Measures of Disease in Clinical Epidemiology

Brendan J. Kelly

Infectious Diseases & Epidemiology

Penn
University of Pennsylvania

updated: 2019-09-24

Disclosures

- No conflicts of interest.
- Too many topics covered.
- May contain GIFs.

Learning Objectives

- Describe types of data and data distributions.
- Use population data in the description of health and disease.
- Explain the meaning of:
    - prevalence
    - incidence (warning: multiple types)
    - relative risk (RR)
    - odds ratio (OR)
- Calculate prevalence, incidence, RR, and OR from study data.

Notes from John:

a. "Incidence rate" crept into the notes and lecture (I missed that again! sorry!). I think we need to remove this term and stick with cumulative incidence and incidence density. "Incidence rate" is still confusing (even to me because I think of it as just the 'colloquial' version of either.)

b. It would be good to have a separate slide on 'person time' and calculating person time.

c. I think we should use the term "Risk Difference" instead of "Absolute risk reduction." This should always be "Ie-Iu" but in one of the examples in the lecture, it's reversed and it's presented as "Iu-Ie" which was confusing to them. We could also define both terms and say RD=Ie-Iu and then absolute risk reduction has to do with the way in which the question is phrased (i.e., improvement is a good outcome or adverse event where the outcome is bad).

d. Calculating annual cumulative incidence on slides 23-24; I think the answer given during lecture was 0.3 but I think it should actually be 0.15. This was confusing to them (and me).

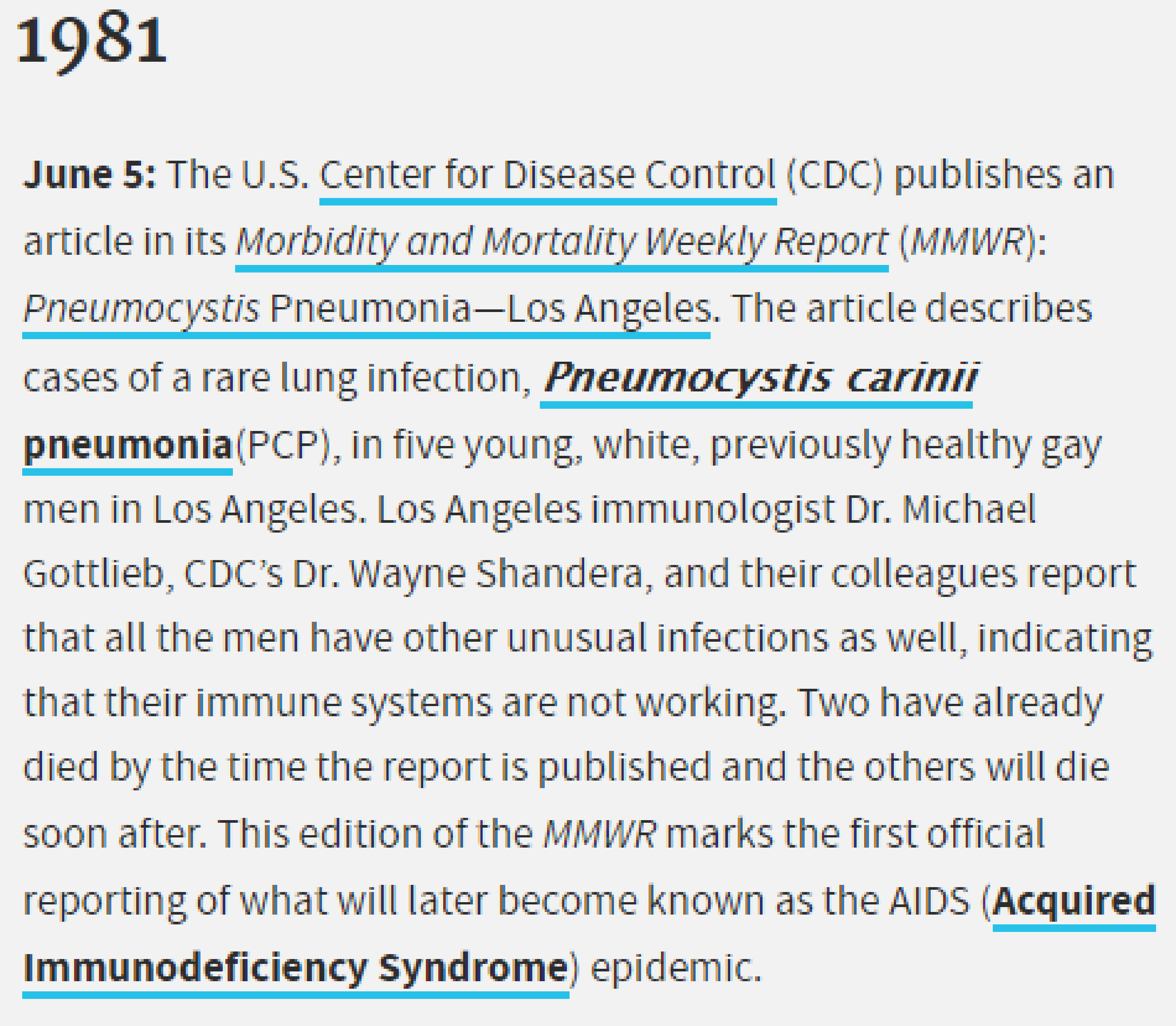
Learning Objectives

- Why use these tools (prevalence, incidence, RR, OR)?
  - inform differential diagnoses & counsel patients
  - design public health interventions & direct new diagnostics/therapies
  - understand distributions and determinants of diseases
- How to use these tools?
  - precise definitions (e.g., RR vs OR, cumulative incidence vs incidence density)
  - a bit of arithmetic

A case from 1981...

- 36-year-old man presents with a 4-month history of fever, dyspnea, and cough.

STARRING
Tom Selleck
AS MAGNUM



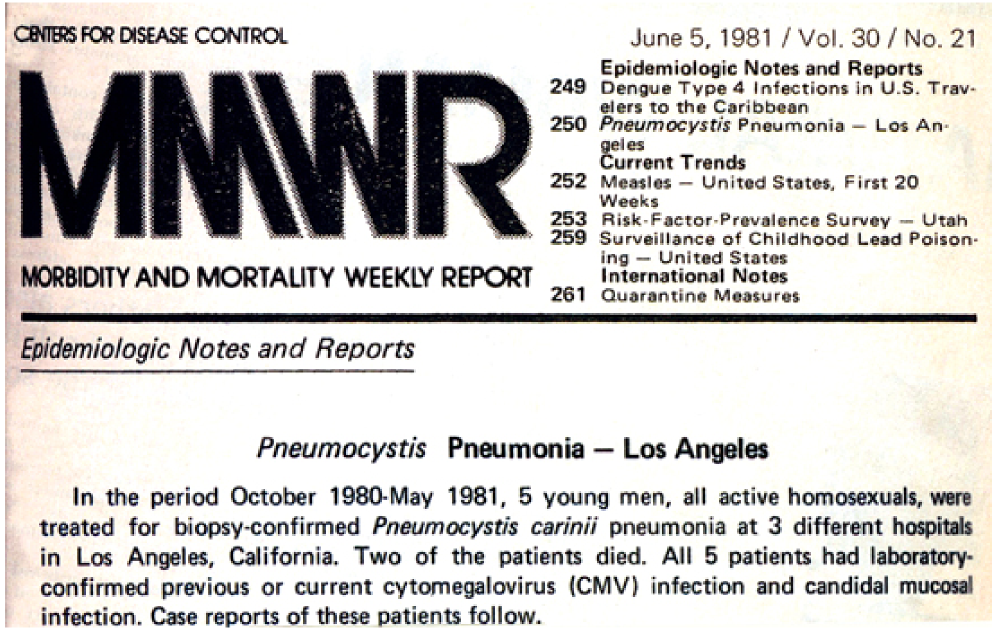
1981

June 5: The U.S. Center for Disease Control (CDC) publishes an article in its *Morbidity and Mortality Weekly Report* (*MMWR*): *Pneumocystis* Pneumonia—Los Angeles. The article describes cases of a rare lung infection, ***Pneumocystis carinii* pneumonia**(PCP), in five young, white, previously healthy gay men in Los Angeles. Los Angeles immunologist Dr. Michael Gottlieb, CDC's Dr. Wayne Shandera, and their colleagues report that all the men have other unusual infections as well, indicating that their immune systems are not working. Two have already died by the time the report is published and the others will die soon after. This edition of the *MMWR* marks the first official reporting of what will later become known as the AIDS (**Acquired Immunodeficiency Syndrome**) epidemic.

MMWR June 5, 1981

All the above observations suggest the possibility of a cellular-immune dysfunction related to a common exposure that predisposes individuals to opportunistic infections such as pneumocystosis and candidiasis. Although the role of CMV infection in the pathogenesis of pneumocystosis remains unknown, the possibility of P. carinii *infection must be carefully considered in a differential diagnosis for previously healthy homosexual males with dyspnea and pneumonia.*

- Gottlieb MS et al MMWR 1981



CENTERS FOR DISEASE CONTROL

June 5, 1981 / Vol. 30 / No. 21

MMWR

MORBIDITY AND MORTALITY WEEKLY REPORT

Epidemiologic Notes and Reports
249 Dengue Type 4 Infections in U.S. Travelers to the Caribbean
250 *Pneumocystis* Pneumonia — Los Angeles
Current Trends
252 Measles — United States, First 20 Weeks
253 Risk-Factor-Prevalence Survey — Utah
259 Surveillance of Childhood Lead Poisoning — United States
International Notes
261 Quarantine Measures

Epidemiologic Notes and Reports

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.

A case from 1981… *Pneumocystis* pneumonia?

- 36-year-old man presents with a 4-month history of fever, dyspnea, and cough.

- What do you want to know and why?

    - history
    - vital signs
    - physical exam
    - laboratory test values
    - radiology

- What is his diagnosis? Does he have *Pneumocystis* pneumonia?

A case from 1981... *Pneumocystis* pneumonia?

- Differential diagnosis must be grounded in understanding:
  - distributions of disease: we'll learn about **prevalence** & **incidence**
  - determinants of disease: we'll learn **2x2 tables** to relate exposures and outcomes
- In 1981, *Pneumocystis* was known to be a low prevalence disease.
- New data would show an increasing incidence.

Note: in infectious diseases, the differential is always evolving

A case from 1981... *Pneumocystis* pneumonia?

- We want to know how the distribution of *Pneumocystis* pneumonia is changing.
- We want to know whether our patient has *Pneumocystis* pneumonia.
- This is a dichotomous measure (yes/no *Pneumocystis*).
- We have to start by thinking about how continuous data become dichotomous.

Data Types & Data Distributions

Data Types

- Vital signs and physical exam:
    - temperature (degrees) - continuous
    - heart / respiratory rate (beats or breaths / min) - continuous
    - oxygen saturation (%) - continuous
- Laboratory values:
    - white blood cell count (cells / uL) - continuous
    - *if it weren't 1981… CD4 cell count (cells / uL) and HIV viral load (copies / mL) - continuous*
- Radiology:
    - ground glass - dichotomous

Data Types

- **Dichotomous**: history of diabetes, history of breast cancer; survival, pneumonia, MI
- **Continuous**: age, height, weight, blood pressure; probability of treatment result
- **Nominal**: race, ethnicity, state of residence
- **Ordinal**: age category, weight category; patient satisfaction

Note: in medicine, we regard diagnoses and clinical decisions as dichotomous

Characterizing Continuous Data

- "Normally distributed" data can be well characterized by their mean and standard deviation.
  - **mean (μ)**:

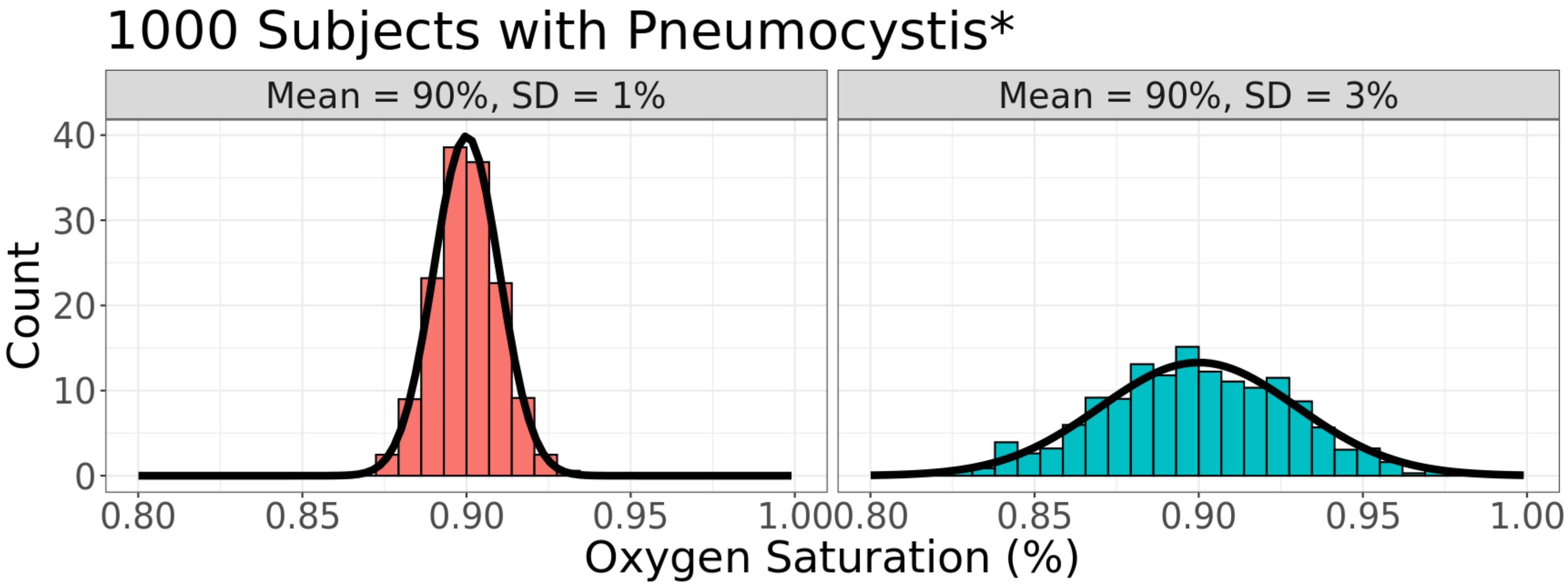
$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

  - **standard deviation (SD or σ)**:

$$\sigma = \sqrt{\frac{1}{N} * \sum_{i=1}^{N} (x_i - \mu)^2}$$

Characterizing Continuous Data

- What makes standard deviation greater?



[*] Data made up.

Reflection Question

What makes standard deviation greater?

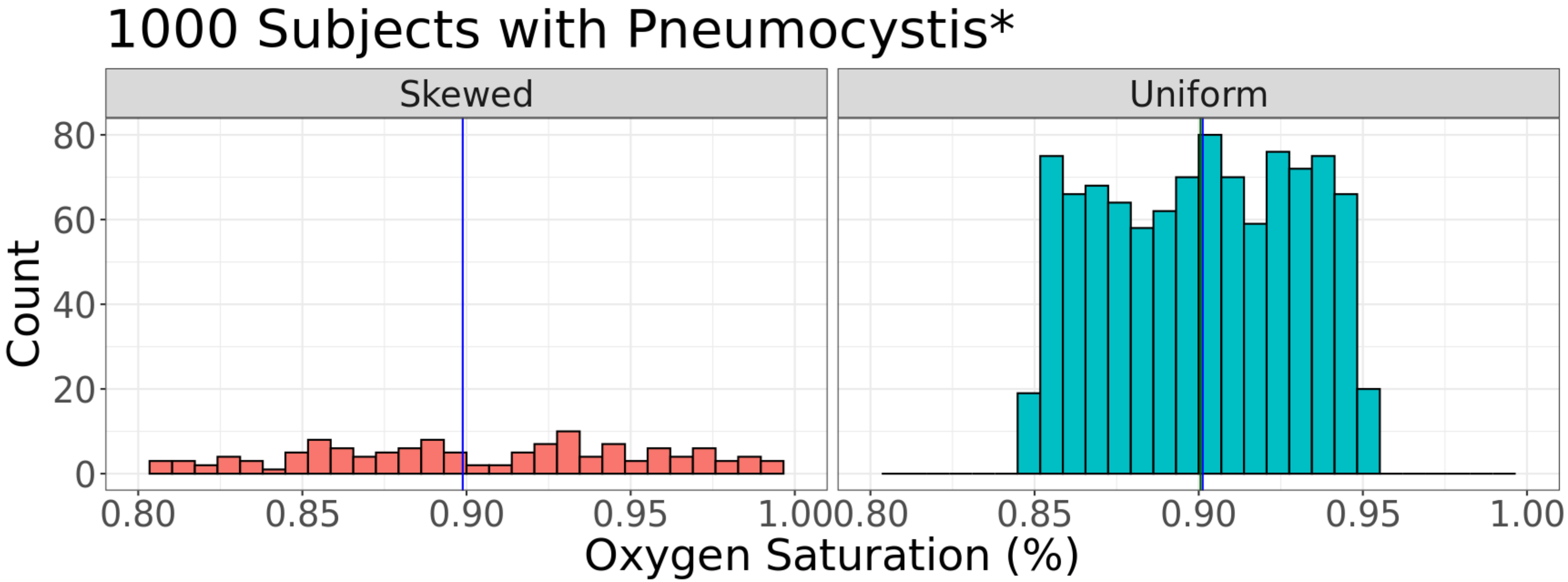
- (A) More subjects?
- (B) Higher mean value?
- (C) Higher maximum value?
- (D) Greater difference between extreme and mean values?

Characterizing Continuous Data

- “Non-parametric” data have a mean and standard deviation, but these parameters do NOT characterize the data well.
- For uniform or skewed data, we prefer to use **median** and **interquartile range (IQR)** to describe the distribution of a continuous variable.

Note: for normal data, mean and median have the same value

Characterizing Continuous Data



[*] Data made up.

Characterizing Continuous Data

- Median and mode depend on ranking the values:
    - median is the “middle value” when data are ordered
    - mode is most frequently occurring value.
- Interquartile range (IQR) also depends on ranking the values:
    - the first quartile is the "middle" value of the first half of the ordered set
    - the third quartile is the "middle" value of the second half of the ordered set
    - IQR is the range of values between first and third quartiles

Reflection Question

Which is least affected by outliers?

- (A) Mean?
- (B) Median?
- (C) Mode?

Characterizing Continuous Data

- Pair a measure of central tendency with a measure of dispersion:
  - mean and SD
  - median and IQR
- In doing so, account for **uncertainty** in measures.
- But if we use a threshold to transform continuous data into dichotomous data, we lose information about uncertainty.
- This is on top of the fundamental uncertainty we face with any epidemiologic measure: does the measured population represent the population of interest?

Dichotomizing Continuous Data

- Imagine diagnosing *Pneumocystis* pneumonia based on threshold of oxygen saturation:

Dichotomania

- As we discuss prevalence, incidence, RR, and OR, we will focus on **dichotomous** exposures and outcomes.
- Remember -- with dichotomous data:
  - information is lost
  - misclassification is common
- But we're doing it anyway 😀 … why?
- Medicine focuses on dichotomous diagnostic and treatment decisions.

Measures of Disease Occurrence

A case from 1981... *Pneumocystis* pneumonia?

- How common is *Pneumocystis* pneumonia?

Prevalence

- How common is *Pneumocystis* pneumonia?
- Prevalence:
    - number with the disease / number in specified population
    - **point prevalence**: at a specific point in time
    - **period prevalence**: during a given period (e.g., 12-month prevalence)
    - a proportion (unitless, ranges from 0-1)
    - numerator includes all people who have the disease, both new and ongoing cases
    - represents a cross-sectional “snapshot” of the population

Prevalence of *Pneumocystis* pneumonia

- In 1967, CDC became the sole supplier of pentamidine in the United States and began collecting data on cases of PCP:
  - period prevalence published in 1974*: 579 cases (194 confirmed) over 3 years.
  - what's the denominator?
  - what's the prevalence?
- **Point prevalence** of *Pneumocystis* would be vanishingly small given limited duration of disease.
- In 1967, even the **period prevalence** is very small.

[*] Walzer PD et al *Annals Int Med* 1974

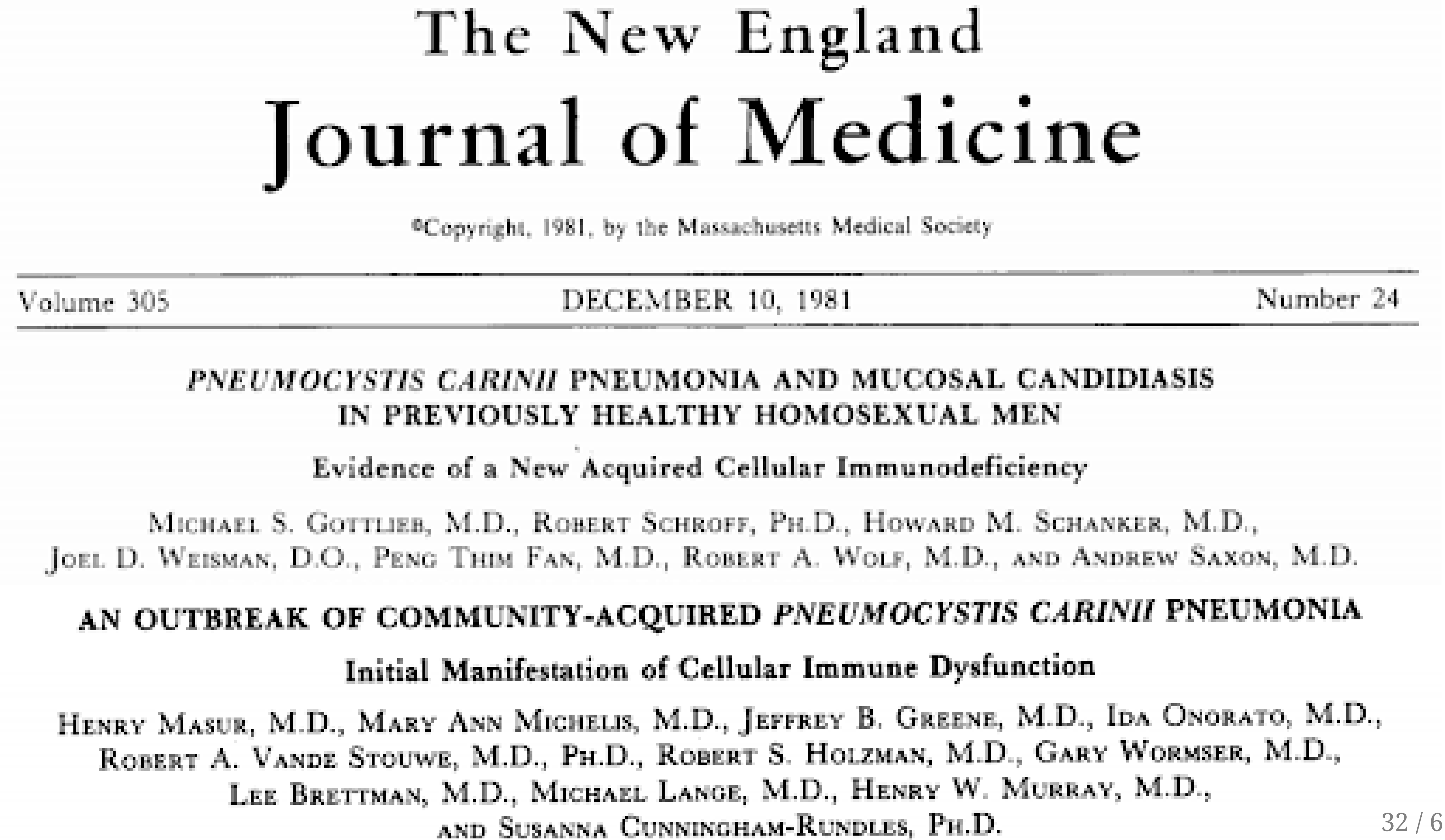
Prevalence

- Prevalence is **NOT** the same as risk.
- Prevalence numerator includes all people who have the disease, both new and ongoing cases, so represents a cross-sectional “snapshot” of the population.
- Prevalence does **NOT** estimate the risk of developing the disease because prevalence does not fully account for time (are the measured cases old cases or new cases?).

Reflection Question

How can an infection have high prevalence if it occurs infrequently?

- (A) the infection is rapidly fatal
- (B) the infection rapidly resolves
- (C) a few children get the infection every year, but the infection persists for the rest of their lives
- (D) the infection results in lifelong protective immunity



The New England

Journal of Medicine

©Copyright, 1981, by the Massachusetts Medical Society

Volume 305 DECEMBER 10, 1981 Number 24

PNEUMOCYSTIS CARINII PNEUMONIA AND MUCOSAL CANDIDIASIS IN PREVIOUSLY HEALTHY HOMOSEXUAL MEN

Evidence of a New Acquired Cellular Immunodeficiency

Michael S. Gottlieb, M.D., Robert Schroff, Ph.D., Howard M. Schanker, M.D., Joel D. Weisman, D.O., Peng Thim Fan, M.D., Robert A. Wolf, M.D., and Andrew Saxon, M.D.

AN OUTBREAK OF COMMUNITY-ACQUIRED *PNEUMOCYSTIS CARINII* PNEUMONIA

Initial Manifestation of Cellular Immune Dysfunction

Henry Masur, M.D., Mary Ann Michelis, M.D., Jeffrey B. Greene, M.D., Ida Onorato, M.D., Robert A. Vande Stouwe, M.D., Ph.D., Robert S. Holzman, M.D., Gary Wormser, M.D., Lee Brettman, M.D., Michael Lange, M.D., Henry W. Murray, M.D., and Susanna Cunningham-Rundles, Ph.D.

Incidence

- Among MSM, *Pneumocystis* pneumonia is occurring more frequently…
- Incidence: occurrence of new cases over a given period of time:
    - **cumulative incidence**: # new cases / # population at risk @ start time interval
    - **incidence density**: # new cases / person-time at risk (more precise)

Notes on Cumulative Incidence

- Cumulative incidence:
  - must specify population consisting of at-risk individuals
  - must specify a time period of observation
  - numerator = all new cases during a specified time period
  - denominator = all individuals at risk in the specified population at the start of the specified time period (does NOT account for deaths due to other causes)
  - ranges from 0 to 1 (a.k.a., “incidence proportion”)
  - like prevalence, is a proportion and therefore has no units (but only makes sense if you specify the time period of observation, e.g., % per year)

Notes on Incidence Density

- Incidence density:
  - in a specified population consisting of at risk individuals over a specified period of observation, more precisely quantifies the person-time at risk
  - numerator = all new cases during a specified time period
  - denominator = the sum, over all individuals in the population, of time at risk until the event of interest, death, loss to follow-up, the end of the study, or when they are no longer at risk for whatever reason
  - not a proportion; range depends on the units of person-time (0 to infinity)
  - **accounts for death from other causes!**

Notes on Population at Risk

- In a population, individuals are at risk of disease if they:

  (1) do not have the disease at baseline (2) are capable of developing the disease□(e.g., have the organ of interest; □have not been successfully immunized against the disease;□haven’t developed lifelong immunity)

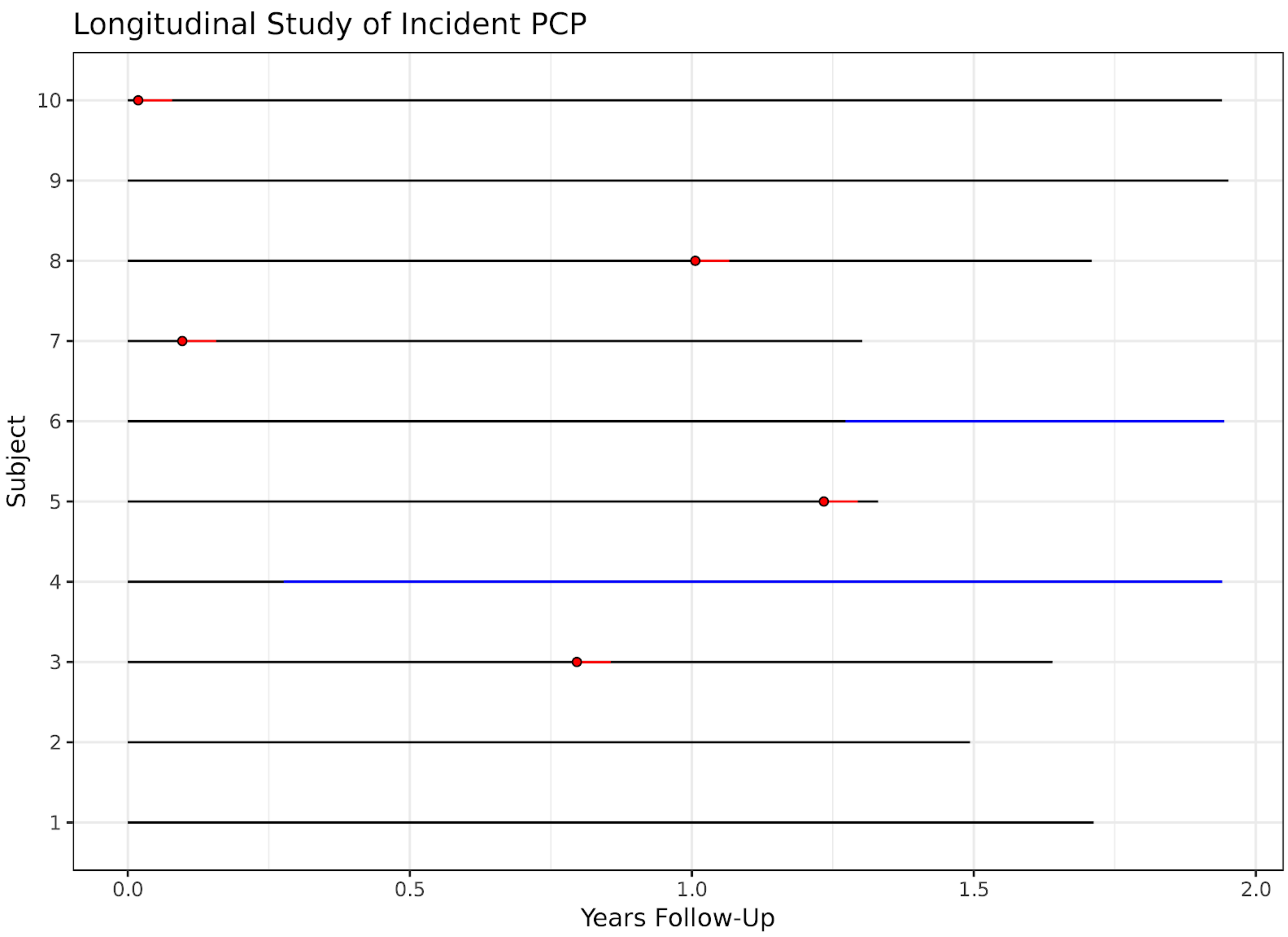
- The difference between cumulative incidence and incidence density is that the latter attempts a more precise quantification of population at risk -- it’s harder to evaluate, but more informative if you can.

Caution with “Incidence Rate”

- “Incidence rate” is used to mean two different things:
    - number new cases / number population at risk @ start (short) time interval (e.g., “annual incidence rate” to mean cumulative incidence over one year)
    - number new cases / person-time at risk (i.e., incidence density, the precise rate)

Incidence: Which Denominator?

- To understand the difference between cumulative incidence and incidence density, imagine a study of persons with HIV at risk for PCP: 10 subjects enrolled at the start of a two-year observation period
    - 5 cases of PCP (red on plot); each receives 3 weeks of antibiotic treatment
    - 2 subjects started on PCP prophylaxis during follow-up (blue on plot)

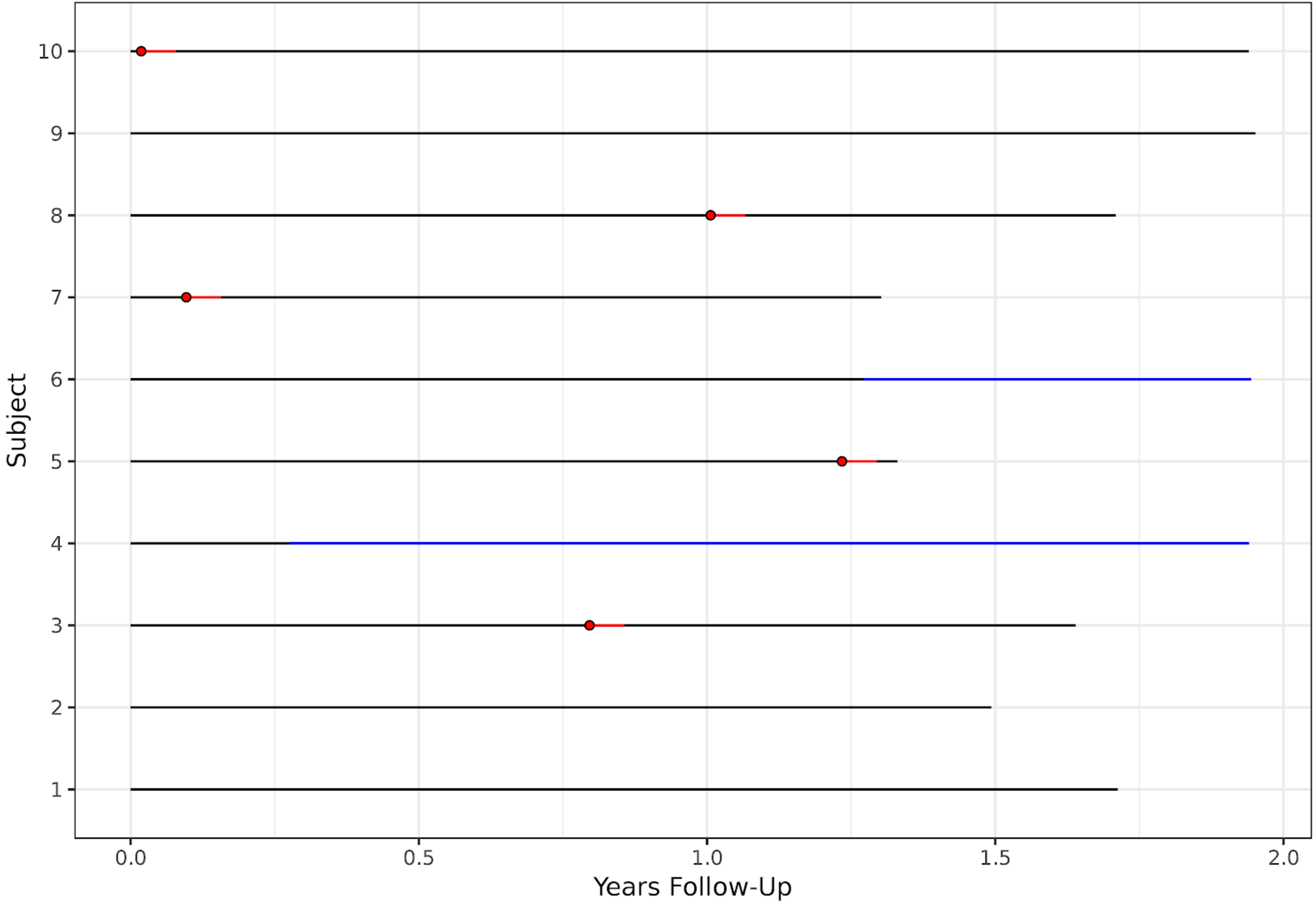

Longitudinal Study of Incident PCP
10
9
8
7
6
5
4
3
2
1
Subject
0.0
0.5
1.0
1.5
2.0
Years Follow-Up

(Data Made Up)

Incidence: Which Denominator?

- To understand the difference between cumulative incidence and incidence density, imagine a study of persons with HIV at risk for PCP: 10 subjects enrolled at the start of a two-year observation period
    - 5 cases of PCP (red on plot); each receives 3 weeks of antibiotic treatment
    - 2 subjects started on PCP prophylaxis during follow-up (blue on plot)
- What’s the annual cumulative incidence of PCP?
    - if you don’t count time on prophylaxis or treatment antibiotics as “time at risk”, how does the incidence density compare to the annual cumulative incidence?
    - what if the end of the black line is death / loss to follow-up?

Longitudinal Study of Incident PCP



(Data Made Up)

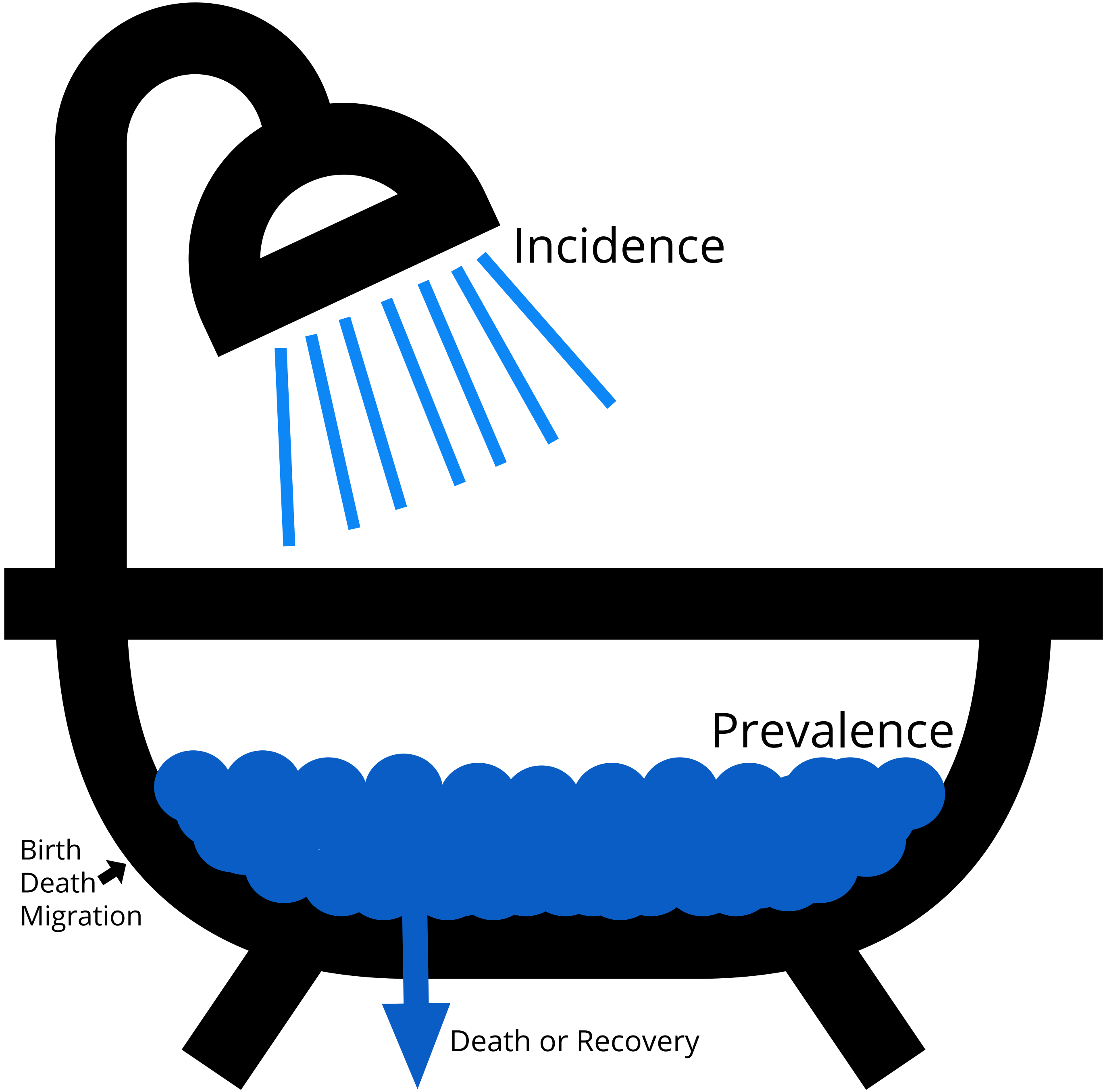
Reflection Question

Your patient with HIV is considering starting prophylactic antibiotics for PCP. You have PCP prevalence, cumulative incidence, and incidence density data available. Which data provide the most precise information on the patient’s risk of PCP off of prophylaxis?

- (A) prevalence
- (B) cumulative incidence
- (C) incidence density

Prevalence vs Incidence

- Can you tell prevalence from incidence?


Incidence
Prevalence
Birth
Death
Migration
Death or Recovery

Prevalence vs Incidence

- Can you tell prevalence from incidence?
- pancreatic cancer versus leukemia:
  - new cases per year: pancreatic 24,120; leukemia 23,370
  - deaths per year: pancreatic 19,850; leukemia 10,240
  - which is more prevalent?
- HIV in Rakai, Uganda 1994-2003*:
  - intensive “ABC” intervention (Abstinence, Be faithful, Condoms)
  - prevalence declined, but incidence remained constant at 1.5% per year
  - what happened?

[*] Wawer M et al CROI 2005; Roehr B BMJ 2005

We're not done yet...

#BROADCITY

… and coming now are the formulas you really need to know…

#BROADCITY

Inference from Epidemiologic Measures

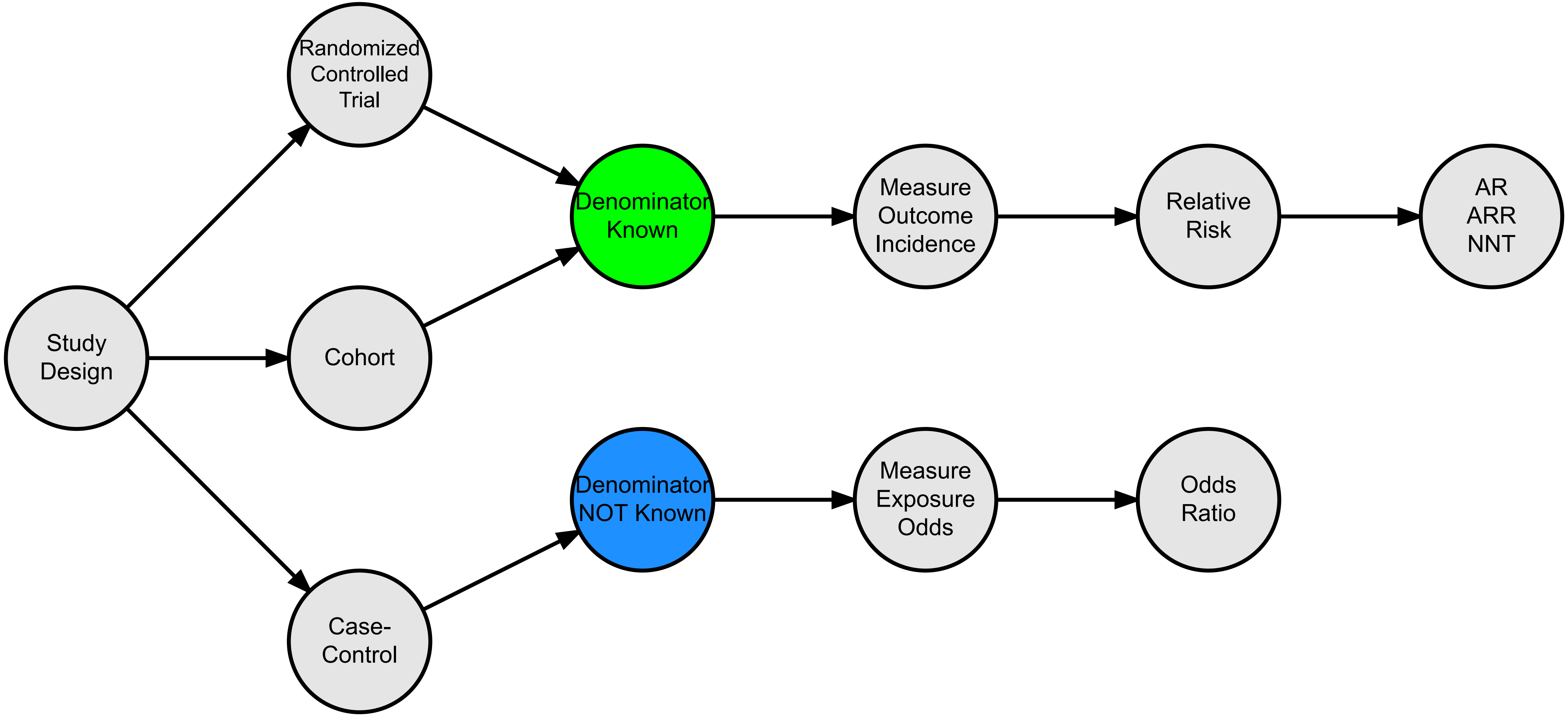
Basics of Study Types

- We want to understand the relationship between risk factors (exposures) and disease (outcomes). For example, between CD4 count and PCP in HIV.

- To calculate incidence need to know how many are in a population:

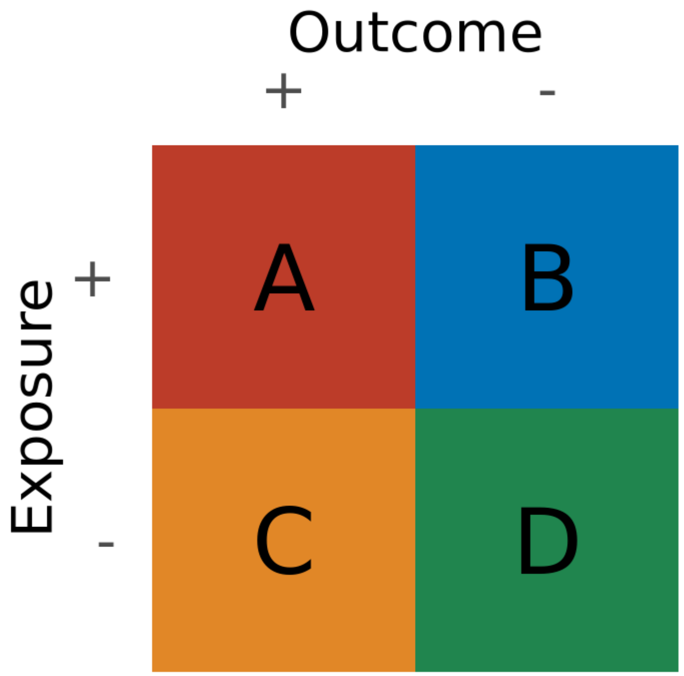
  - randomized trials: pick the population, randomize, control the treatment, and measure the outcome
  - cohort studies: pick the population, divide into preselected exposure (treatment or risk factor) groups, and measure the outcome

- But do NOT know this generally in case control studies: pick the cases and control groups, then measure rates of exposure (do NOT know how many in population).


Randomized Controlled Trial
Study Design
Cohort
Case-Control
Denominator Known
Denominator NOT Known
Measure Outcome Incidence
Measure Exposure Odds
Relative Risk
Odds Ratio
AR ARR NNT

2x2 Table

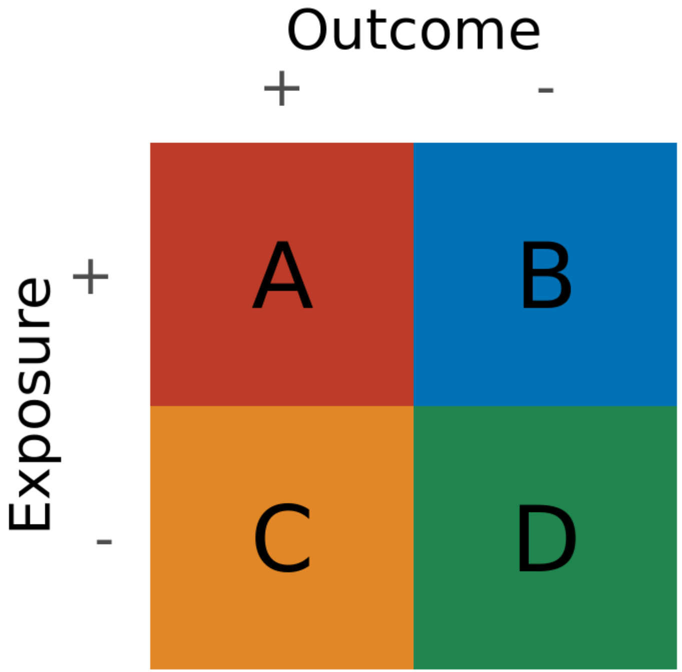
- Dichotomous exposures and outcomes.
- Examine relationships between exposures and outcomes. Goal: inference about larger world.
- Caution!!!
  - need to know whether study is RCT/cohort vs case-control
  - can always calculate a relative risk (RR) from 2x2 table but only appropriate for RCT/cohort
  - can always calculate an odds ratio (OR) from 2x2 table but only appropriate for case-control study
    (can do better with RR if RCT/cohort)



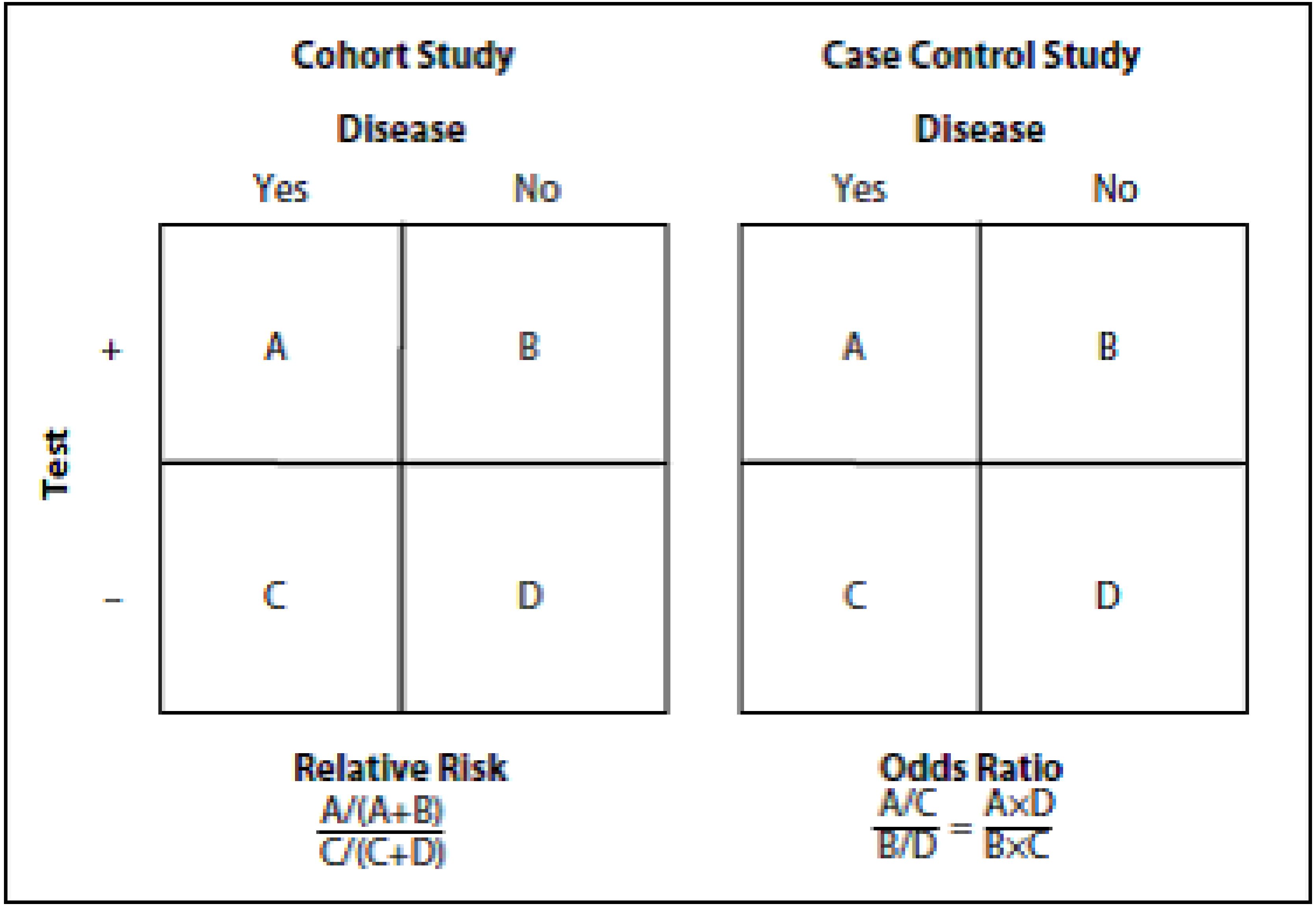
		Outcome	
		+	-
Exposure	+	A	B
	-	C	D

2x2 Table: Calculation for Cohort/Trial

- Relative Risk (RR)= [A/(A+B) / C/(C+D)]
- Absolute Risk Reduction (ARR) = [A/(A+B) – C/(C+D)]
- Number Needed to Treat (NNT)= 1/ARR



		Outcome	
		+	-
Exposure	+	A	B
	-	C	D



Cohort Study

Test	Disease: Yes	Disease: No
+	A	B
–	C	D

Relative Risk

$$\frac{A/(A+B)}{C/(C+D)}$$

Case Control Study

Test	Disease: Yes	Disease: No
+	A	B
–	C	D

Odds Ratio

$$\frac{A/C}{B/D} = \frac{A \times D}{B \times C}$$

Figure 1: In a case-control study, the odds ratio can be used to approximate the relative risk under the assumption that the disease is rare.

Relative Risk (RR)

- Compares risk between two groups of people:
  - if 2 in 10 are cured in the control group and 3 in 10 in the treatment group, the RR is (2/10)/(3/10) = 0.66 □(i.e. 0.33 less likely to have the disease after treatment)
  - can also be calculated as the inverse: if 3 in 10 are cured in the treatment group and 2 in 10 in the control group, the RR is (3/10)/(2/10) = 1.5 □(i.e. 0.5 times more likely to be cured if treated)

Absolute Risk (AR) & Reduction (ARR)

- Absolute risk (AR): risk of developing a given disease over a period of time
    - this is the incidence!!!
    - if you have a 1 in 10 chance of developing skin cancer in your lifetime, you are said to have a 10% absolute risk
- Absolute risk reduction (ARR): difference in risk between the treatment/ exposure group and the control group
    - if 2 in 10 are cured in the control group and 3 in 10 in the treatment group, the ARR is 3/10 – 2/10 = 10%

Number Needed to Treat (NNT)

- Number needed to treat (NNT): Number of patients who need to be treated for one person to benefit from the treatment (= 1/ARR)
    - using ARR numbers above, NNT = 1/ARR = 1/0.1 = 10
    - in this example, you need to treat 10 people to prevent one bad outcome
- Return to example of CD4 count and PCP… imagine a pill that can maintain CD4 count above 250… what's the NNT to prevent one case of PCP?
    - RR = (26 / 50) / (2 / 50) = 13
    - ARR = (26 / 50) - (2 / 50) = 0.48
    - NNT = 1 / 0.48 = 2.08

2x2 Table: Calculation for Case-Control

- Odds Ratio (OR)= (A/C)/(B/D) = AD / BC
- Absolute Risk Reduction (ARR) = not appropriate to calculate
- Number Needed to Treat (NNT) = not appropriate to calculate
- **############## add table**

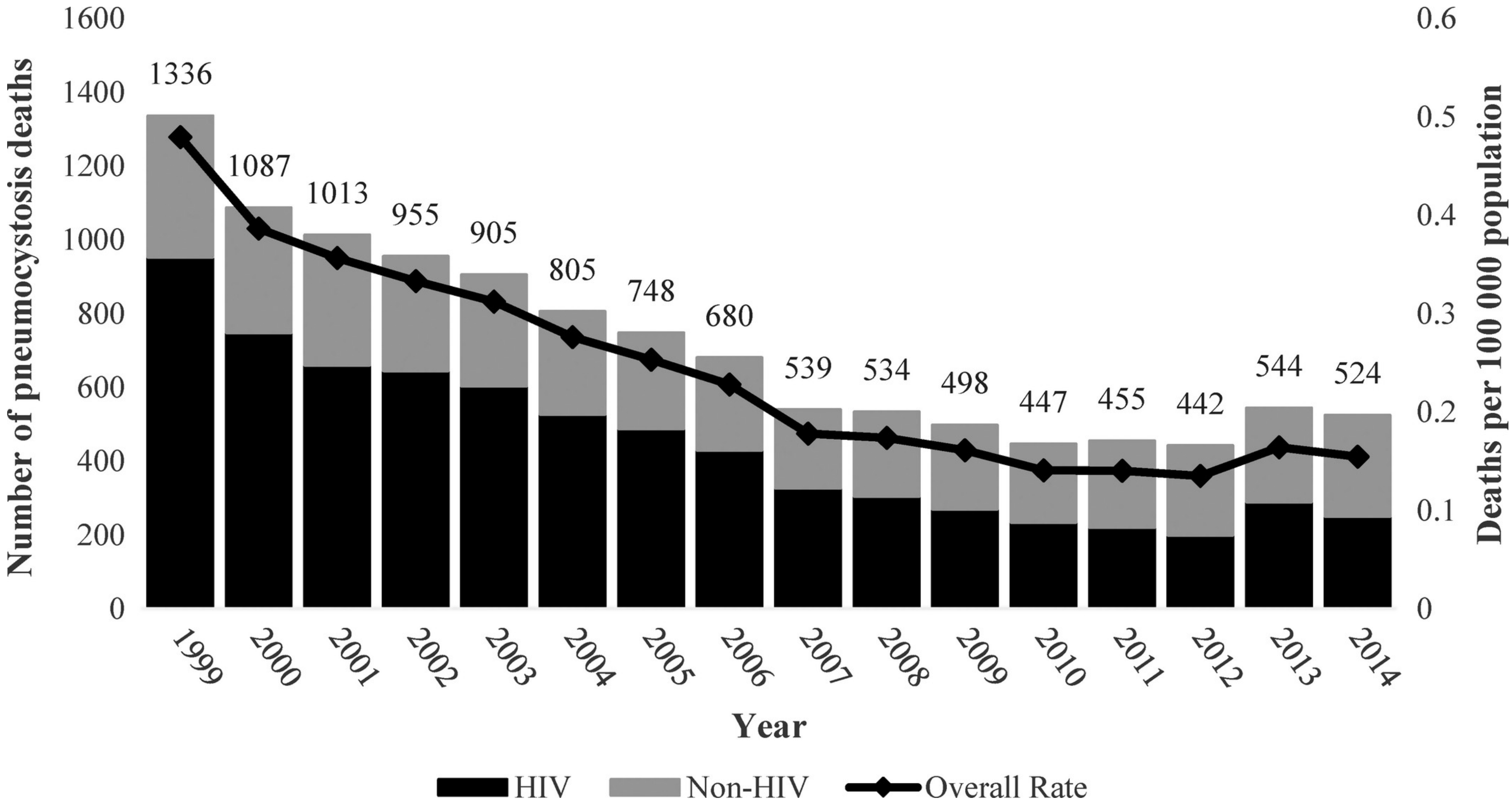
Utility of Odds Ratio (OR) & Case-Control

- if the disease incidence is low, then:
- $A + B \sim B$ & $C + D \sim D$
- $RR = (A / A + B) / (C / C + D) \sim (A / B) / (C / D) = AD / BC = OR$
- OR will be close to RR if outcome occurs infrequently (<15%).
- If outcome is more common, OR will differ increasingly from RR:

Conclusions

Measures of Disease in Clinical Epidemiology

- Prevalence is determined by incidence and survival time.
- Distribution of data determines how we describe them: mean and SD vs median and IQR.
- Relative risk (RR) and odds ratio (OR) are measures of a difference between the incidence of the outcome for two or more exposures or treatments.
- RR and OR approximate each other when outcome is rare.
- NNT can be a clinically useful number.



Wickramasekaran et al *Mycoses* 2017