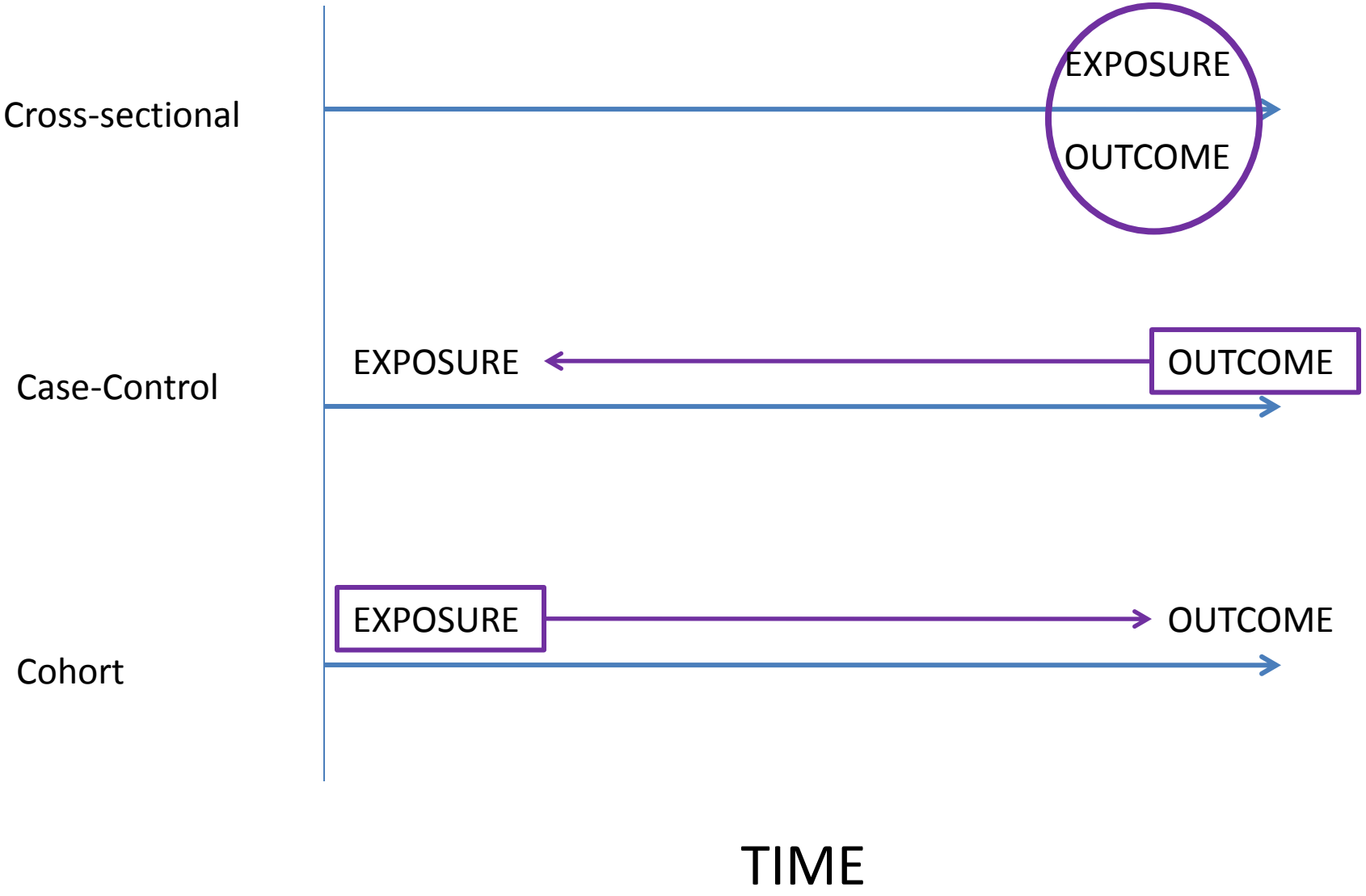
- Depends on:
  - **The research question and hypotheses**
  - Resources and time available for the study
  - Type of outcome of interest
  - Type of exposure of interest
  - Ethics

Study Design in Epidemiology

- Includes:
  - The research question and hypotheses
  - Measures and data quality
  - Time
  - Study population
    - Inclusion/exclusion criteria
    - Internal/external validity

Epidemiologic Study Designs

- Descriptive studies
  - Seeks to measure the frequency of disease and/or collect descriptive data on risk factors
- Analytic studies
  - Tests a causal hypothesis about the etiology of disease
- Experimental studies
  - Compares, for example, treatments


Cross-sectional
EXPOSURE
OUTCOME
Case-Control
EXPOSURE
OUTCOME
Cohort
EXPOSURE
OUTCOME
TIME

Cross-sectional studies

- Measure existing disease and current exposure levels at one point in time
- Sample without knowledge of exposure or disease
- Ex. Prevalence studies

Cross-sectional studies

- Advantages
  - Often early study design in a line of investigation
  - Good for hypothesis generation
  - Relatively easy, quick and inexpensive…*depends on question*
  - Examine multiple exposures or outcomes
  - Estimate prevalence of disease and exposures

Cross-sectional studies

- Disadvantages
  - Cannot infer causality
  - Prevalent vs. incident disease
  - May miss latent disease
  - May be subject to recall bias

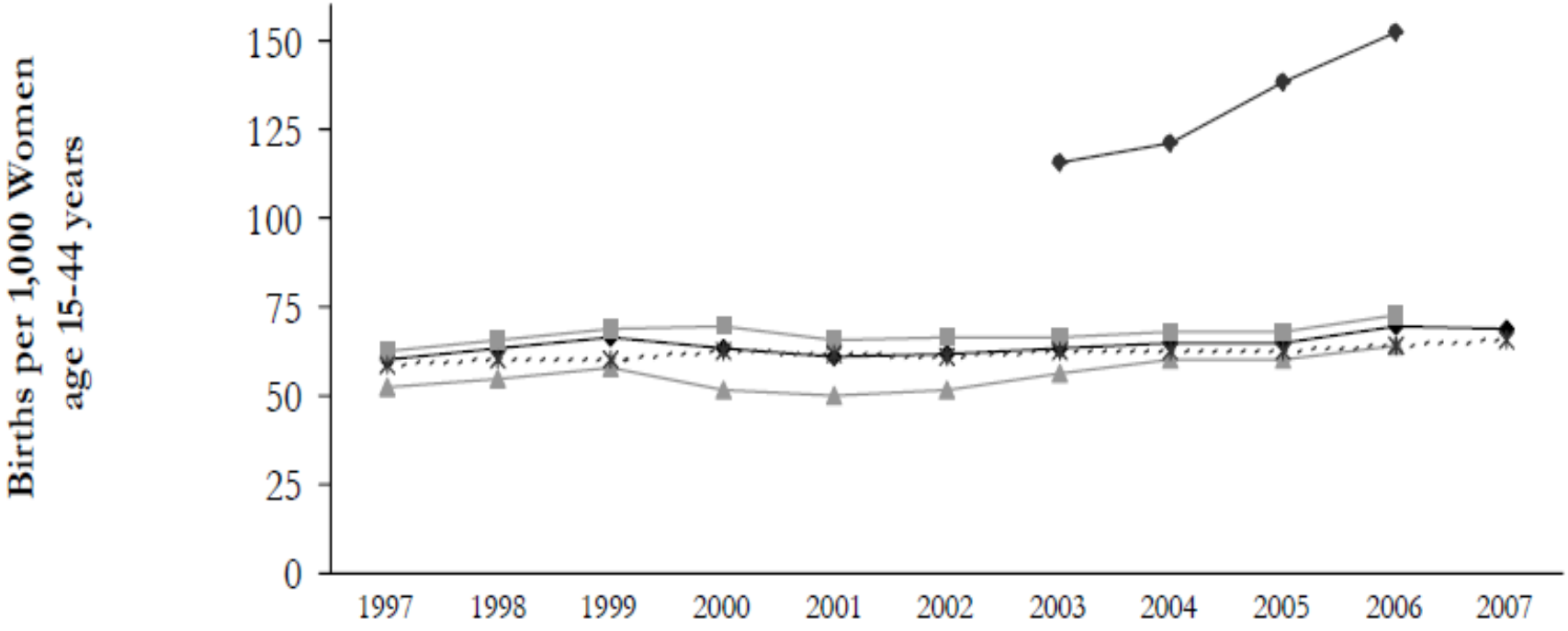
Research Question

- Determine whether there are differences in rates of stroke and myocardial infarction by gender and race among patients.

Hypothesis

- *There will be differences in rates of stroke by gender and race.*
- *There will be differences in rates of myocardial infarction by gender and race.*

General Fertility Rate, Baltimore City by Race and Maryland 1997-2007



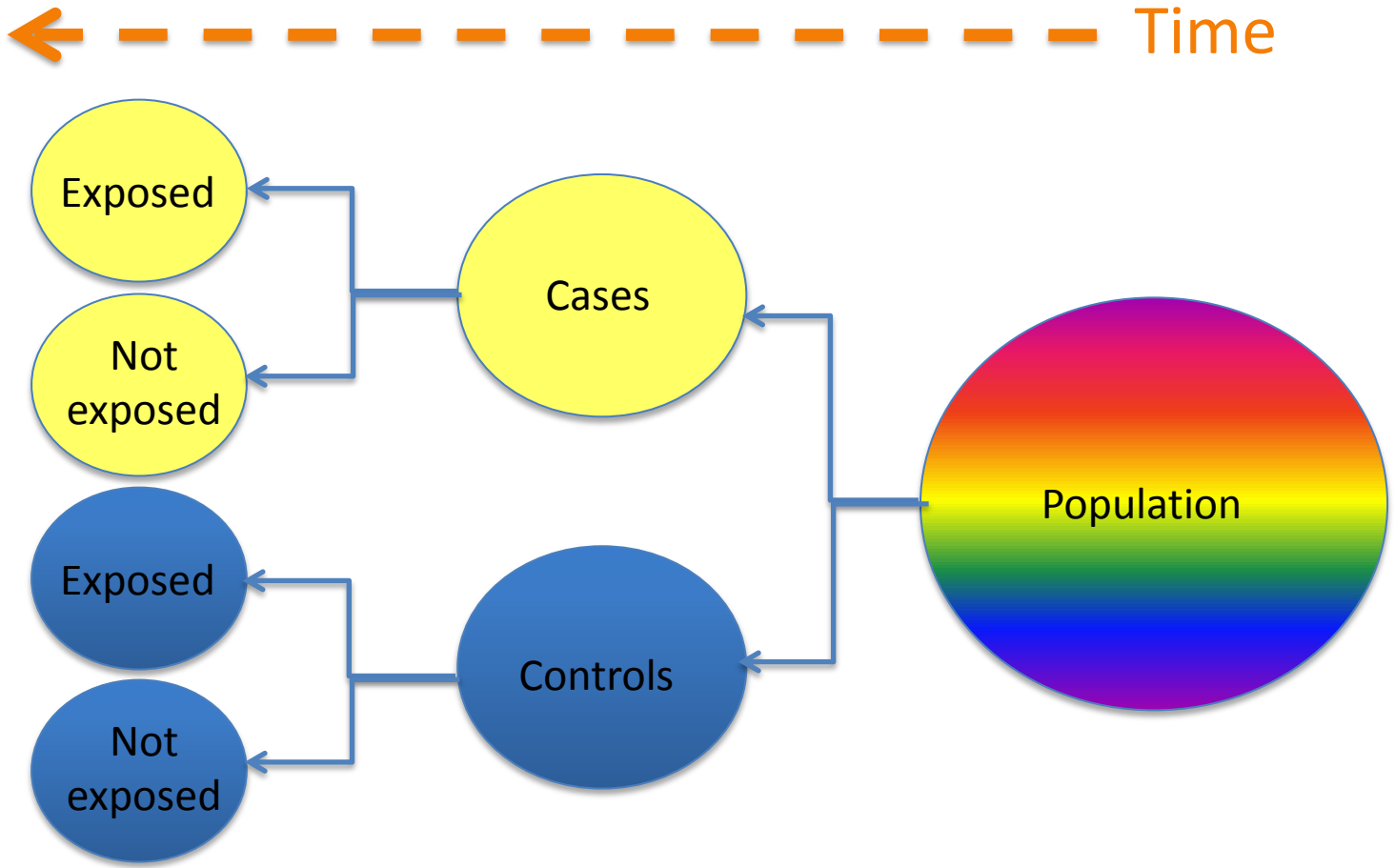
	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Baltimore City All Races	60.0	63.0	66.0	63.1	60.7	61.5	63.0	64.6	65.1	69.3	68.5
African American	62.1	65.7	68.4	69.6	65.6	66.3	66.0	67.7	68.2	72.7	
White	52.6	54.9	57.7	51.2	50.3	51.3	55.9	59.9	60.1	64.0	
Hispanic*							115.9	121.0	138.5	152.4	
Maryland All Races	58.8	60.0	60.3	62.4	61.5	60.6	62.3	62.2	62.7	64.1	65.5

Source: Maryland Department of Health and Mental Hygiene, Vital Statistics Annual Report (2007 data are preliminary and not yet available by race/ethnicity)

**Includes all births to mothers of Hispanic origin of any race, data not available prior to 2003*

Case-Control studies

- Identify individuals with existing disease/s and retrospectively measure exposure



Case-Control studies

- Advantages
  - Good design for rare, chronic and long latency diseases
  - Relatively inexpensive (population size and time)
  - Allows for the examination of multiple exposures
  - Estimate odds ratios
  - Hospital-based studies and outbreaks

Case-Control studies

- Disadvantages
  - Multiple outcomes cannot be studied
  - Recall bias
  - Sampling bias
  - Cannot calculate prevalence, incidence, population relative risk or attributable risk
  - Beware of reverse causation

Neonatal Abstinence Syndrome (NAS) and Drug Exposure

Research question

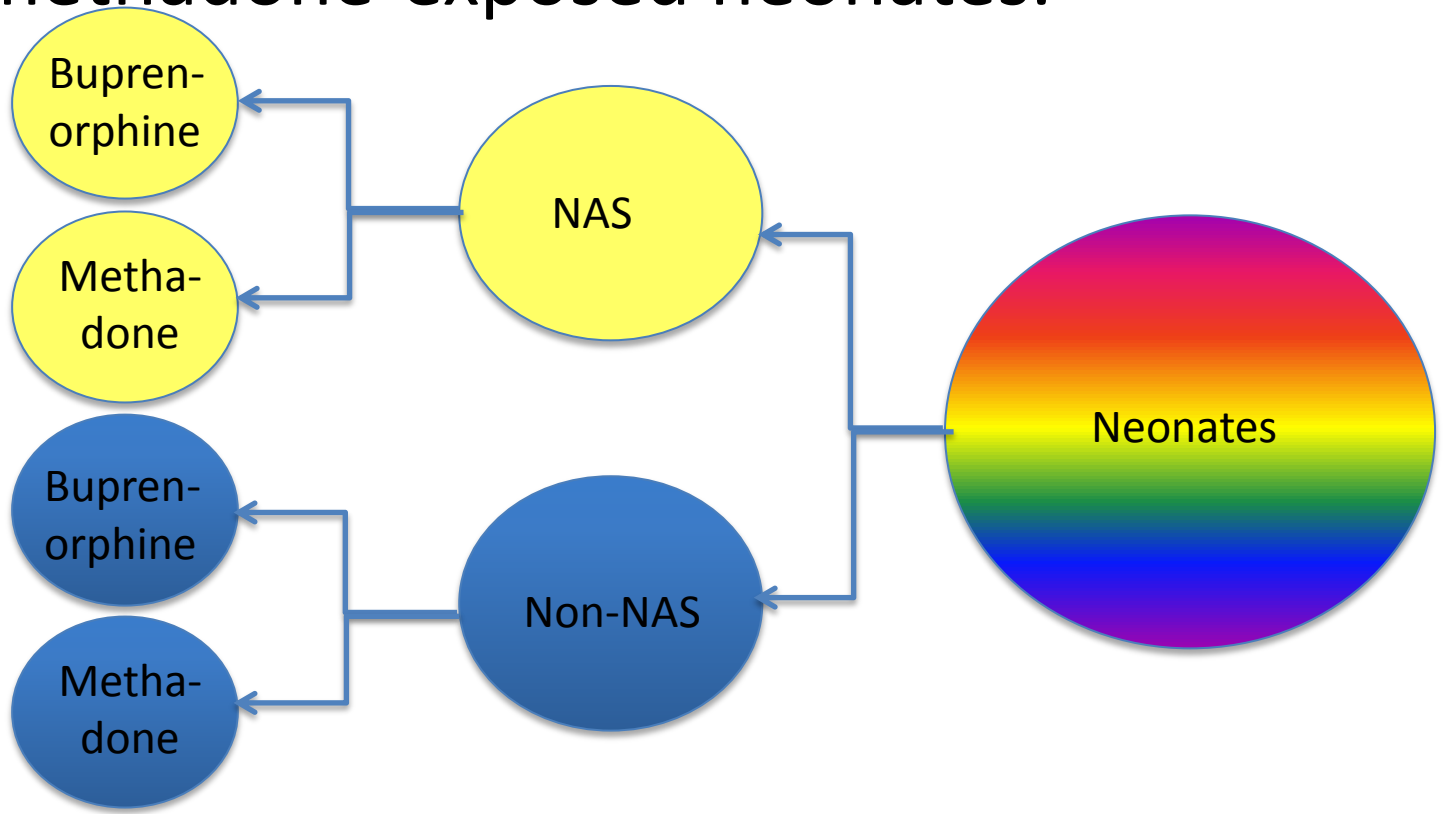
?

Hypothesis 1

Buprenorphine-exposed neonates will exhibit less NAS than methadone-exposed neonates.

Case-Control Study Example

- Hypothesis 1: Buprenorphine-exposed neonates will exhibit less NAS than methadone-exposed neonates.

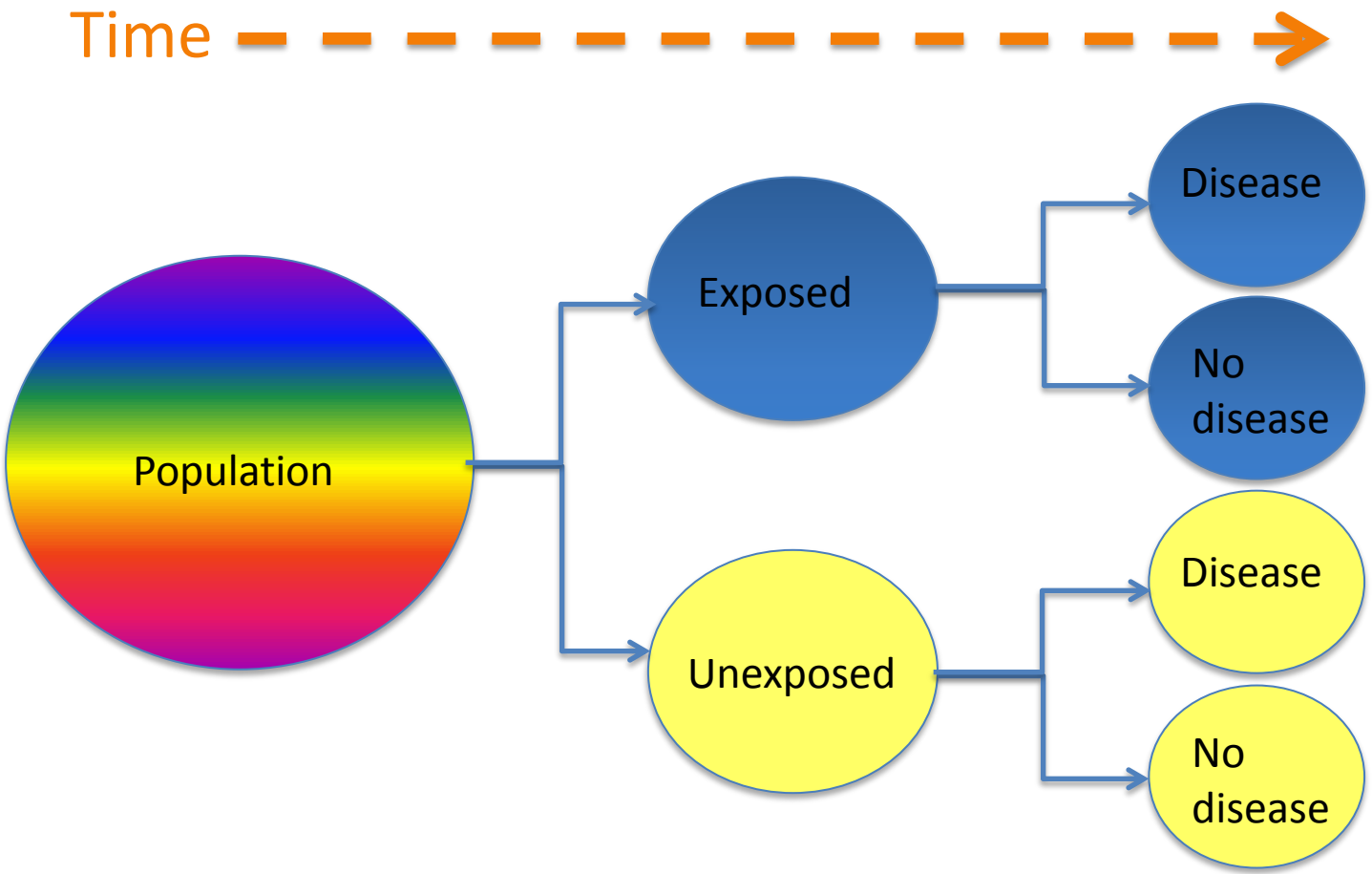


Challenges in Case-Control Studies

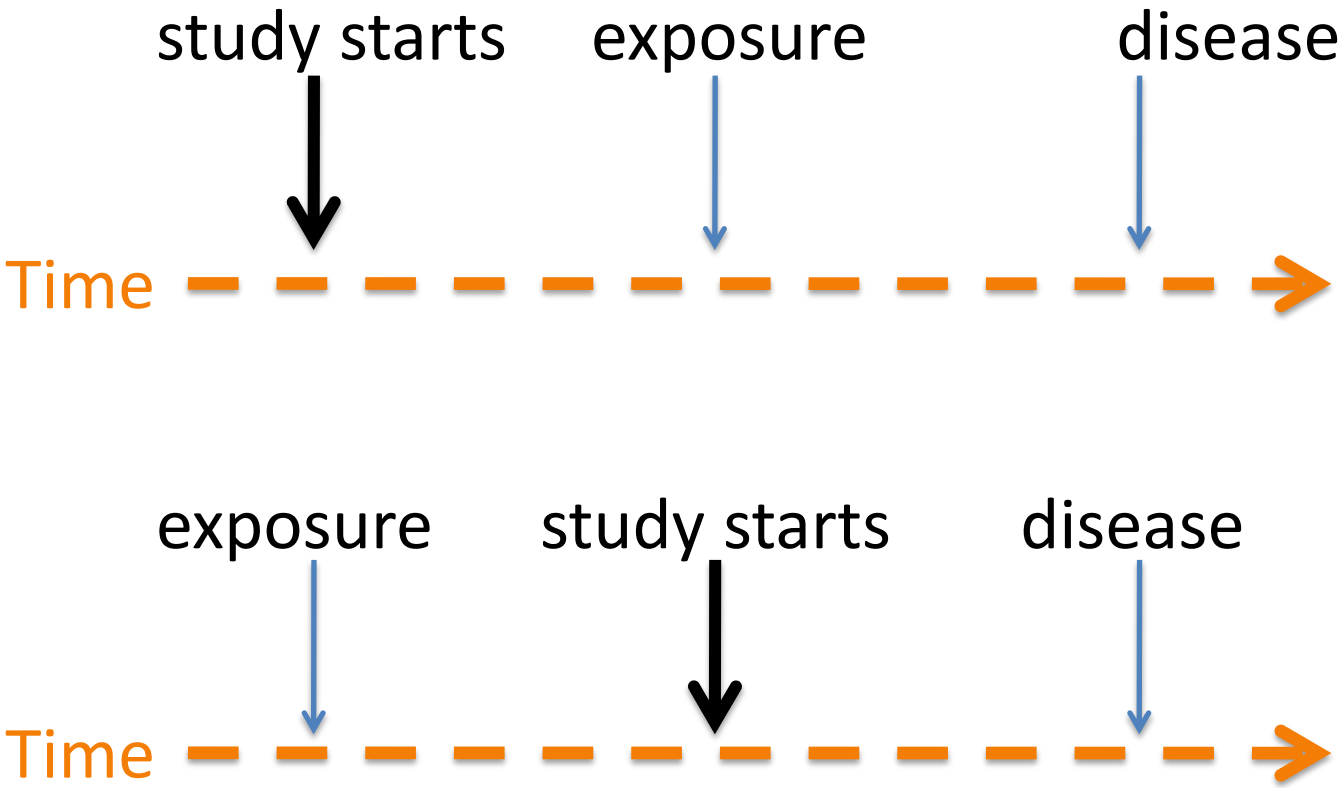
- Selection of Controls
  - Sample size
  - Matching (group or individual)
- Selection of Cases
  - Incident or prevalent disease
- Nested case-control study

Cohort Studies

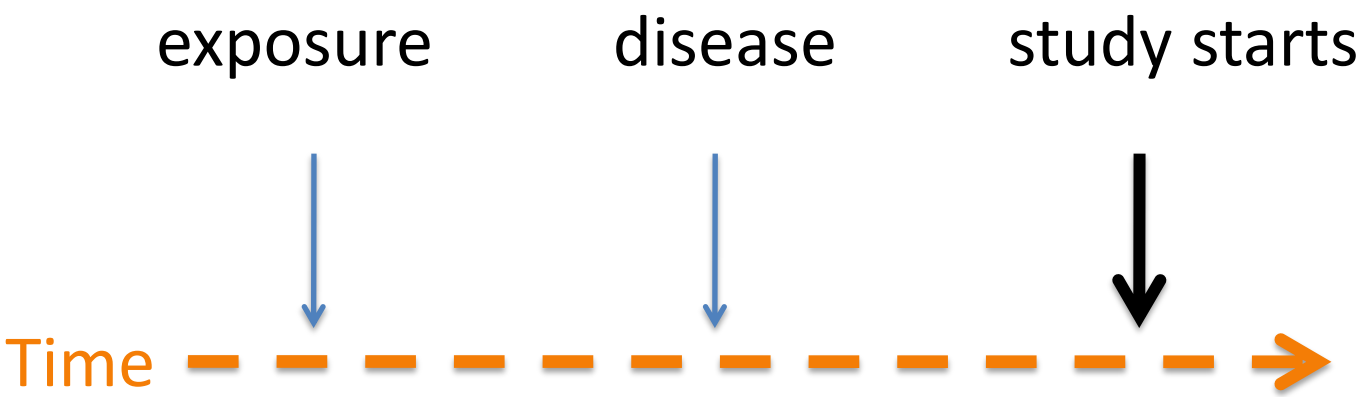
- Identify exposed and unexposed individuals and follow them over time measuring outcome/s (Prospective)



Prospective Cohort Study



Retrospective Cohort Study



Cohort Studies

- Advantages
  - Measure population-based incidence
  - Relative risk and risk ratio estimations
  - Rare exposures
  - Temporality
  - Less likely to be subject to biases (recall and selection as compared to Case-control)
  - Possible to assess multiple exposures and/or outcomes

Cohort Studies

- Disadvantages
  - Impractical for rare diseases and diseases with a long latency
  - Expensive
    - Often large study populations
    - Time of follow-up
  - Biases
    - Design - sampling, ascertainment and observer
    - Study population – non-response, migration and loss-to-follow-up

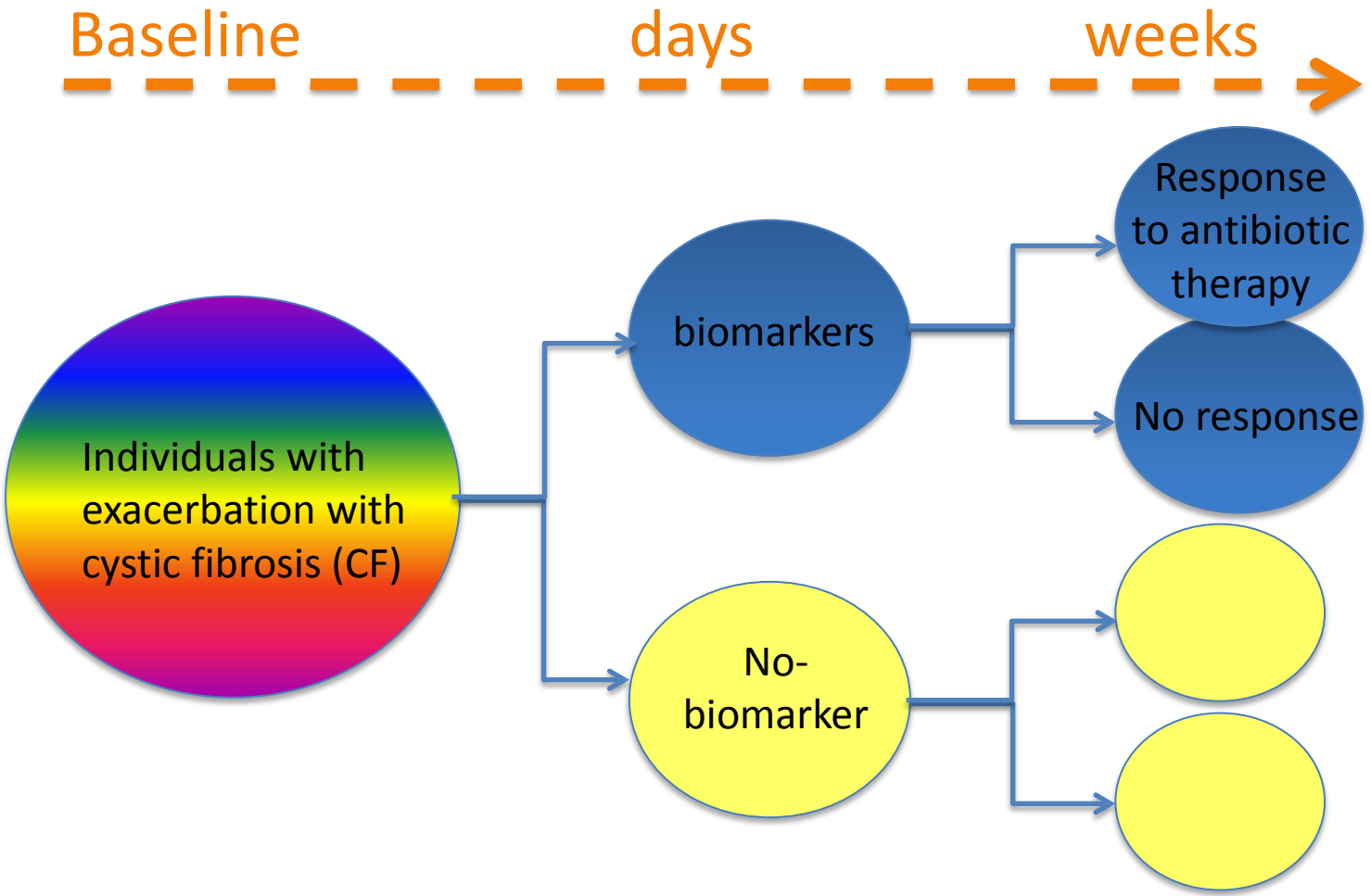
Research Question

Determine whether circulating biomarkers (i.e. C-reactive protein; exhaled breath condensate - pH, hydrogen peroxide, 8-isoprostene, nitrite, nitrate levels; sputum - TNF-α, IL-6, IL-8, IL-1β, neutrophil elastase; and fractional exhaled nitric oxide) predict individuals who will benefit from initiation of antibiotic therapy for the treatment of a mild decrease in FEV_1.

Hypothesis

Biomarkers at the time of presentation with a mild increase in pulmonary symptoms or small decline in FEV_1 can be used to identify which patients require antibiotics to recover.

Cohort Study



Important features

- How much selection bias was present?
  - Were only people at risk of the outcome included?
  - Was the exposure clear, specific and measureable?
  - Were the exposed and unexposed similar in all important respects except for the exposure?
- Were steps taken to minimize information bias?
  - Was the outcome clear, specific and measureable?
  - Was the outcome identified in the same way for both groups?
  - Was the determination of the outcome made by an observer blinded to treatment?

Important features

- How complete were the follow-up of both groups?
  - What efforts were made to limit loss to follow-up?
  - Was loss to follow-up similar in both groups?
- Were potential confounding factors sought and controlled for in the study design or analysis?
  - Did the investigators anticipate and gather information on potential confounding factors?
  - What methods were used to assess and control for confounding?

Randomized Controlled Trials (RCTs)

- Experimental: exposure is assigned
- Randomization assignment
  - Random allocation of exposure or treatment
  - Results (or should result!) in two equivalent groups on all measured and unmeasured confounders
- Gold Standard for causal inference

Randomized Controlled Trials

- Advantages
  - Least subject to biases of all study designs (**IF** designed and implemented well…!)

Randomized Controlled Trials

- Disadvantages
  - Intent-to-treat
  - Loss-to-follow-up
  - Randomization issues
  - Not all exposures can be “treatments”, i.e. are assignable
  - Note: for reporting of RCTs see
    Altman DG, et al. CONSORT GROUP (Consolidated Standards of Reporting Trials). Ann Intern Med. 2001 Apr 17;134(8):663-94.

Research Question

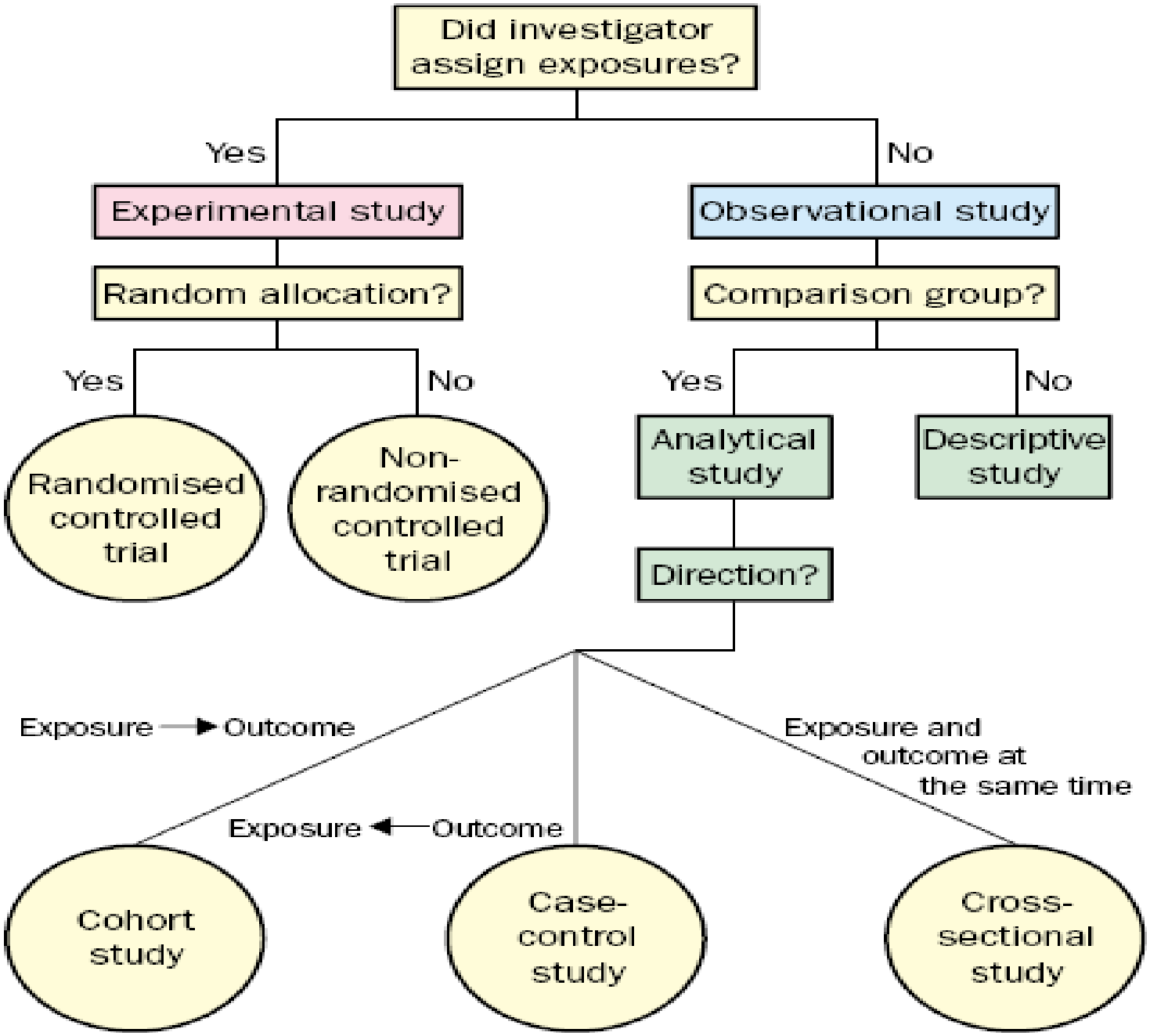
- *To determine whether resident's attitudes and skills in diabetes management and counseling change after a curricular intervention.*
- *To determine whether patient outcomes related to diabetes (i.e. weight, smoking status) change after a curricular intervention among residents.*

Hypothesis

- *Attitudes and skills related to diabetes management and counseling will improve among residents after a curricular intervention.*
- *Fewer patients with diabetes will smoke over time after a curricular intervention among residents.*

Randomization Strategies

- Randomly assigned
- Quasi-randomization
- Block randomization – method of randomization that ensures that at any point in the trial, roughly equal numbers of participants have been allocated to the comparison groups



Grimes & Schulz, 2002

Study Design

- **Must be defensible**
- **Drives conclusions:**
  What do you want to be able to say at the end of the study?

Exploratory Data Analyses

Jacky M Jennings, PhD, MPH

Objectives

- To identify some basic steps in data analyses
- To understand the reason for and methods of exploratory quantitative data analysis
- To learn some statistical tools for inferential statistics

Research Questions

- Testable hypotheses
- Measureable – exposure and outcome
- Time - how is time incorporated
- Study population

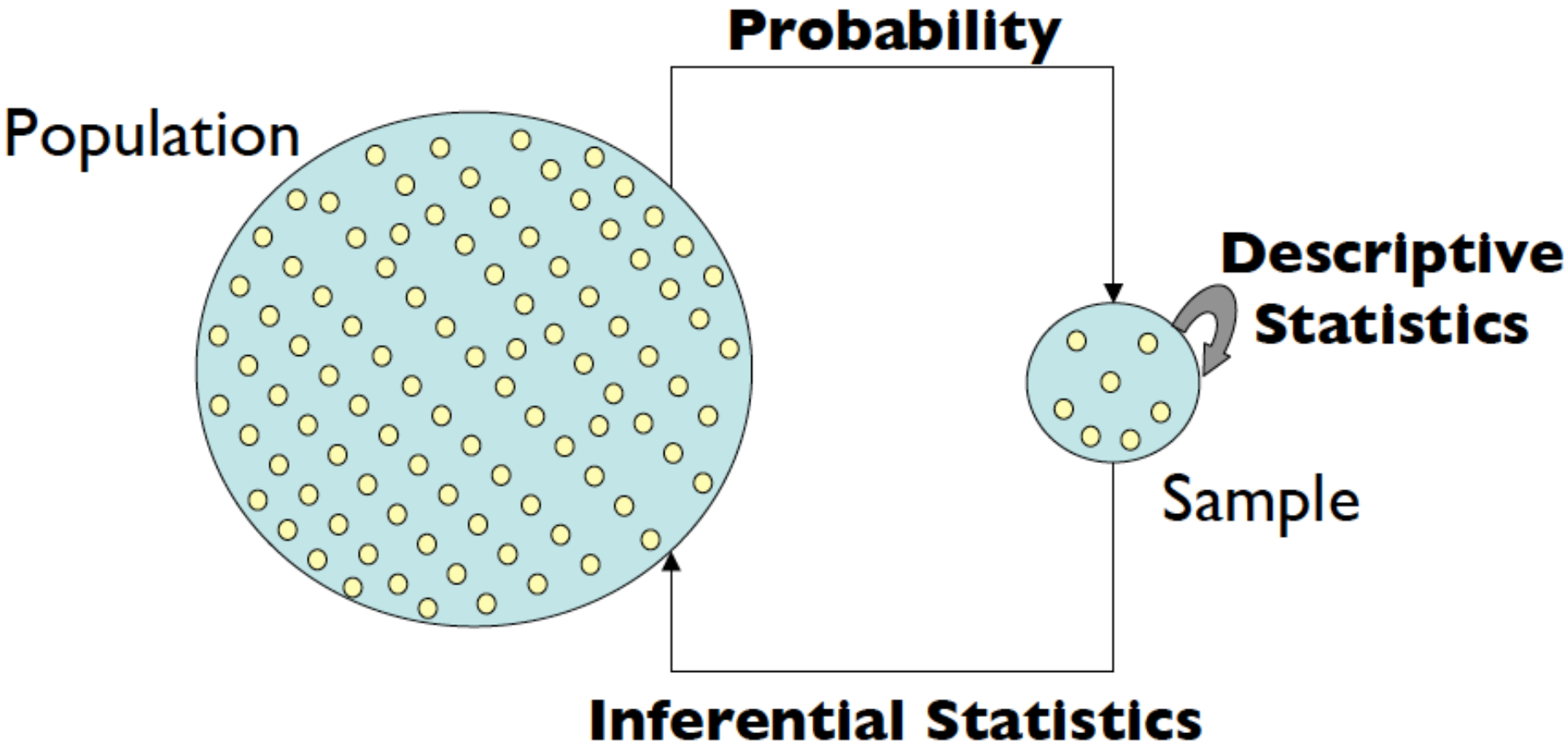
Taking Stock of your Data

- How was the data measured?
  - Type of data
    (i.e. continuous, dichotomous, categorical, etc.)
  - Single item, multiple items, new/previously validated measure
  - Cross-sectional vs. cohort study (i.e. one measure in time vs. multiple measures over time)

Descriptive Statistics

- Exploratory data analysis (EDA)
- Basic numerical summaries of data (i.e. Table 1 in a paper)
- Basic graphical summaries of data
- Goal: to visualize relationships and generate hypotheses

Basis of Statistics



http://www.gs.washington.edu/

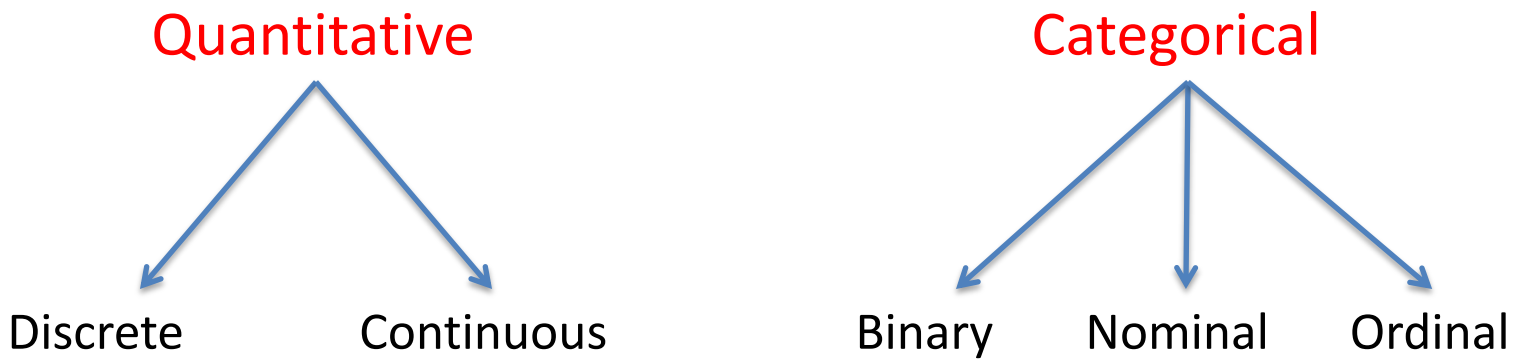
Exploratory Data Analysis (EDA)

- Essential first step of data analysis
- Helps to:
  - Identify errors
  - Visualize distributions and relationships
  - See patterns, e.g. natural or unnatural
  - Find violations of statistical assumptions
  - Generate hypotheses

Look

GFAP	gfaptime 0	1	2	3	4	Total
0	19	30	29	26	20	124
.04	3	1	1	0	2	7
.042	1	0	0	0	0	1
.046	0	0	0	1	1	2
.048	1	0	0	0	0	1
.049	0	0	1	0	0	1
.052	0	0	0	1	0	1
.053	0	1	0	0	0	1
.054	0	0	0	0	1	1
.063	0	0	0	0	1	1
.065	0	1	0	0	0	1
.069	1	0	0	0	0	1
.074	1	0	0	0	0	1
.08	0	0	0	0	1	1
.081	0	1	0	0	0	1
.089	1	0	0	0	0	1
.092	0	0	0	0	1	1
.095	0	0	0	1	0	1
.098	1	0	0	0	0	1
.102	0	0	1	0	0	1
.105	1	0	0	0	0	1
.106	0	0	0	1	0	1
.11	0	0	1	0	0	1
.119	0	0	0	0	1	1
.12	0	0	0	1	0	1
.137	1	0	0	0	0	1
.138	0	1	0	0	0	1
.141	1	0	0	0	0	1
.164	0	1	0	1	0	2
.172	0	0	0	0	1	1
.204	0	0	0	0	1	1
.223	1	0	0	0	0	1
.262	0	0	1	0	0	1
.29	0	0	0	1	0	1
.303	0	0	0	1	0	1
.328	0	0	0	1	0	1
.35	0	0	0	0	1	1
.566	0	0	1	0	0	1
.574	0	0	1	0	0	1
.651	0	0	1	0	0	1
.904	0	1	0	0	0	1
.985	0	1	0	0	0	1
1.03	0	0	0	0	1	1
1.236	0	0	1	0	0	1
.	10	4	4	7	10	35
Total	42	42	42	42	42	210

Types of Data



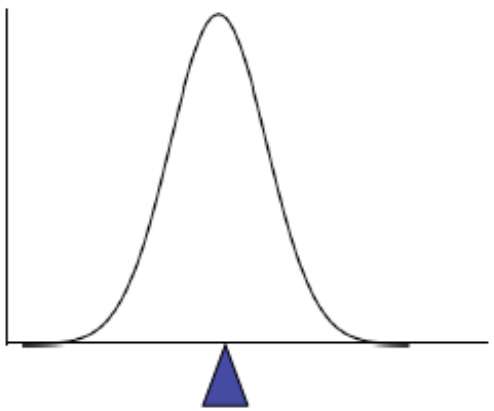
Numerical Summaries of Data

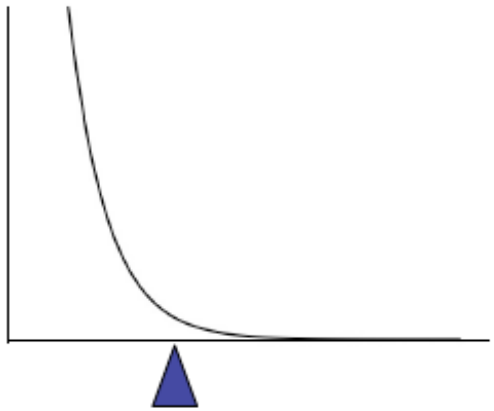
- Central tendencies measures
  - Calculated to create a “center” around which measurements in the data are distributed
- Variation or variability measures
  - Describe how far away (or data spread) measurements are from the center
- Relative standing measures
  - Describe the position (or standing) of specific measurements within the data

Location: Mean

- The average of a set of observations
- Add values and divide by the number of observations

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



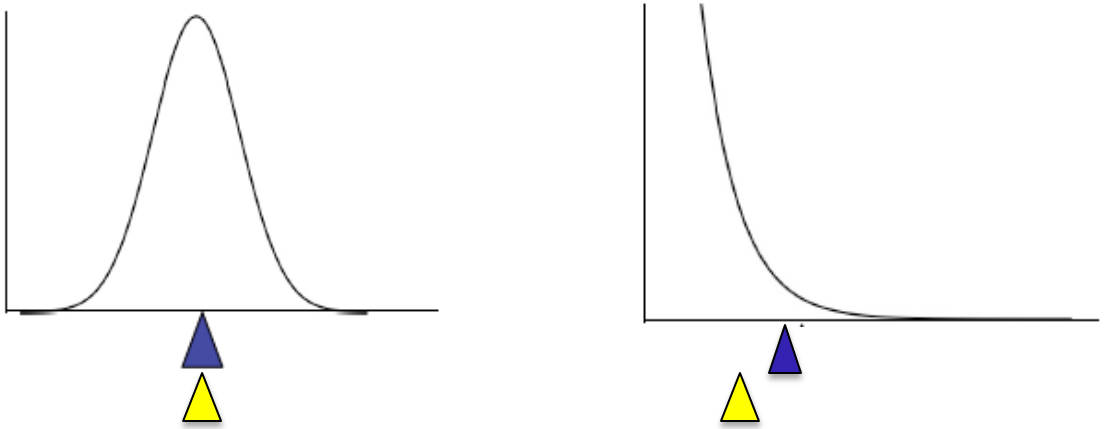


Location: Median

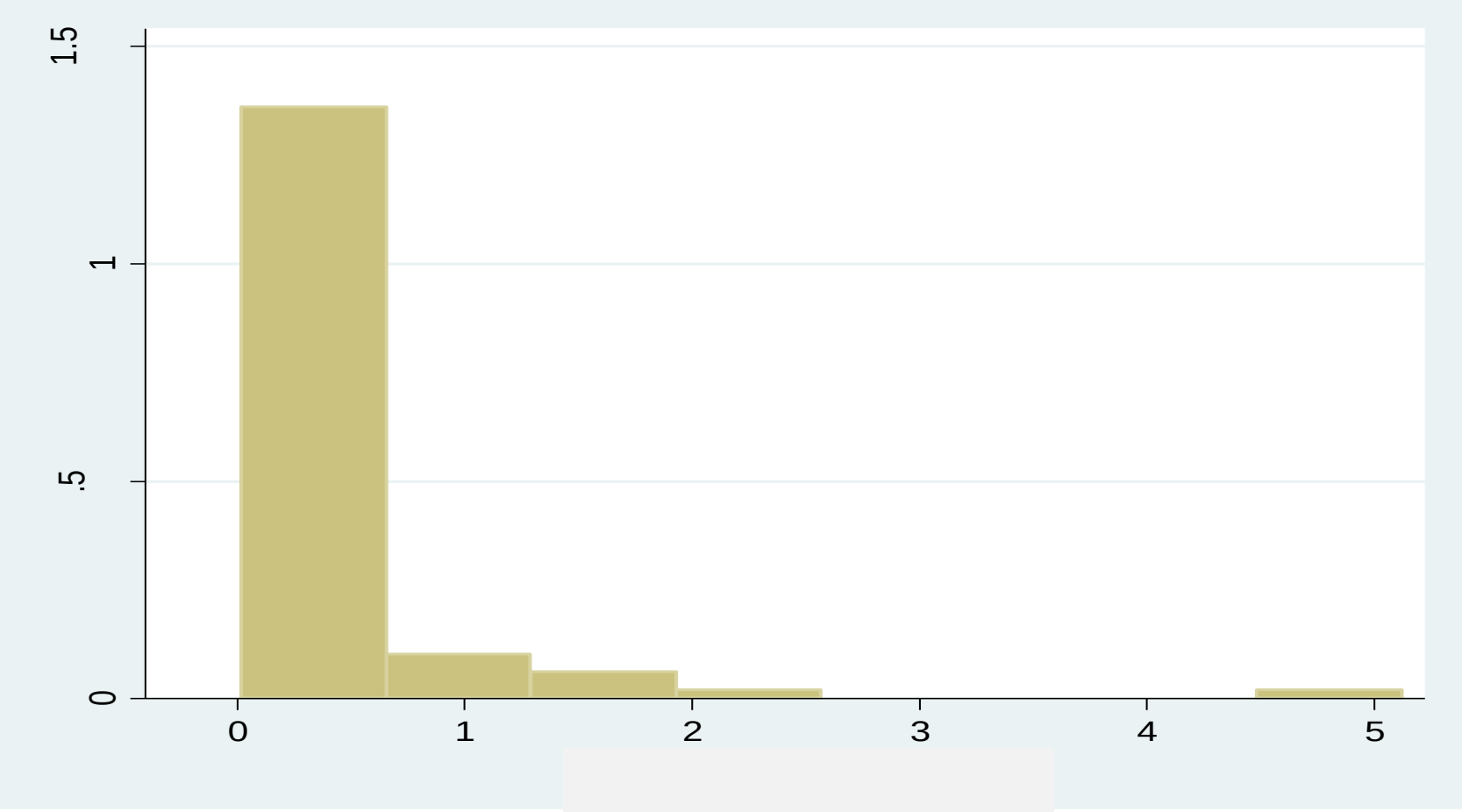
- The exact middle value, i.e. 50th percentile
- Number of observations
  - Odd: find the middle value
  - Even: find the middle two values and average them
- Example
  - Odd: 5, 6, 10, 3, 4, median = 10
  - Even: 5, 6, 10, 8, 3, 4, median = 10+8/2= 9

Which Measure is Best?

- Mean
  - best for symmetric (or normal) distributions
- Median
  - Useful for skewed distributions or data with outliers



Biomarker – one time point



Examples of Numerical Summaries

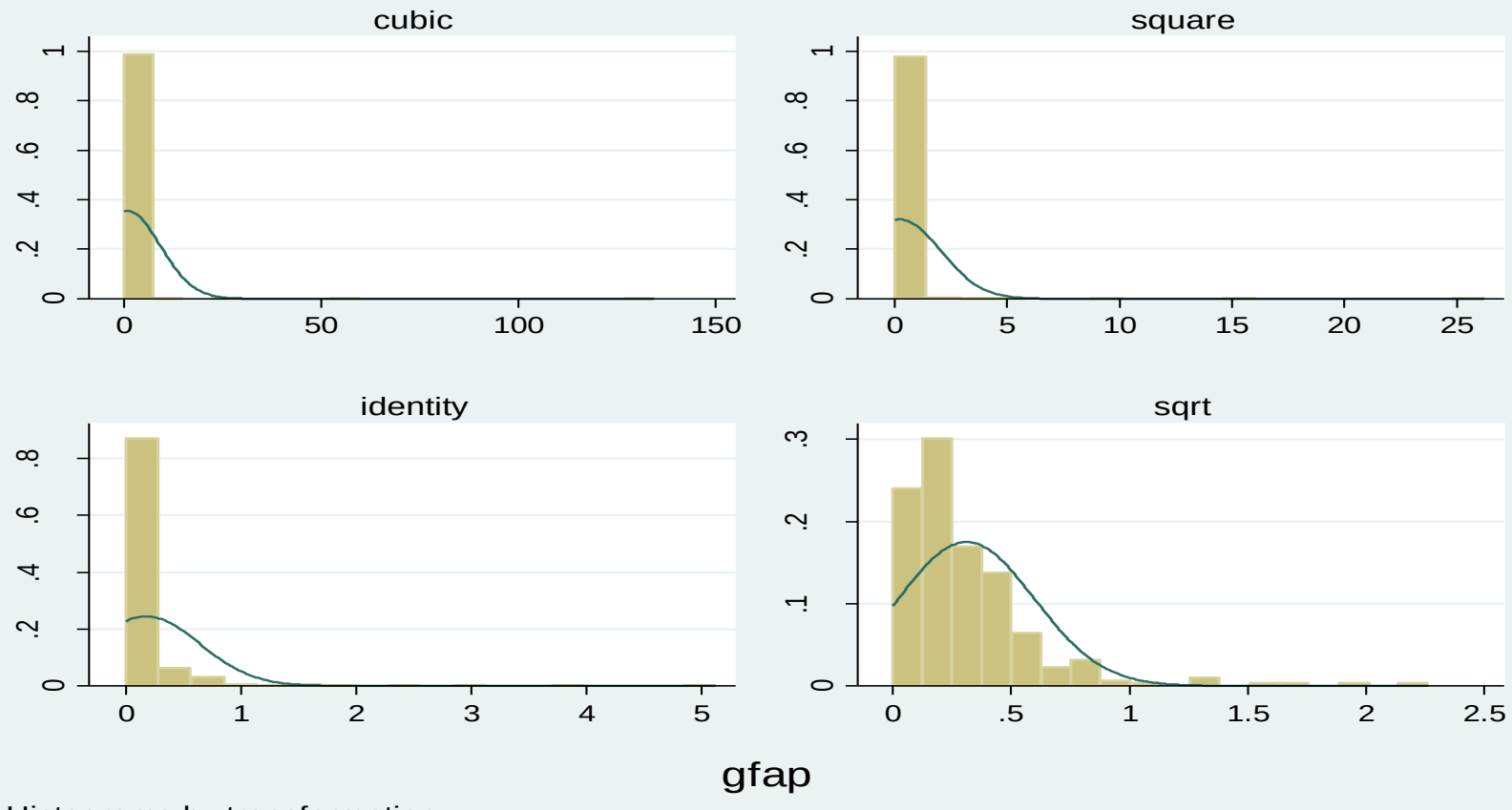
-> pvl = control

Variable	Obs	Mean	Std. Dev.	Min	Max
gfap0	16	.0231875	.0357122	0	.105
gfap1	17	.0061765	.0179869	0	.065
gfap2	18	0	0	0	0
gfap3	18	0	0	0	0
gfap4	14	.0106429	.0216603	0	.063

-> pvl = case

Variable	Obs	Mean	Std. Dev.	Min	Max
gfap0	16	.0484375	.0686838	0	.223
gfap1	21	.1107143	.281544	0	.985
gfap2	20	.1795	.3286394	0	1.236
gfap3	17	.0884706	.1164072	0	.328
gfap4	18	.1189444	.2465624	0	1.03

Transformation



Scale: Variance

- Average of the squared deviations of values from the mean
- Example, sample variance

$$\hat{\sigma}^2 = \frac{\sum_{i}^{n} (x_i - \bar{x})^2}{n-1}$$

Scale: Standard Deviation

- Variance is somewhat arbitrary
- Standardizing helps to bring meaning to deviation from the mean
- Standard deviations are simply the square root of the variance
- Example, sample SD

$$\hat{\sigma} = \sqrt{\frac{\sum_{i}^{n} (x_i - \bar{x})^2}{n-1}}$$

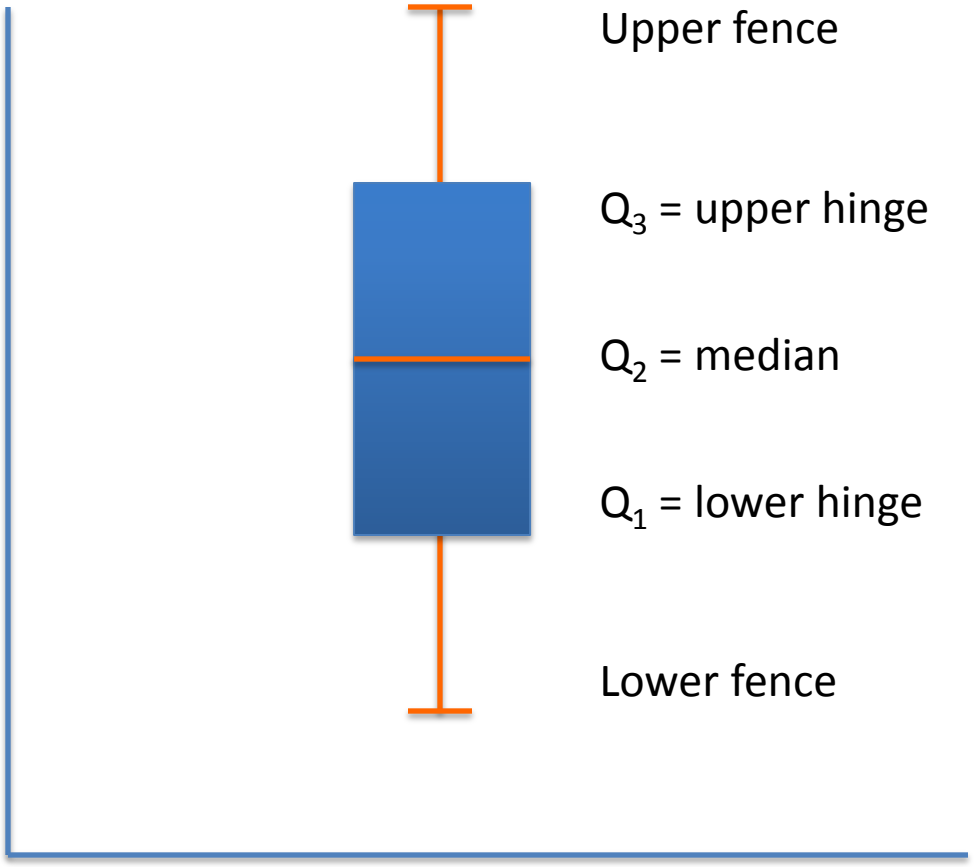
Scale: Quartiles and Inter Quartile Range (IQR)

- Quartiles or percentiles (order data first)
  - Q_1 (1st quartile) or 25th percentile is the value for which 25% of the observations are smaller and 75% are greater
  - Q_2 is the median or the value where 50% of the observations are smaller and 50% are greater
  - Q_3 is the value where 75% of the observations are smaller and 25% are greater

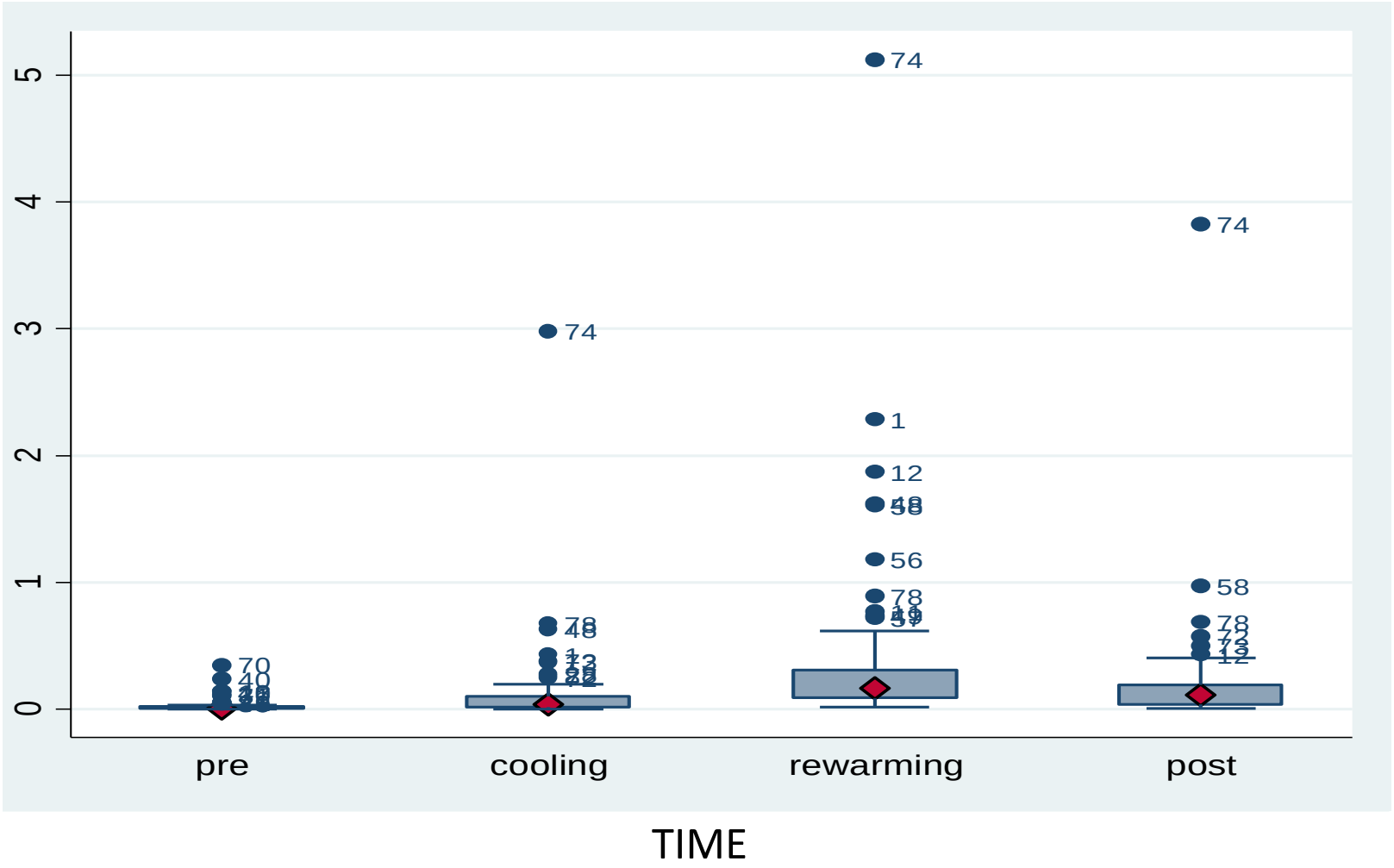
Graphical Summaries of Data: Box Plots and Histograms

- Box plot (i.e. box-and-whisker plots)
  - Shows frequency or proportion of data in categories, i.e categorical data
  - Visual of frequency tables
- Histogram
  - Shows the distribution (shape, center, range, variation) of continuous variables
  - Bin size is important

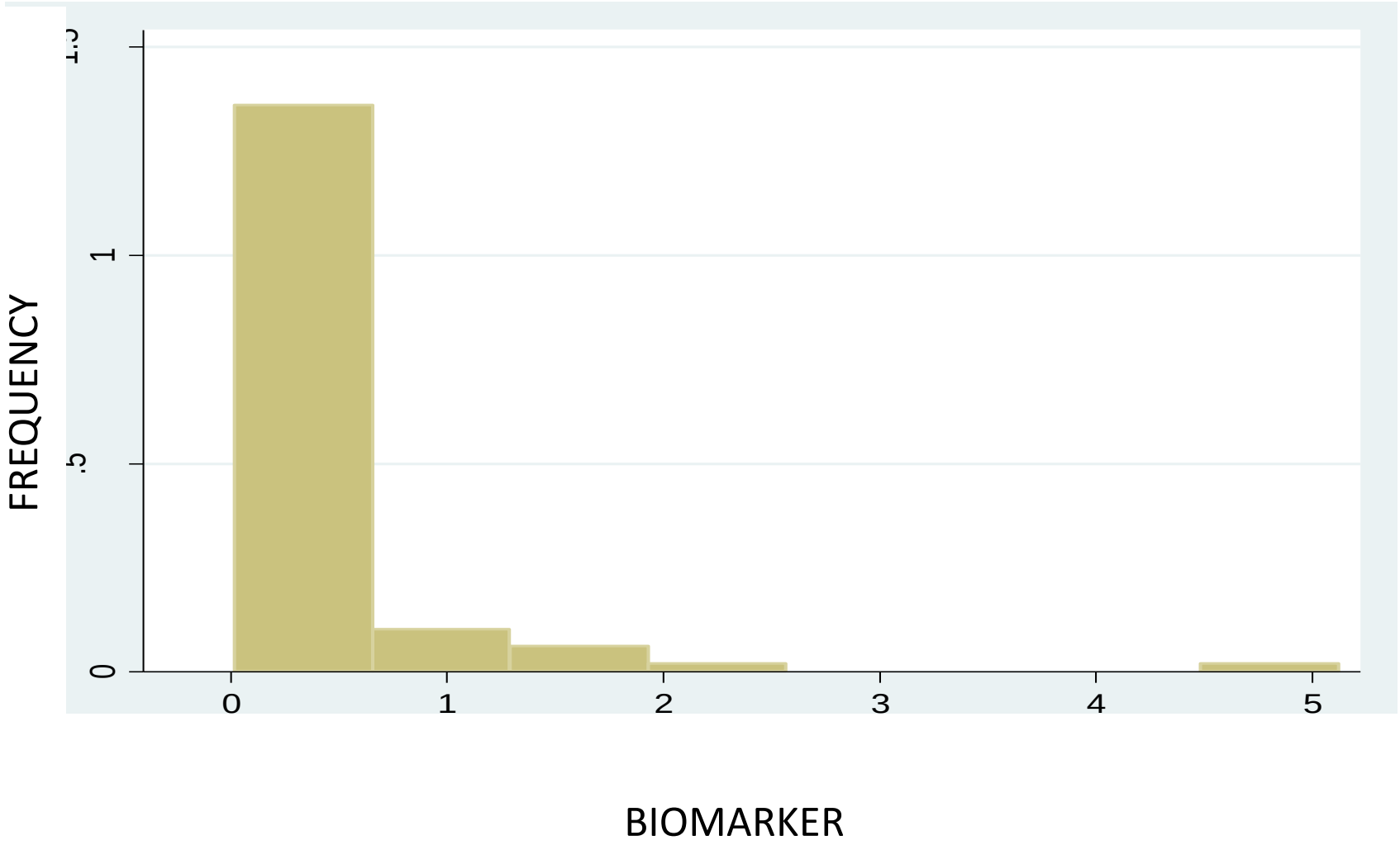
Box Plot


Upper fence
Q_3 = upper hinge
Q_2 = median
Q_1 = lower hinge
Lower fence

Box Plot



Histogram



Examples of Numerical Summaries

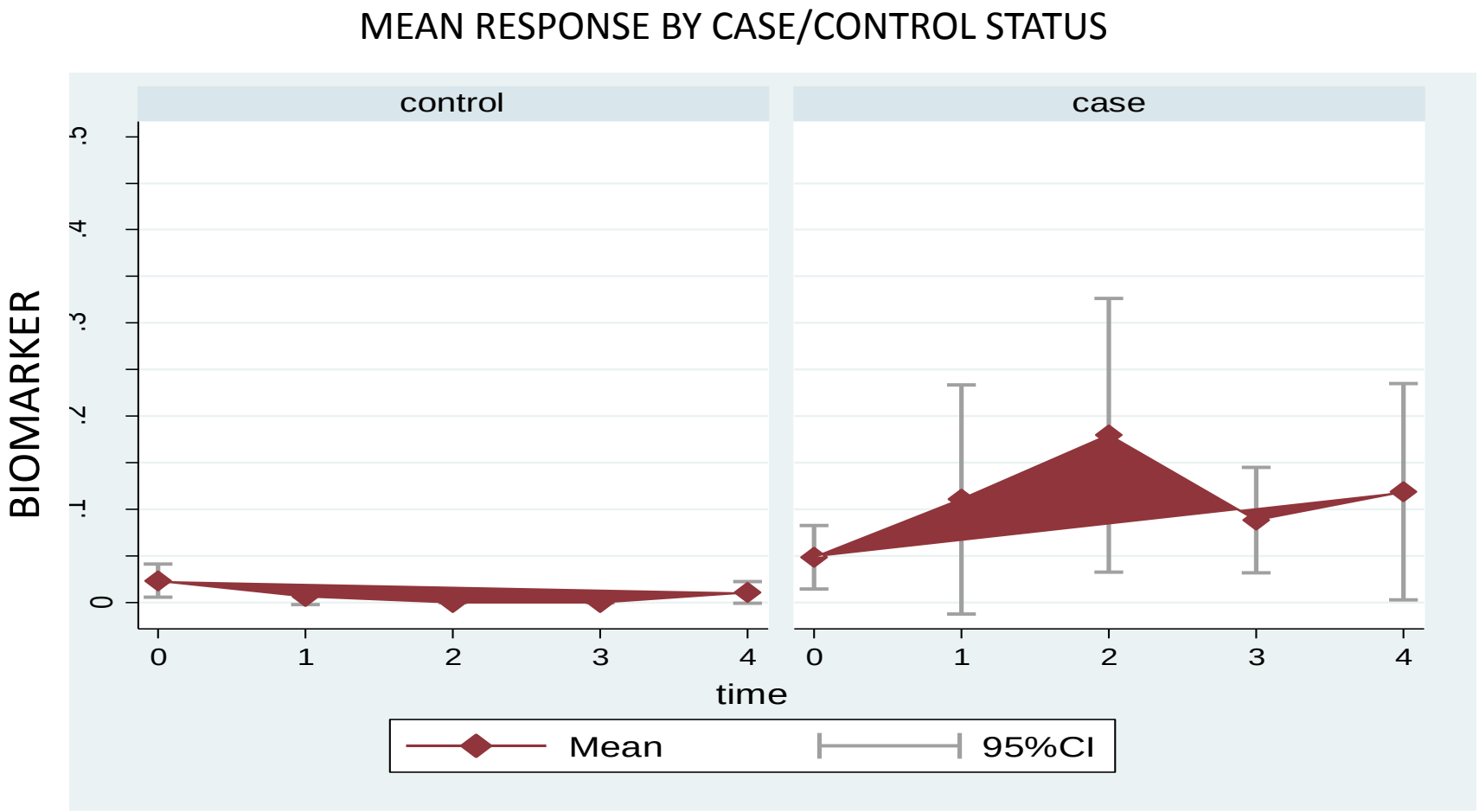
CONTROL

Variable	Obs	Mean	Std. Dev.	Min	Max
gfap0	16	.0231875	.0357122	0	.105
gfap1	17	.0061765	.0179869	0	.065
gfap2	18	0	0	0	0
gfap3	18	0	0	0	0
gfap4	14	.0106429	.0216603	0	.063

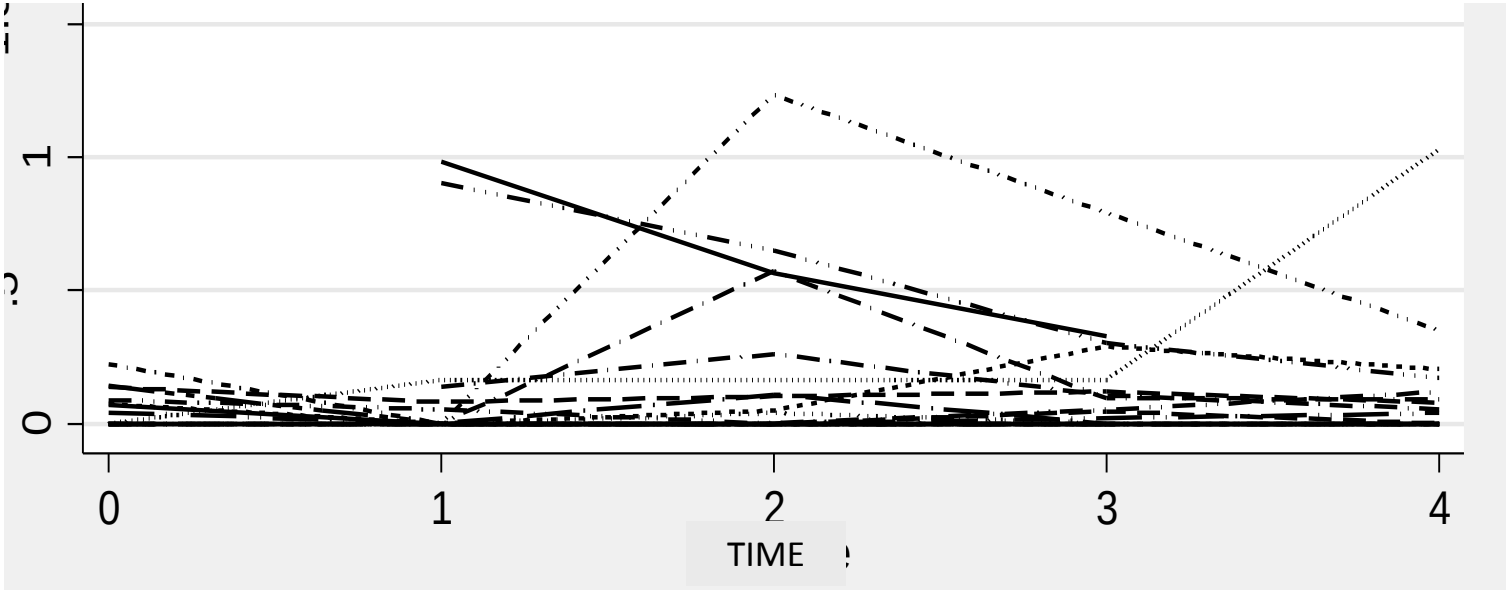
CASE

Variable	Obs	Mean	Std. Dev.	Min	Max
gfap0	16	.0484375	.0686838	0	.223
gfap1	21	.1107143	.281544	0	.985
gfap2	20	.1795	.3286394	0	1.236
gfap3	17	.0884706	.1164072	0	.328
gfap4	18	.1189444	.2465624	0	1.03

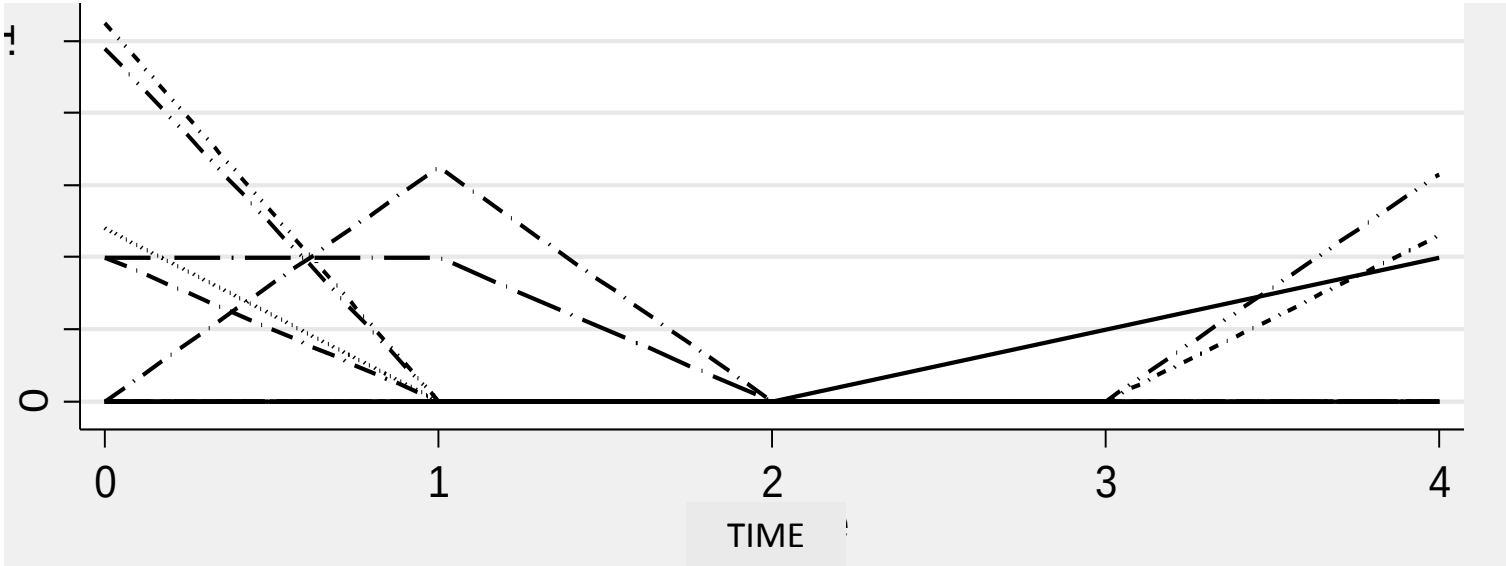
Another Way to Visualize



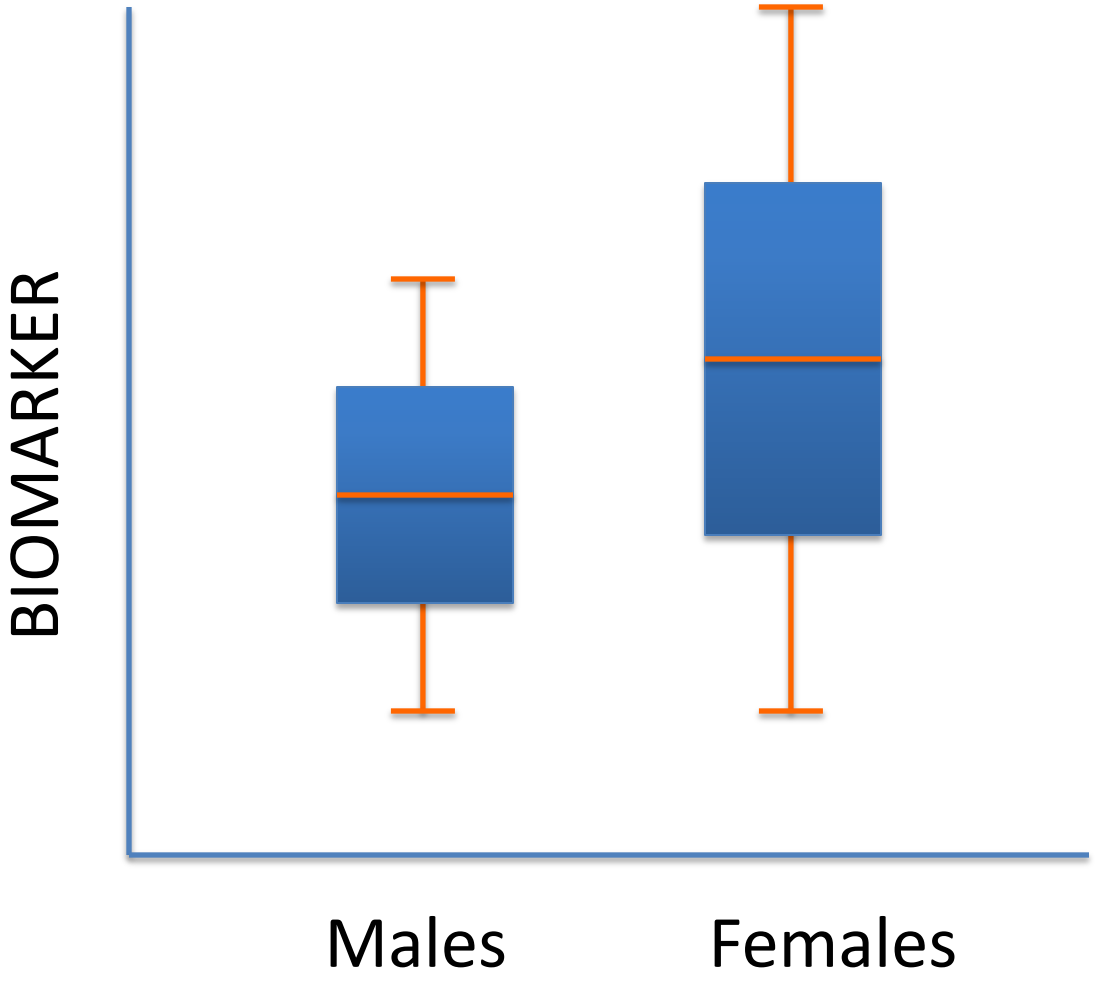
INDIVIDUAL BIOMARKER LEVEL CHANGE OVER TIME AMONG CASES



INDIVIDUAL BIOMARKER LEVEL CHANGE OVER TIME AMONG CONTROLS



Side-by-Side Box Plot



Bivariate Data

Variable 1	Variable 2	Display
Categorical	Categorical	Crosstabs Stacked Box Plot
Categorical	Continuous	Boxplot
Continuous	Continuous	Scatterplot Stacked Box Plot

Dos and Do Nots of Graphing

- Goal of graphing
  - To portray data accurately and clearly

- Rules of graphing
  - Label and appropriately scale axis
  - Simplify, display only the necessary information
  - Stay away from pie charts

Take Homes

- Important basic steps in data analyses
  - Include exploratory data analyses and summary statistics
- Main rationale for exploratory quantitative data analysis
  - Get to know your data so that your methods and inferences will be appropriate
- Statistical tools for inferential statistics
  - They are vast, we covered just a few