Study Design and Analysis in Epidemiology:

Where does modeling fit?

Clinic on the Meaningful Modeling of Epidemiological Data, 2016
African Institute for Mathematical Sciences
Muizenberg, South Africa

Steve Bellan, PhD, MPH
Center for Computational Biology & Bioinformatics
University of Texas at Austin

Defining Epidemiology

"The study of the distribution and determinants of health related states and events in populations, and the application of this study to control health problems."

John M Last
Dictionary of Epidemiology

Varieties of Infectious Disease Epidemiology

- Risk Factors & Intervention Epidemiology

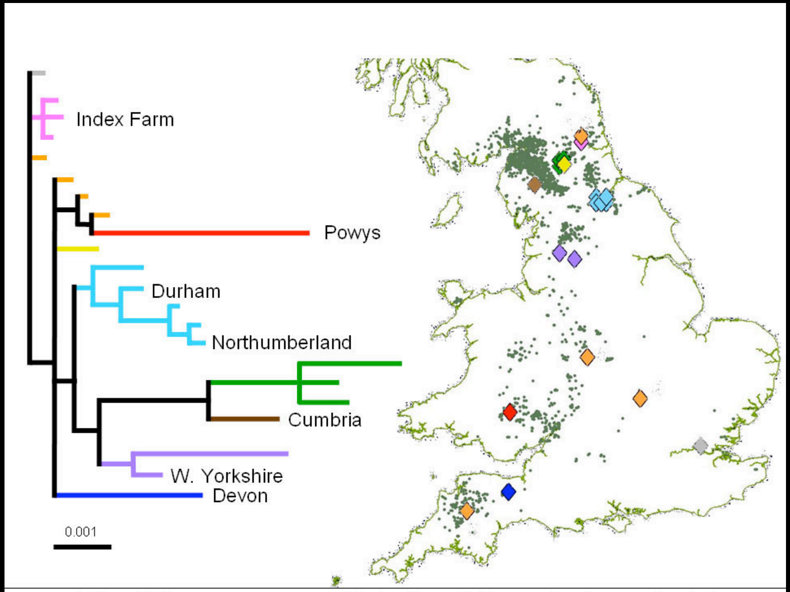
  Risk Factor: A characteristic that is correlated with a measure of disease.

  - Often used synonymously with *covariate*.

  - Protective factors: Risk factors that are negatively associated with disease

Varieties of Infectious Disease Epidemiology

- Risk Factors & Intervention
- Outbreak



Use of full genome sequences of foot-and-mouth disease virus to reconstruct individual transmission events that occurred during the 2001 outbreak.

- Clinical
- Molecular & Genetic
- Surveillance

How does mathematical modeling fit?

- Subfield:
  Linking pattern with process across scales

BUT ALSO

- Methodologies used in other epi subfields

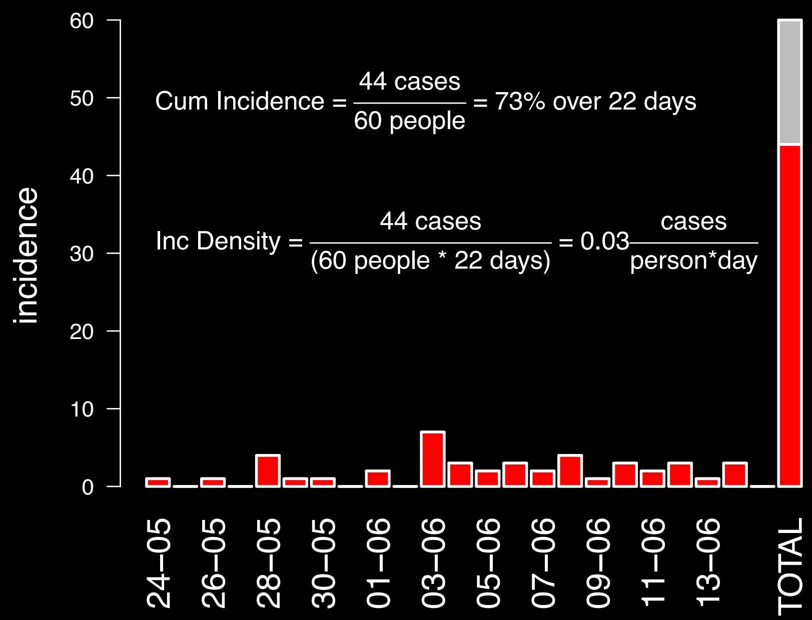
Importance of knowledge breadth

What do *Introductory Epidemiology* courses teach?

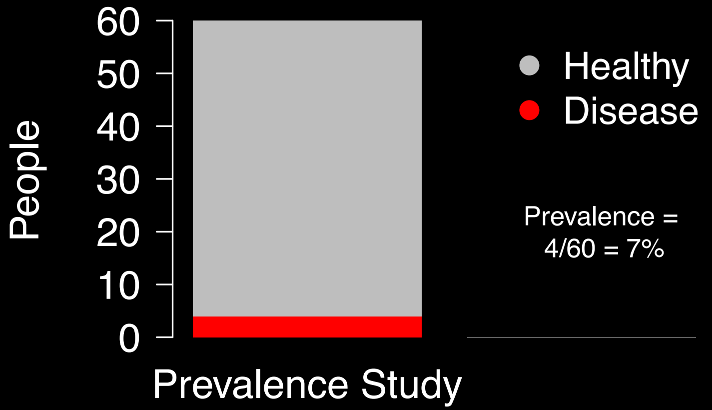
- Measures of Disease
- Measures of Effect (of a risk factor)
- Study Designs for Measuring Effects
  - Dealing with random error
  - Dealing with confounding
  - Dealing with bias
- Biostatistical analyses for analyzing data

Measures of Disease

- Incidence
  - Cumulative Incidence
  - Incidence Density



- Prevalence
  - Point Prevalence
  - Period Prevalence



- Survivorship (time to event, e.g. death)

Measures of Covariates (risk factors)

- Binary: gender, smoker, circumcised
- Nominal/Categorical: geographic region
- Continuous: birth weight, T-cell count
- Ordinal: education, socioeconomic status (SES)

Measures of Effect

- How do you measure the effect of a risk factor on a disease?

Example

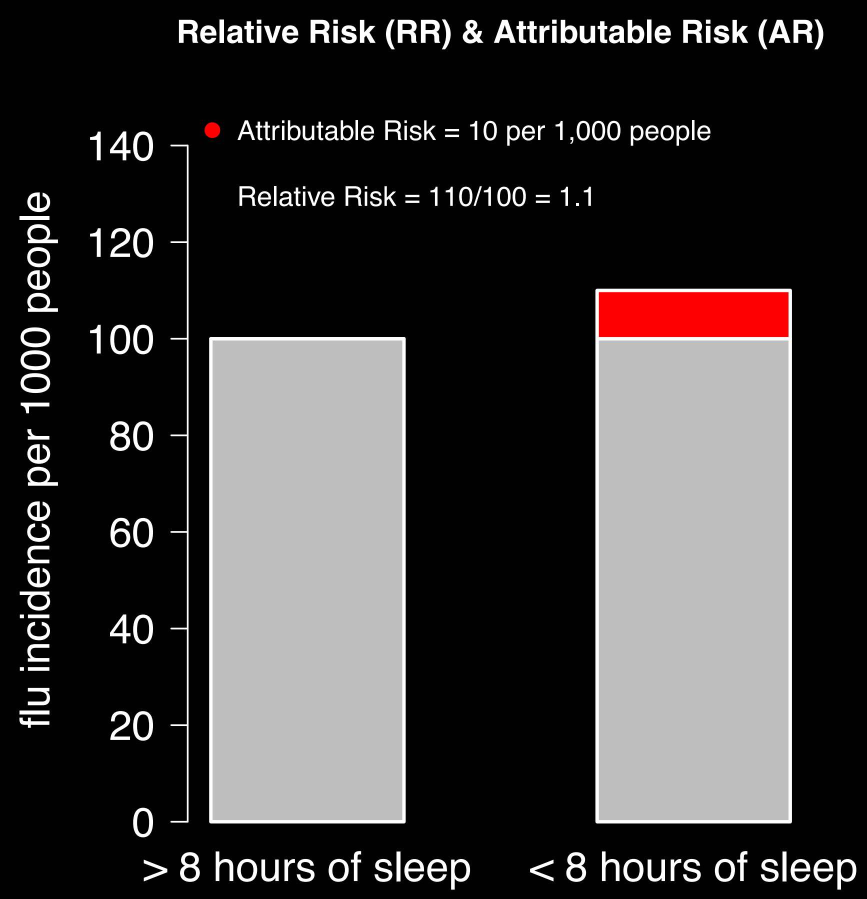
How could you measure whether circumcision reduces the risk of HIV infection?

Measures of Effect

- Compare measure of disease across levels/values of risk factors

- Relative Risk
  Ratio of rates or proportions
    - Prevalence Ratio
    - Cum. Incidence Ratio
    - Incidence Density Ratio
    - Odds Ratio

- Attributable Risk
  Subtract rates or proportions



Contingency Tables: Relative Risk (RR)

	Disease	No Disease	Total (Margins)
Exposed	a	b	a+b
Not exposed	c	d	c+d
Total (Margins)	a+c	b+d	a+b+c+d

Cumulative Incidence Ratio (CIR):
cumulative incidence in exposed population divided by cumulative incidence in unexposed population.

$$\mathrm{CIR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

CIR < 1 exposure correlates with reduced risk of disease
CIR > 1 exposure correlates with increased risk of disease

Epidemiologic Studies

- Descriptive Epidemiology
  - Baseline data on distribution of disease
  - Surveillance
- Analytic Epidemiology – Measure Effect
  - Prospective Cohort Studies
  - Cross-sectional Studies
  - Retrospective Case-Control Studies
  - Ecologic Studies
  - Randomized Controlled Trials

Observational (Prospective Cohort, Cross-sectional, Retrospective Case-Control, Ecologic Studies)

Experimental (Randomized Controlled Trials)

Cohort Studies

- Follow a selected population through time
  - Establishes temporal relationships
  - Can measure incidence
- Takes lots of resources, money, & time!
- Poor design for rare diseases.

Cohort Data and Person-Time

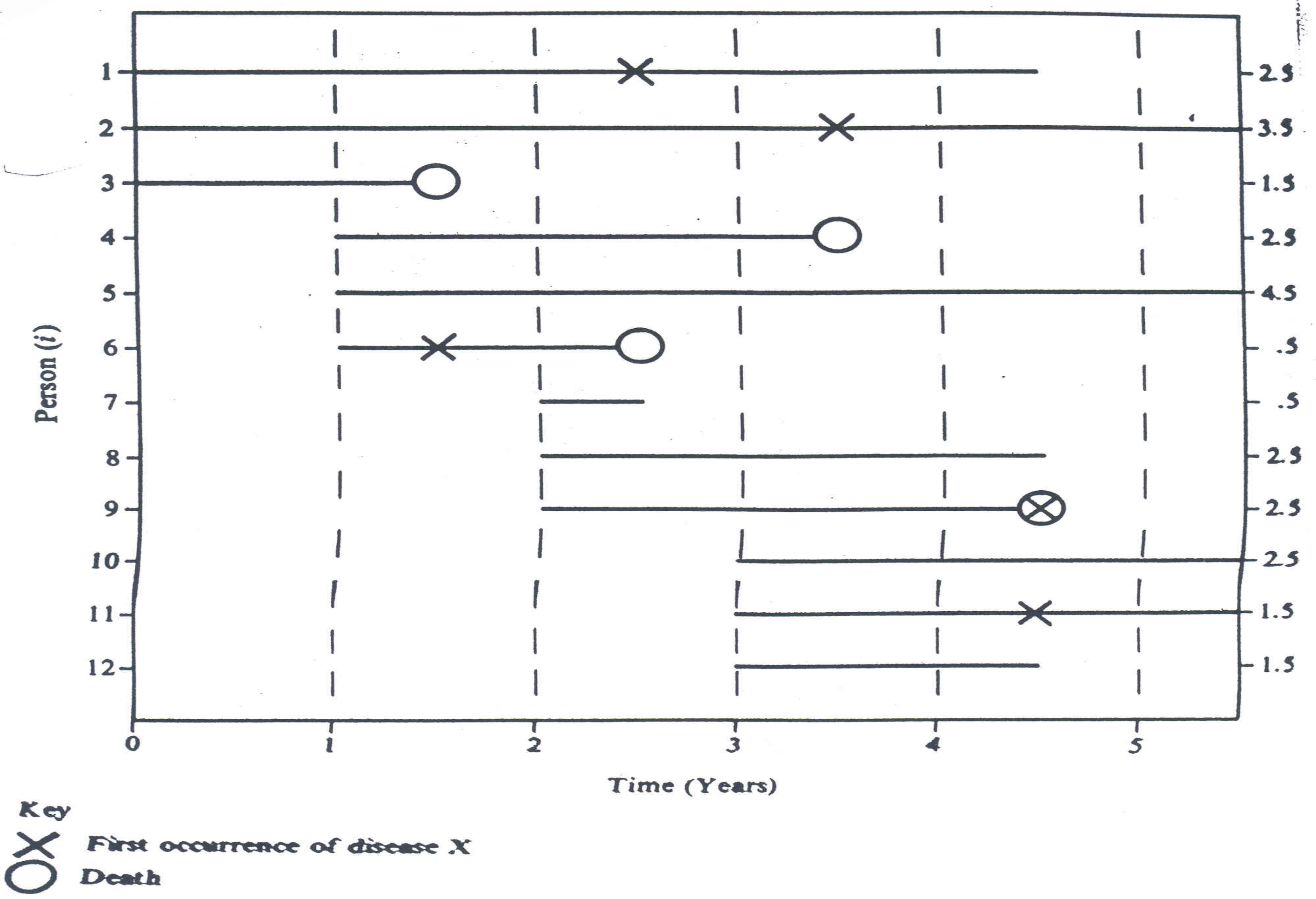
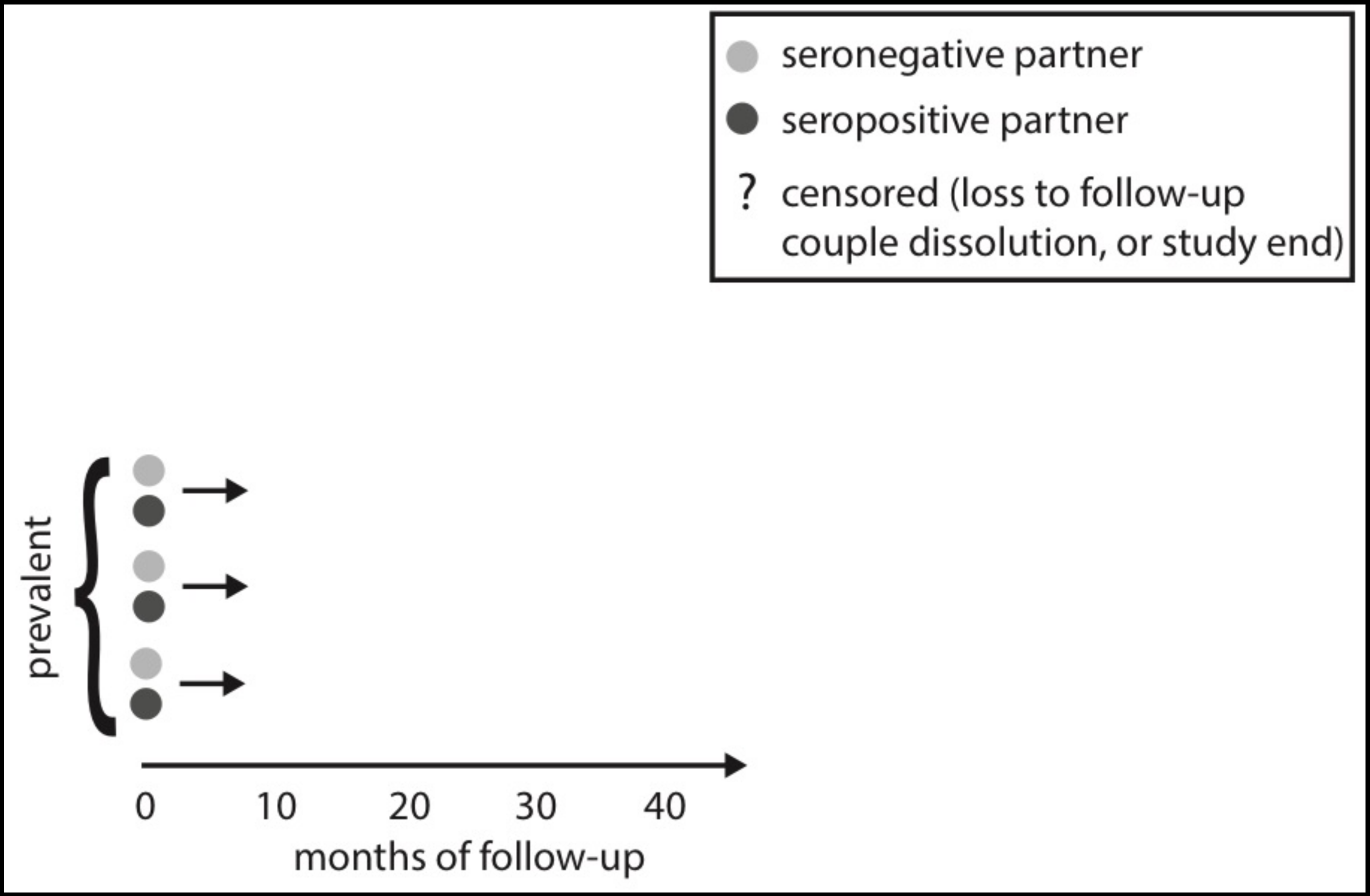


Figure 6.1 **Diagrammatic Representation of the 5.5-Year Follow-up of a Hypothetical Cohort of 12 Subjects Initially Free of Disease X**

Couples Cohort Data

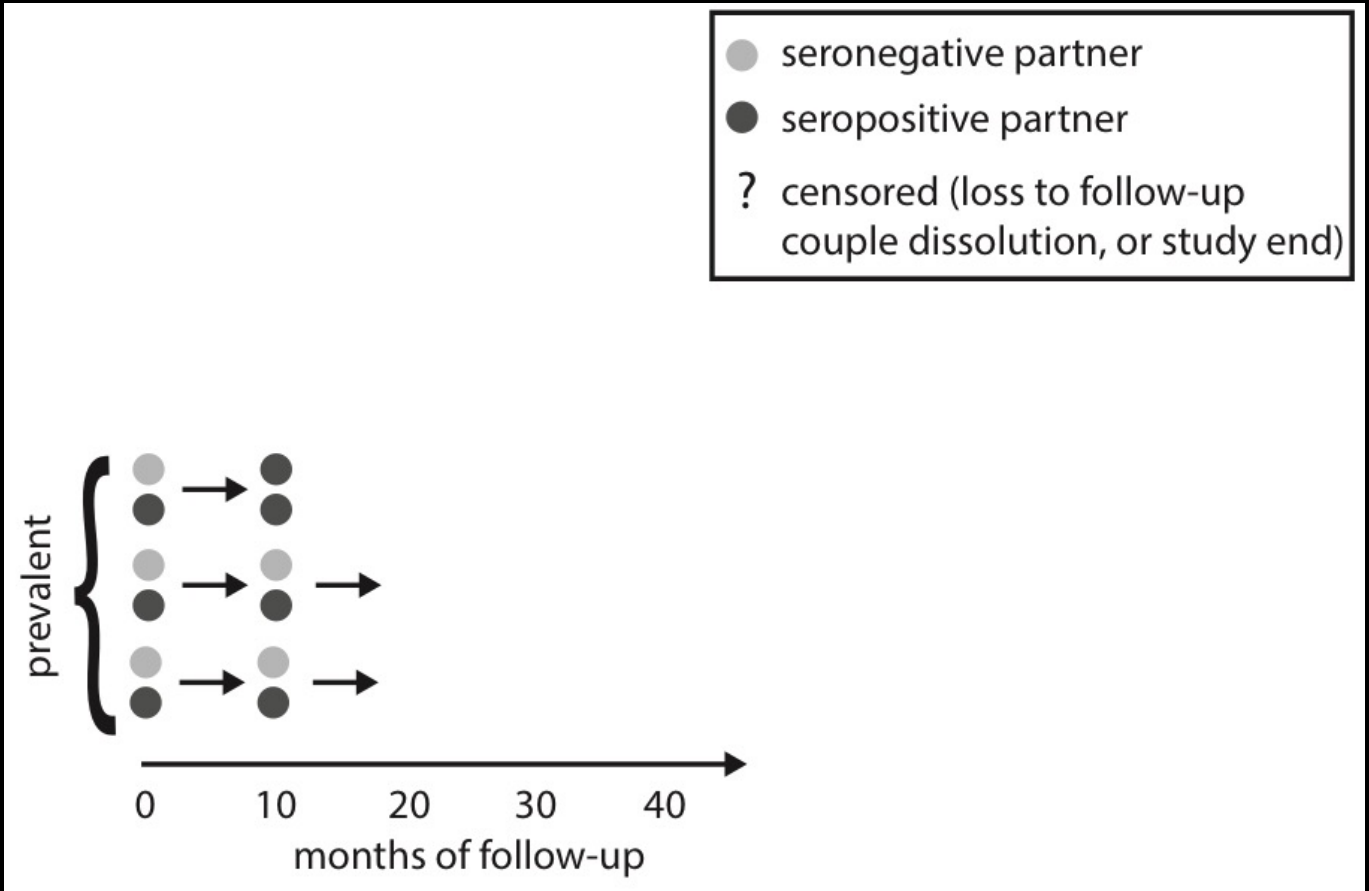
chronic



Bellan et al. (2015)

Couples Cohort Data

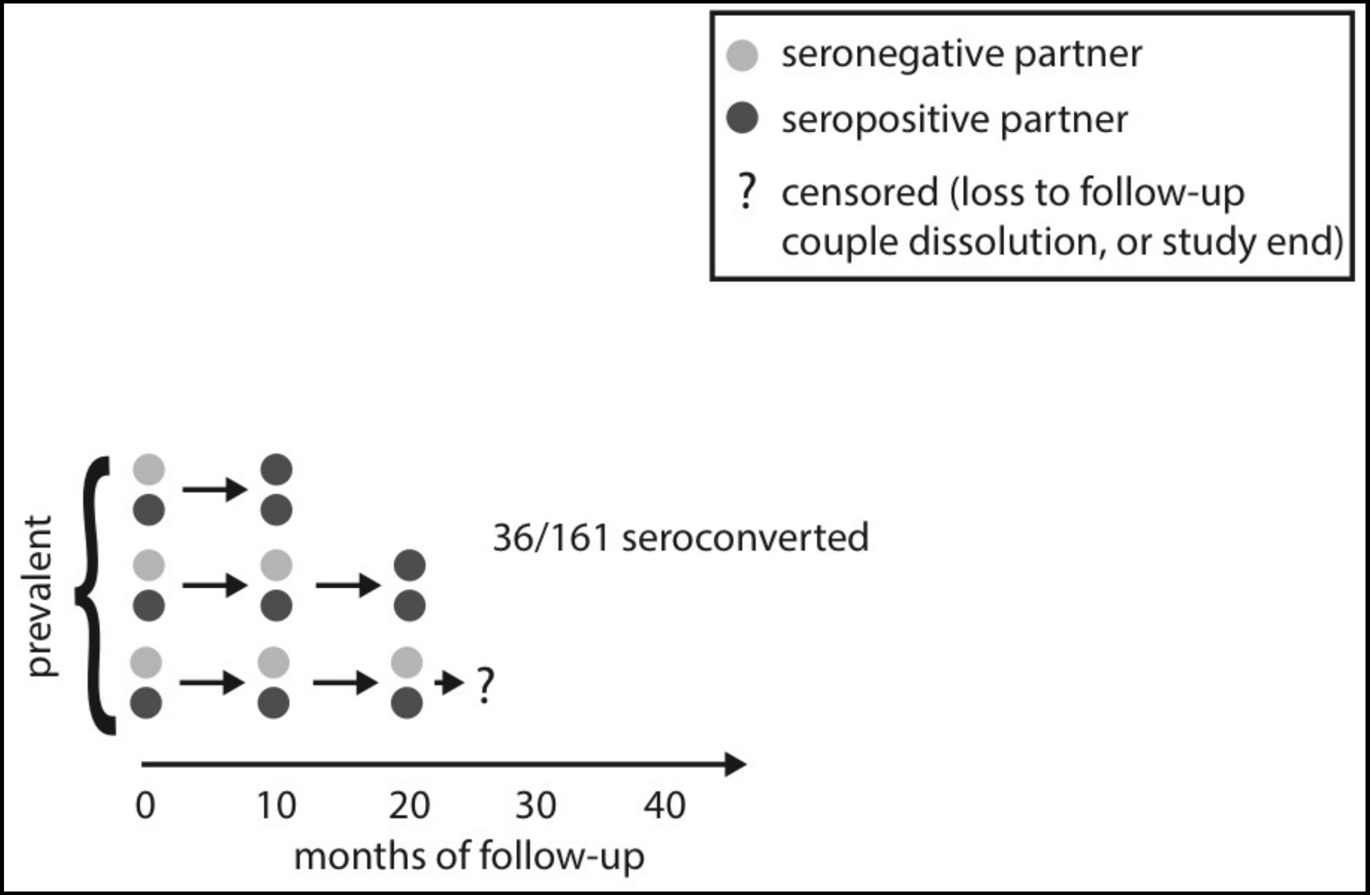
chronic



Bellan et al. (2015)

Couples Cohort Data

chronic

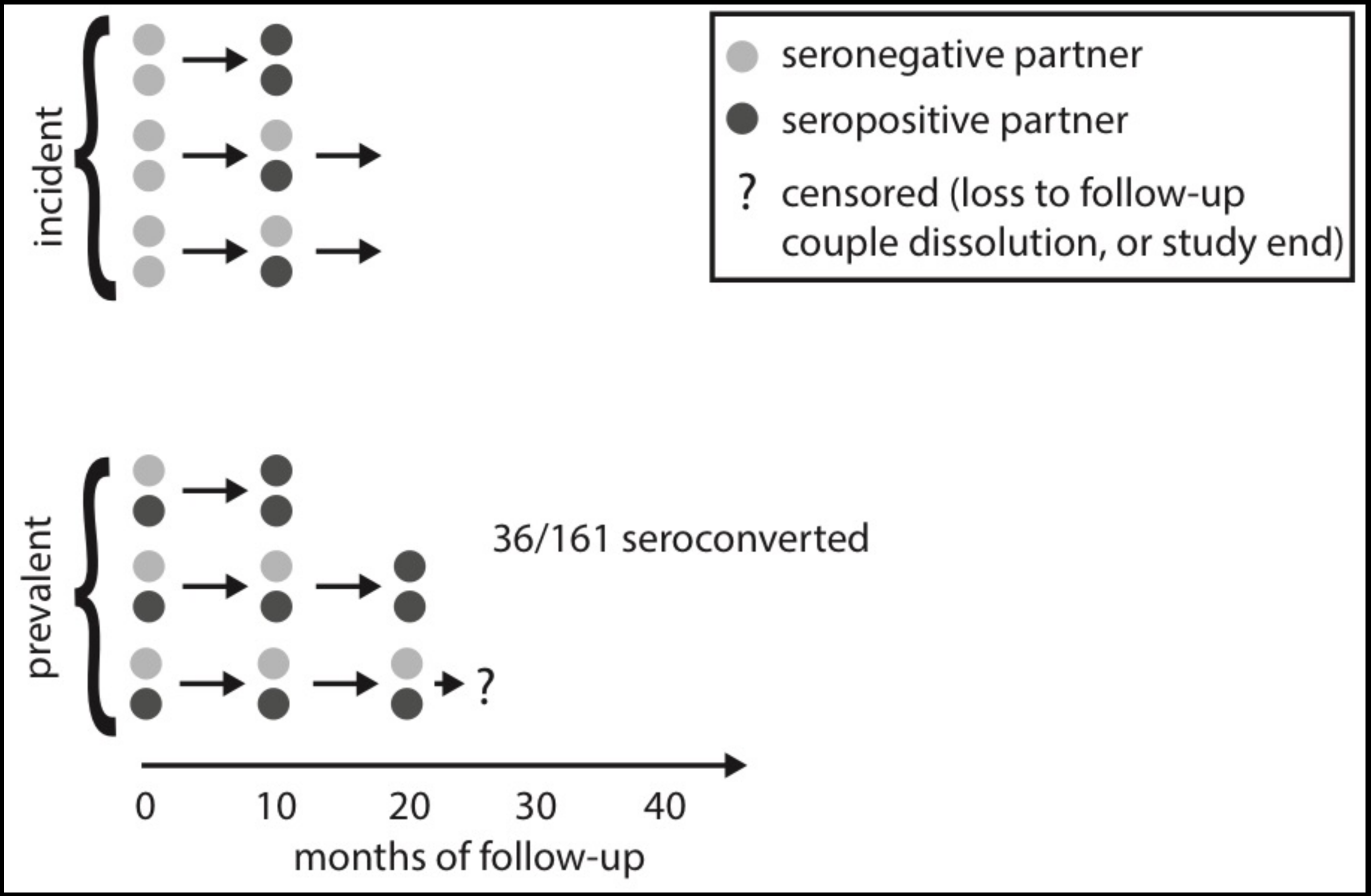


Bellan et al. (2015)

Couples Cohort Data

acute

chronic



Bellan et al. (2015)

Relative Risk: Incidence Density Ratios

	Disease	No Disease	Total (Margins)
Exposed	a	-	PY_e
Not exposed	c	-	PY_0
Total (Margins)	a+c	-	$PY_e + PY_0$

Incidence Density Ratio is the ratio of incidence density of the exposed population to that of the unexposed population.

$$\text{IDR} = \frac{\frac{a}{PY_e}}{\frac{c}{PY_0}}$$

IDR < 1 means exposure correlates with reduced risk of disease
IDR > 1 means exposure correlates with increased risk of disease

Cross-Sectional Studies

- Snapshot of diseases & risk factors.
- Cannot establish temporal relationship.
- Relatively cheap & easy.
- Population must be large to study rare disease
- Not great for diseases of short duration. Why?

Case-Control Studies

- Compare diseased individuals to chosen controls.
  - Quality of study depends entirely on how controls are chosen.
- Good for rare diseases.
- Relatively cheap & quick.

Case Control Studies: Odds Ratios

$$\mathrm{PR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Controls: Number chosen by researcher.

	Disease	No Disease	Total (Margins)
Exposed	a	b	a+b
Not exposed	c	d	c+d
Total (Margins)	a+c	b+d	a+b+c+d

Odds ratio is the ratio of odds in the diseased population divided by the odds in the non-diseased population.

$$\mathrm{OR} = \frac{a/c}{b/d}$$

OR < 1 means exposure correlates with reduced risk of disease
OR > 1 means exposure correlates with increased risk of disease

Randomized Controlled Trials

- Experimental or Intervention Studies
- Establishes temporal relationships
- Addresses confounding (more to come)

Ecologic Studies

- Measurements made at population rather than individual level.

- Weaker inference, but easier to gather data.

Measures of Covariates (risk factors)

- Binary: gender, smoker, circumcised
- Nominal/Categorical: geographic region
- Continuous: birth weight, T-cell count
- Ordinal: education, socioeconomic status (SES)

What do *Introductory Epidemiology* courses teach?

- Measures of Disease
- Measures of Effect (of a risk factor)
- Study Designs for Measuring Effects
  - Dealing with random error
  - Dealing with confounding
  - Dealing with bias
- Biostatistical analyses for analyzing data

Random Error

- How many people must be in a study for the measure of effect to believable?

- Statistical Approach: Assign probabilities to our findings being a product of random error rather than a real phenomenon.

Bias

Difference between observed value and true value due to all causes other than random error.

Bias does not go away with greater sample size!

Bias must be dealt with during study design!

Selection Bias

Error due to systematic differences between those who take part in the study and those who do not.

John Last, Dictionary of Epidemiology

Information Bias

A flaw in measuring exposure or outcome data that results in different quality (accuracy) of information between comparison groups.

John Last, Dictionary of Epidemiology

Confounding

	HIV+	HIV-
Literate	660	340
Illiterate	180	820

$$PR = \frac{660/_{1000}}{180/_{1000}} = 3.67$$

What if some of the study population were much younger than others?

Confounding

Pooled	HIV+	HIV-
Literate	660	340
Illiterate	180	820

$$PR_{all} = \frac{660/_{1000}}{180/_{1000}} = 3.67$$

6-15 years old	HIV+	HIV-
Literate	30	270
Illiterate	90	810

$$PR_{6-15yrs} = \frac{30/_{300}}{90/_{900}} = 1$$

16-24 years old	HIV+	HIV-
Literate	630	70
Illiterate	90	10

$$PR_{16-24yrs} = \frac{630/_{700}}{90/_{100}} = 1$$

6-15 year olds: Literacy = 300/1200 = 25%

16-24 year olds: Literacy = 700/800 = 87.5%

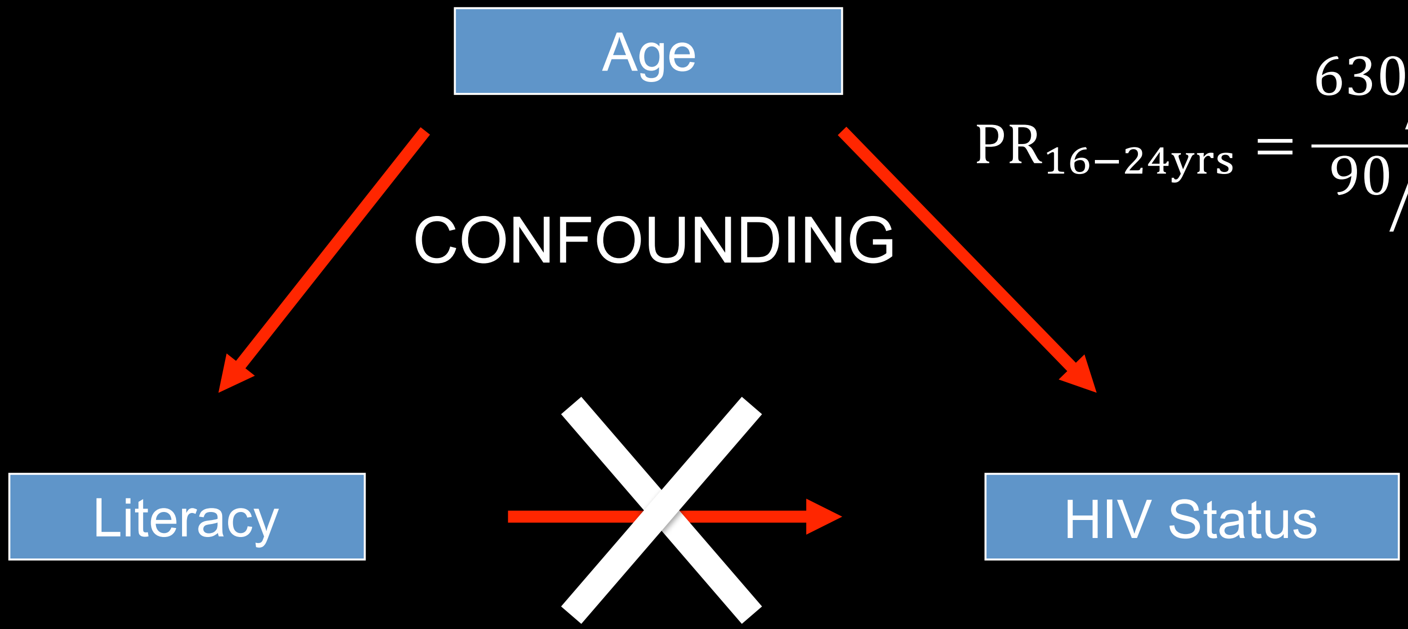
Confounding

	HIV+	HIV-
Literate	660	340
Illiterate	180	820

$$PR_{all} = \frac{660/1000}{180/1000} = 3.67$$

$$PR_{6-15yrs} = \frac{30/300}{90/900} = 1$$

$$PR_{16-24yrs} = \frac{630/700}{90/100} = 1$$



Biostatistical Analyses

- Permutation Tests
- Chi Squared Test
- Generalized Linear (Mixed) Models
  - Normal Regression
  - Logistic Regression
  - Poisson Regression
  - Negative Binomial Regression
- Survival Analysis

Statistical Models

- Account for bias and random error to find correlations that may imply causality.
- Often the first step to assessing relationships.
- Assume independence of individuals (at some scale, *i.e.* clusters).

Dynamic Models

- Systems Approach: Explicitly model multiple mechanisms to understand their interactions.
- Links observed relationships at different scales.
- Explicitly focuses on dependence of individuals

By developing dynamic models in a probabilistic framework we can account for dependence, random error, and bias while linking patterns at multiple scales.

Questions in Epidemiology

Statistical Models

- Is HIV status positively associated with the risk of TB infection?

Dynamic Models

- Based on increased TB risk due to HIV, how much should we expect TB notification rate to increase for a given HIV prevalence?

Questions in Epidemiology

Statistical Models

- Are Insecticide Treated Bednets (ITNs) or Indoor Residual Spraying (IRS) more effective for controlling malaria?

Dynamic Models

- How do we expect the age-distribution of malaria incidence to change after implementing ITNs or IRS?

Computer Labs

- Schedule says lab 5 – Intro to Likelihood lab
- if you are ***very*** comfortable with the likelihood material, you can choose between things that you think are higher priority for you
- There are many others on the list (stochastic algorithms etc)