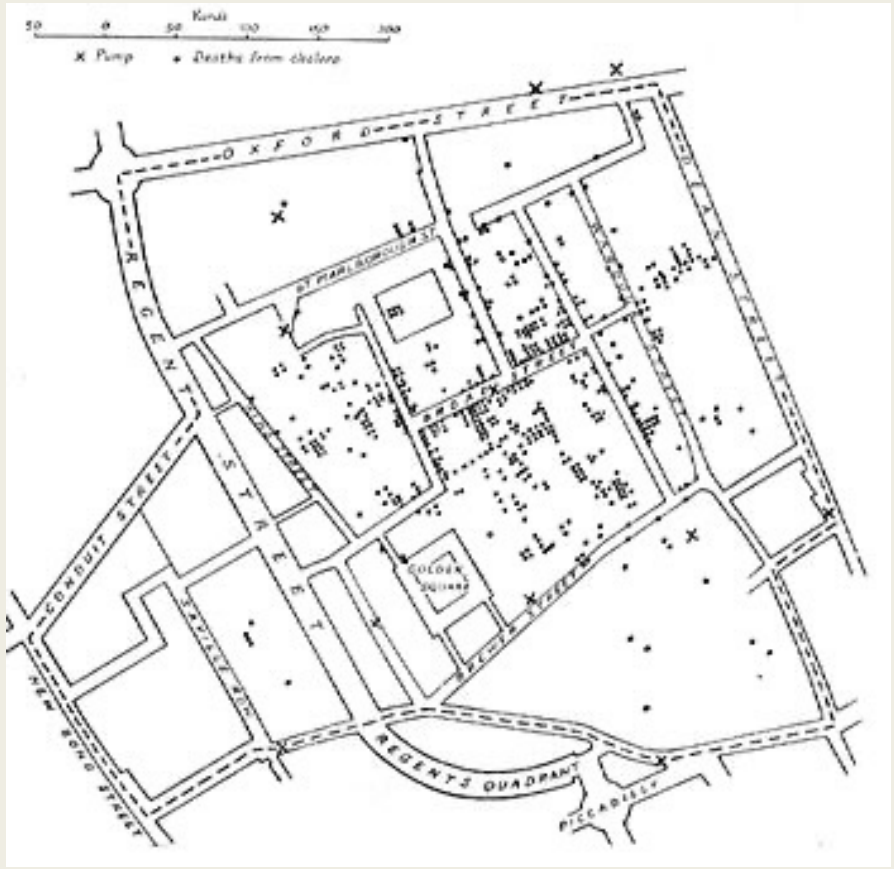

Yards
50 0 50 100 150 200
× Pump • Deaths from cholera
OXFORD STREET
REGENT STREET
CONDUIT STREET
BROAD STREET
GOLDEN SQUARE
REGENT'S QUADRANT
PICCADILLY
NEW BOND STREET

EPI 101

Intro

- Palwasha Khan

- Stephen Olivier

Epidemiology *vs* biostatistics *vs* data science

- Epidemiologists receive substantial training in the science of study design, measurement, and the art of causal inference *including statistical methodology*

- Biostatisticians are well versed in the theory and application of statistical methodological techniques *including study design and causal inference*

- Data scientists receive equivalently rigorous training in computational and visualization approaches for high-dimensional data *including statistical methodology*

Collaboration and cross-training provides opportunity to share and learn:

- Constructs
- Frameworks
- Theories
- Methods

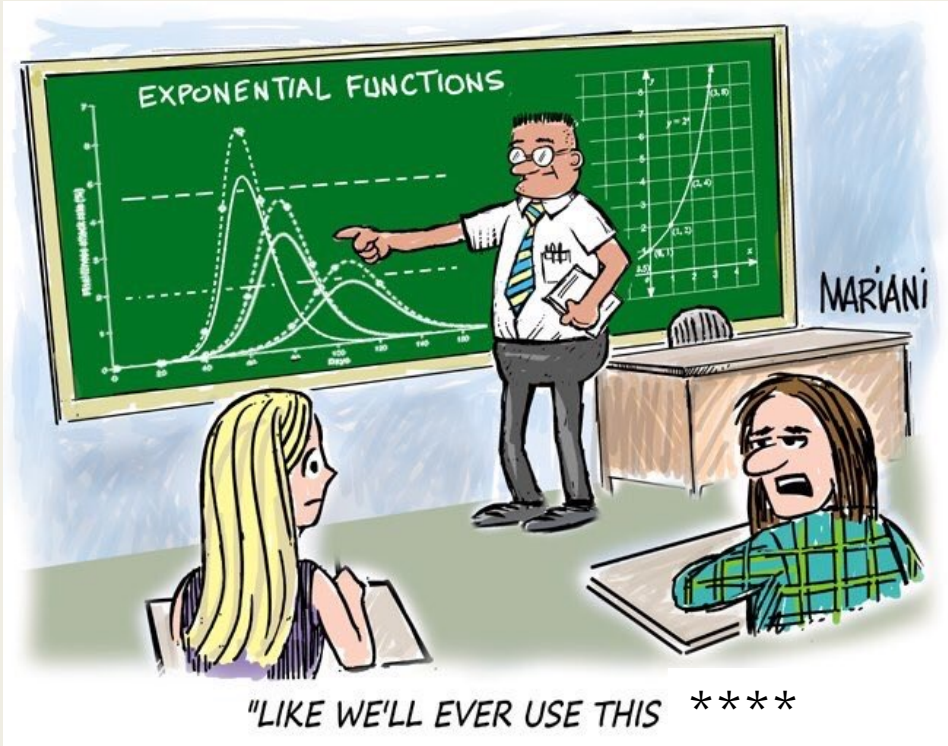
→ fresh and innovative perspectives for tackling challenging problems in health and health care

Goldstein et al. On the Convergence of Epidemiology, Biostatistics, and Data Science. Harv Data Sci Rev. 2020 PMCID: PMC8734556

TEACHING EPI IN 2 DAYS

Intended learning outcomes

- Introduce you to epidemiological concepts
- Provide you with a scaffold for further learning
- Impart an appreciation of the discipline and the important role it plays in health data science



What is epidemiology?

- Morris 1957
  - *Epidemiology defined as the study of health and disease of population and groups in relation to their environment and ways of living*

- Morris 1957
  - *Epidemiology defined as the study of health and disease of population and groups* ***in relation to their environment and ways of living***

- Morris 1957
  - *Epidemiology defined as the study of health and disease of population and groups* ***in relation to their environment and ways of living***

- Lilienfeld & Lilienfeld 1980
  - *Epidemiology is concerned with the patterns of disease occurrence in human populations and of the factors that influence these patterns*

- Kelsey, Thompson & Evans 1986
  - *Epidemiology is the study of the occurrence and distribution of disease and other health-related conditions in populations*

- Oakes & Kaufman 2006
  - *Epidemiology is the study of distribution and determinants of states of health in populations*

- Morris 1957
  - *Epidemiology defined as the study of health and disease of population and groups* ***in relation to their environment and ways of living***
- Lilienfeld & Lilienfeld 1980
  - *Epidemiology is concerned with the patterns of disease occurrence in human populations and of the factors that influence these patterns*
- Kelsey, Thompson & Evans 1986
  - *Epidemiology is the study of the occurrence and distribution of disease and other health-related conditions in populations*
- Oakes & Kaufman 2006
  - *Epidemiology is the study of distribution and determinants of states of health in populations*

A HISTORICAL, POPULATION & CONTEXTUAL SCIENCE

Complements clinical and laboratory sciences
(Nancy Krieger 2007)

Morris's Seven "Uses of epidemiology" (1957)

1. In ***historical study*** of community health, rise and fall of disease in population (useful projections)
2. For ***community diagnosis*** of the presence, nature and distribution of health and disease among the population (changing prevalence, incidence)
3. To study the ***workings of the health services***
4. To estimate the ***individual's chances and risks*** of disease
5. To help ***complete the clinical picture***
6. In ***identifying syndromes***
7. In the ***search for causes*** of health and disease

A METHOD FOR THINKING

What is your research question?

Main roles of epidemiological research

1. Descriptive role (surveillance and observation)
   - a **measure** of the changing burden of disease within and between populations
   - explain local disease patterns
   - *measure of outcome occurrence*

2. Explanatory and analytic role (hypothesis testing and experiments)
   - Explains the **cause**(s) of disease states (even when the biology is not fully understood)- "aetiological epidemiology"
   - Assesses the effectiveness of interventions and modes of healthcare delivery
   - *measure of exposure effect*

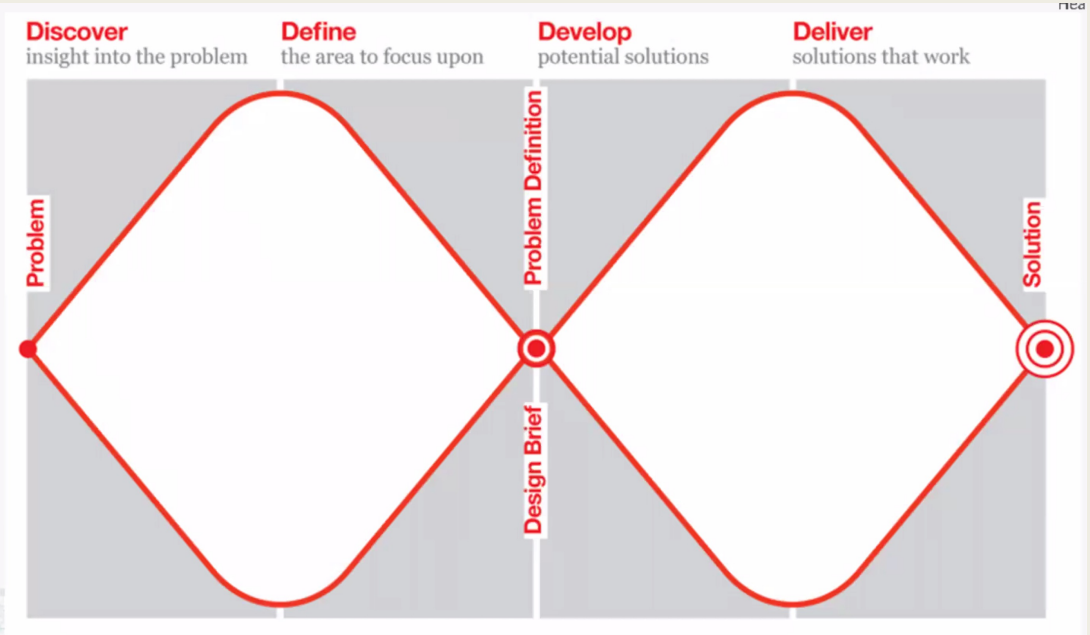
FUNDAMENTAL TO PUBLIC HEALTH

EXPOSURE  OUTCOME

What is your research question?

PROCESS

1. Discover
2. Define
3. Develop
4. Deliver



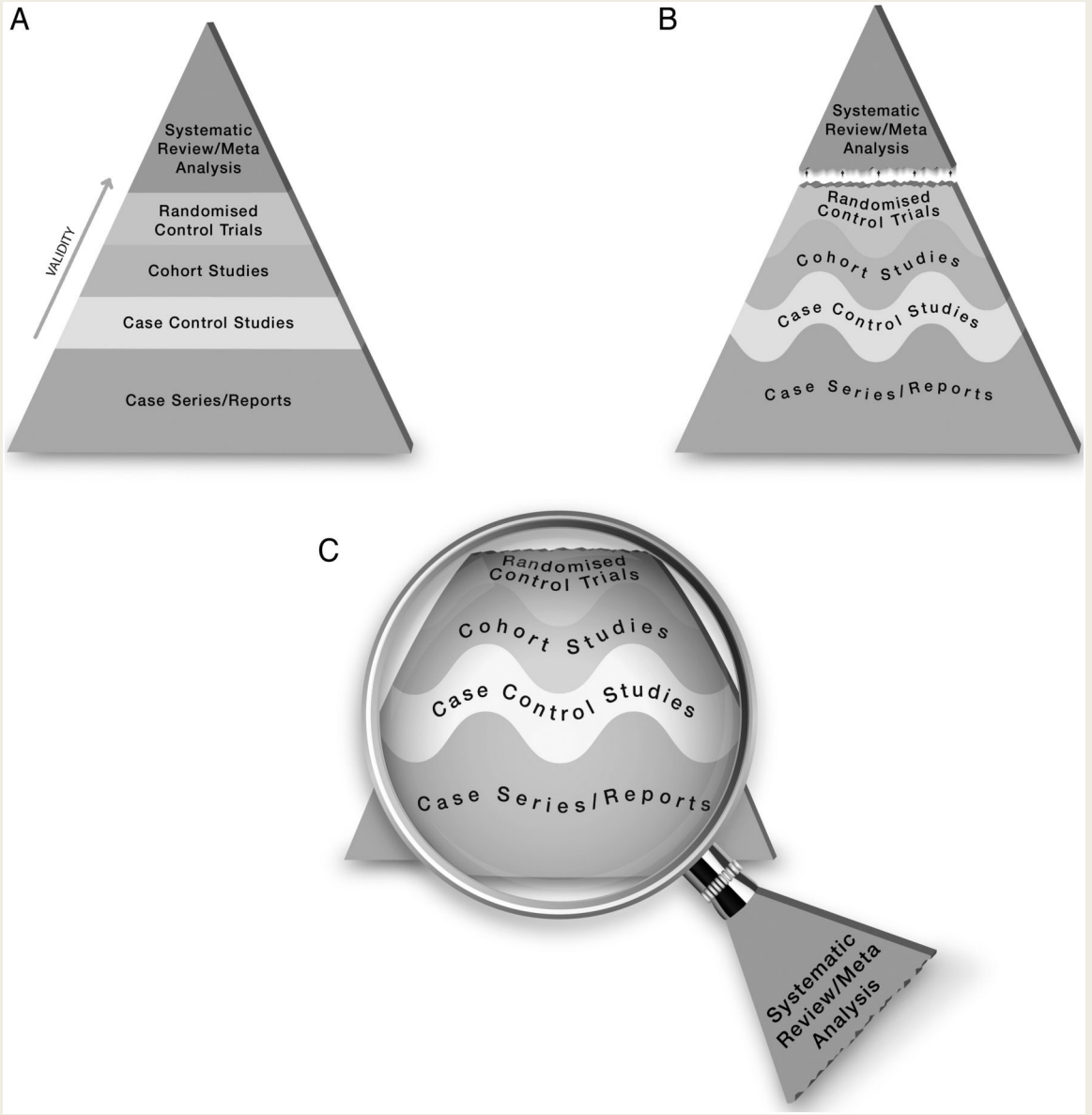
Start with your area of interest, the question, the problem:

1. **Discover** insights so your scope widens but then you have to focus down on a very defined question
2. **Define** the research question very clearly and specifically
3. **Develop** the design brief and again scope widens as you think through different study designs, solutions and finally
4. **Deliver** a study plan on how to address your research question

STUDY DESIGN

"Hierarchy" of study designs

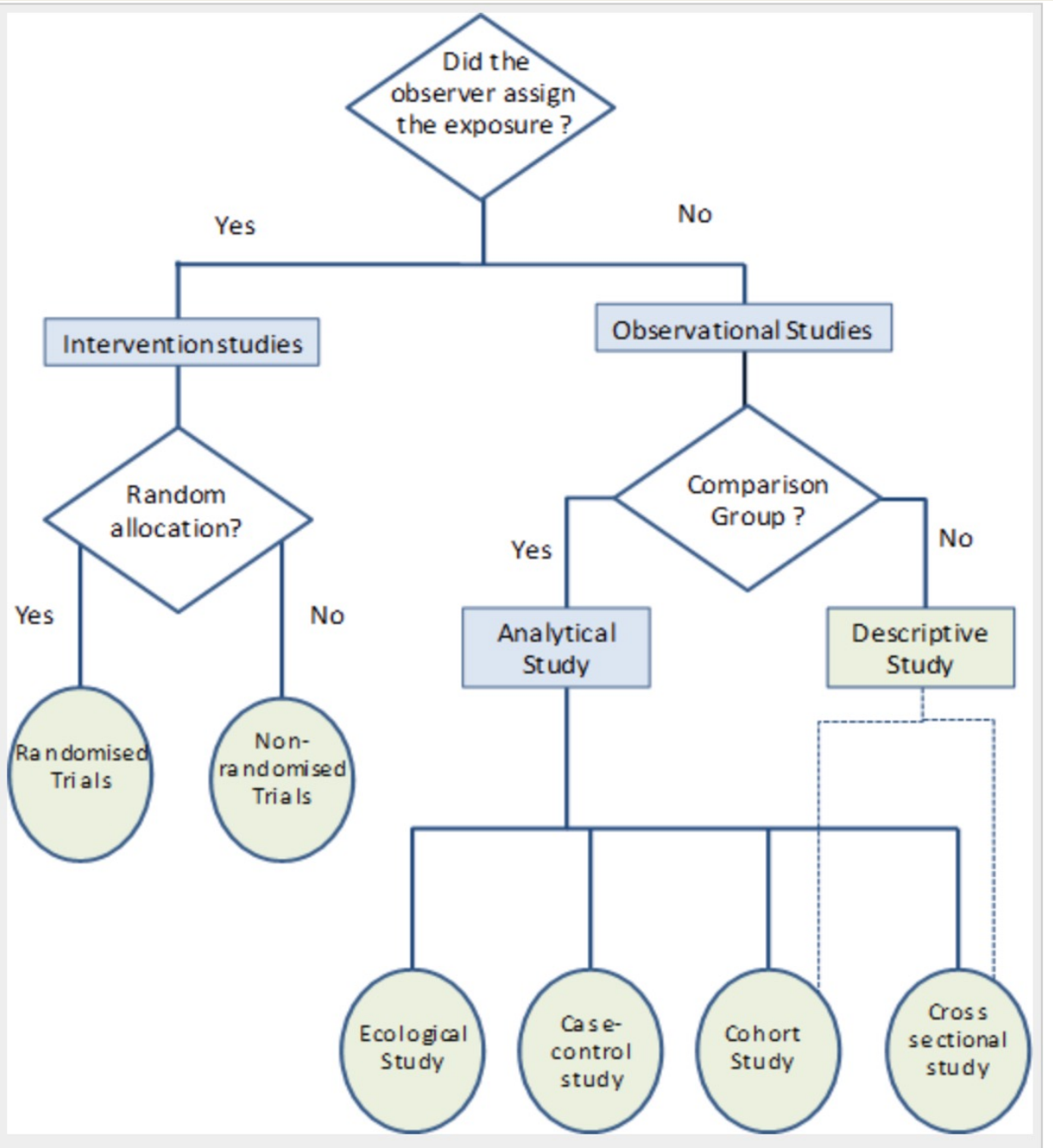
- Different study designs are best for different questions
- **Target validity** = Internal validity + external validity
  - RCT conducted in a white MSM under-30 vs observational study conducted in a sample of population where all groups are represented
- What kind of evidence are you aiming to produce?
  - If you are producing evidence more relevant to public health policy change then need to consider internal validity AND external validity



Murad *et al*. New evidence pyramid. *BMJ Evidence-Based Medicine* 2016;**21:**125-127

Overview

- Epidemiological studies broadly classified as either:
  - *OBSERVATIONAL*
    - Descriptive
    - Analytical
  - *INTERVENTION*



Individual-level data

Cross-sectional studies

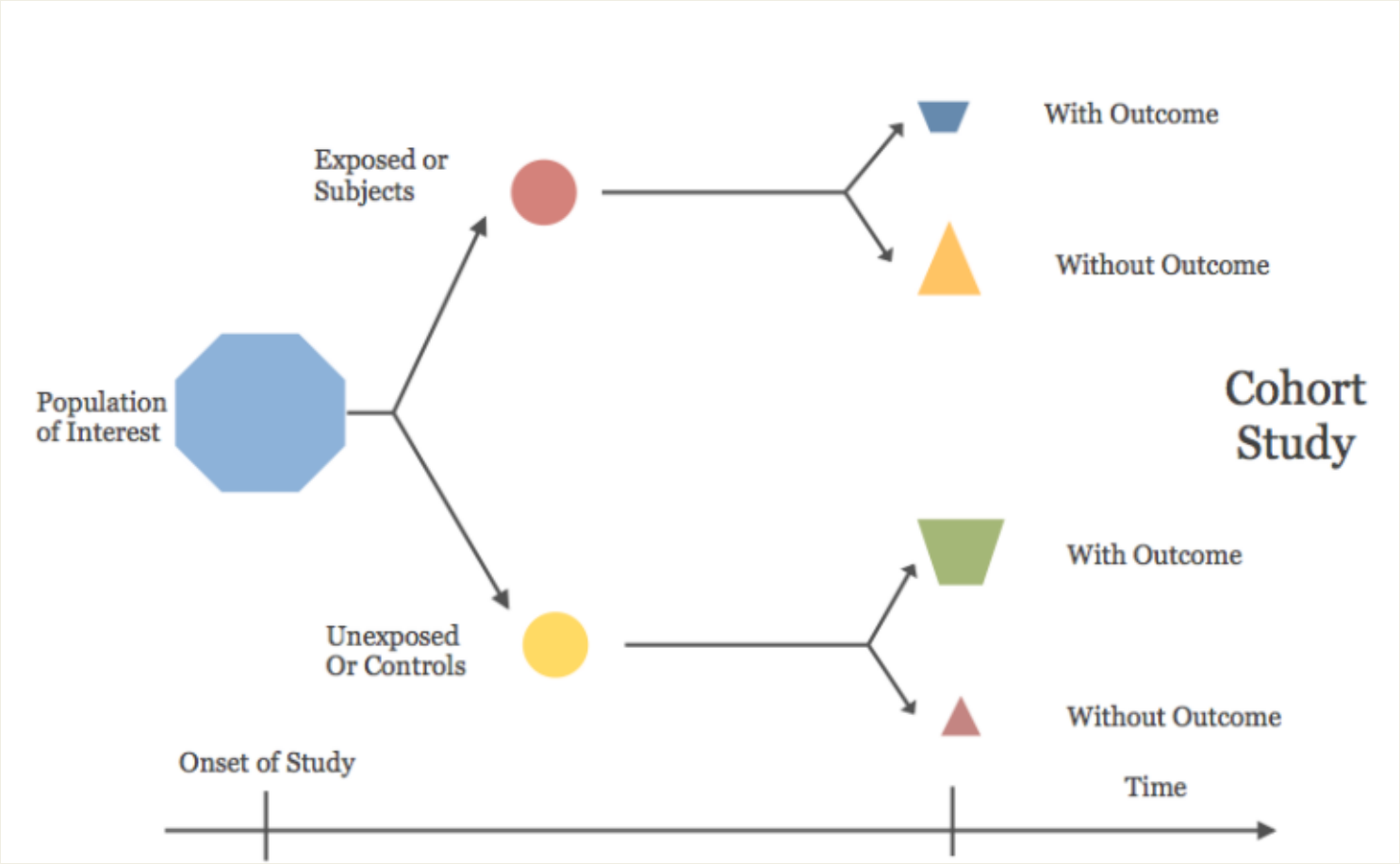
- *Descriptive*: measure frequency of a particular exposure(s) and/or outcome(s) in a defined population at a **particular point in time**

- *Analytical*: collect information on both outcome and exposure at same point in time →make a comparison
  - *frequency of outcome in the people exposed with the frequency in those unexposed*
  - *Direction of **causality** is not always clear*

Examples include pulmonary TB prevalence survey to estimate burden of infectious TB

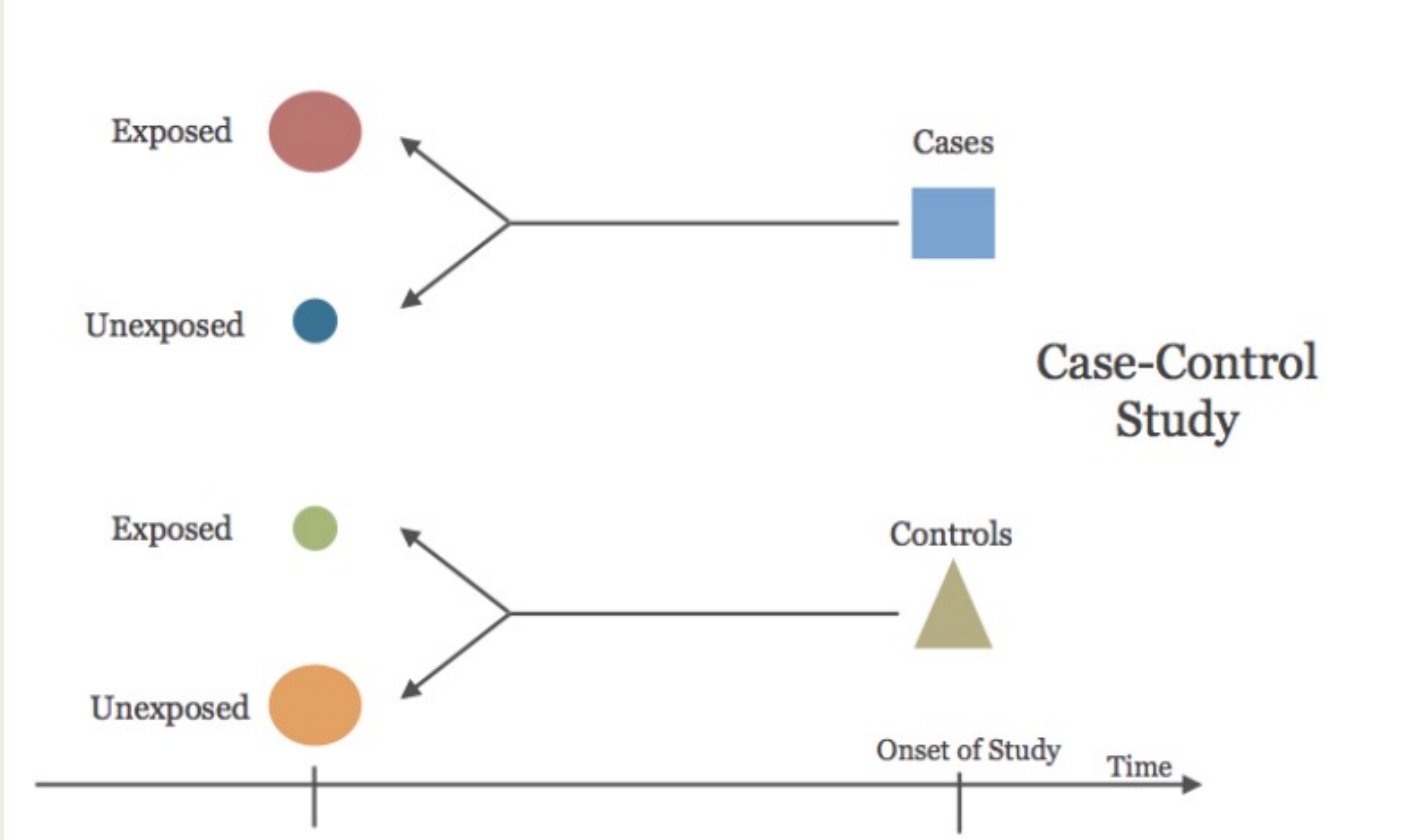
Cohort studies

- Classify members according to **exposure status** (exposed or not)
- Follow up the whole cohort over time
- Compare incidence of the **outcome** in the exposed and in the unexposed
- *Adv*: Calculate incidence, clearly show time of exposure and development of outcome
- *Disadv*: resource intensive, loss to follow-up



Case control studies

- Start by identifying individual cases of the **outcome** of interest
- Identify a representative group of individuals who do not have the outcome = controls
- Compare cases and controls to assess whether there were any difference in past **exposure** to one or more risk factors
- ADV: quick to execute, good for rare outcomes
- DISADV: bias (selection, recall), design may not allow calculation of incidence

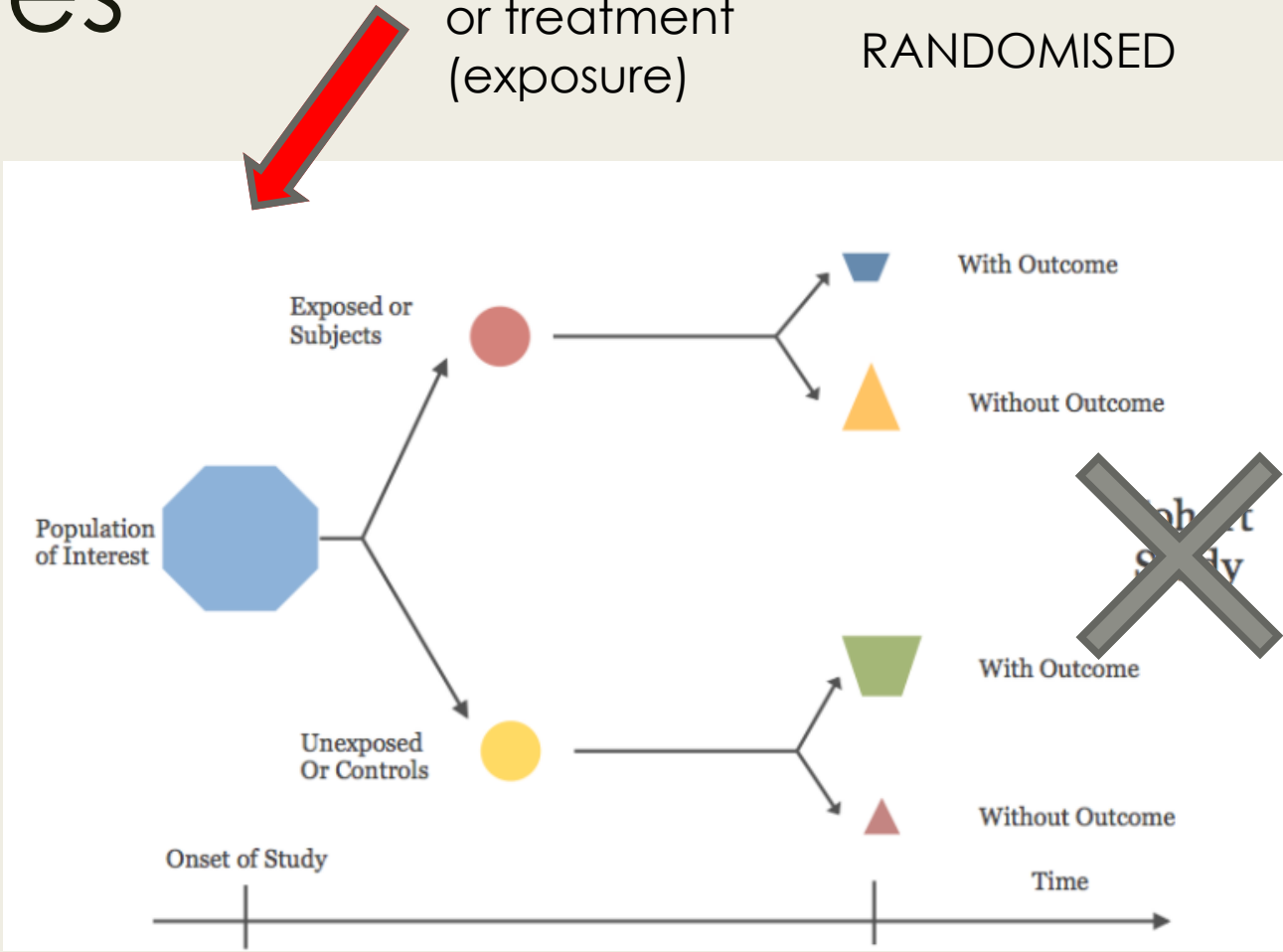


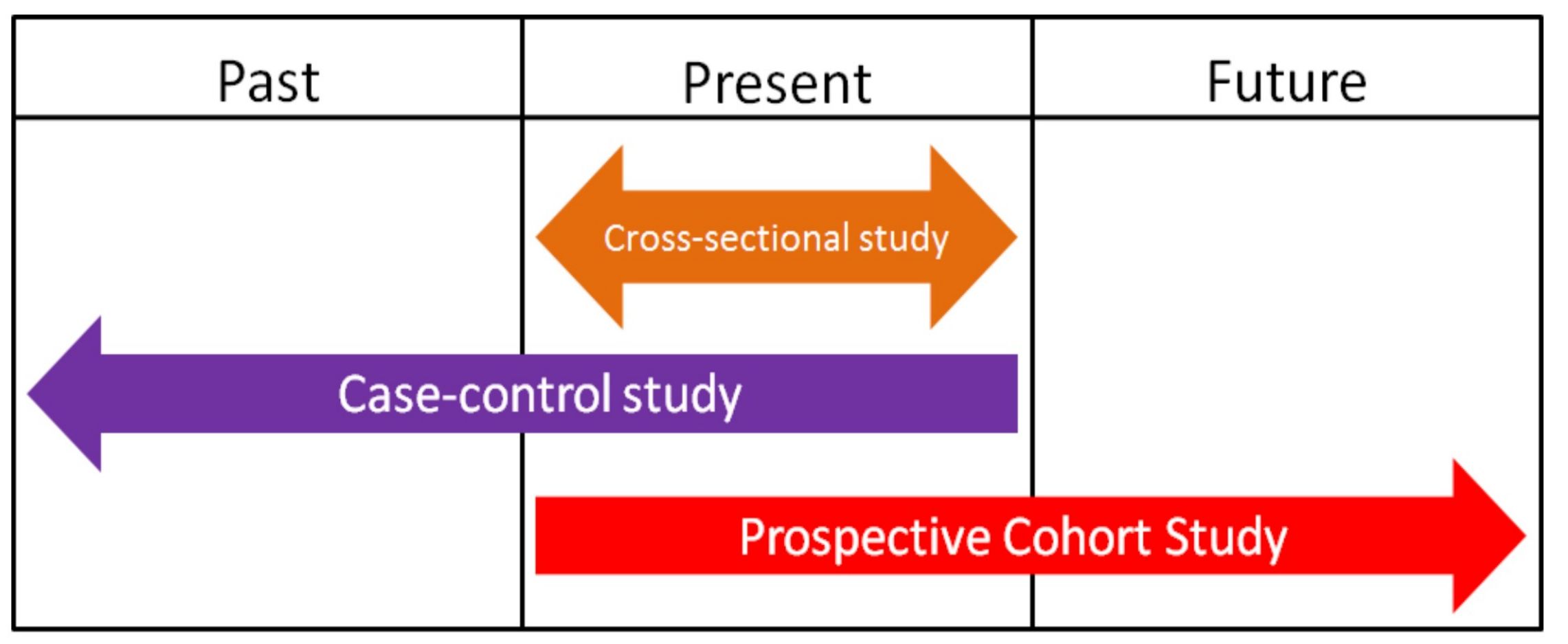
Am I comparing apples with apples?

Or am I really comparing apples with oranges?

Intervention studies

- **Allocate** the exposure or intervention to one of the study groups
- The other group acts as a control group
- Followed-up over period of time
- Compare the frequency of outcome in the two groups
- Random allocation = RCT (randomized controlled trial)
- ADV: strong evidence about causation with random allocation
- DISADV: not always ethical to conduct





Beware of terminology

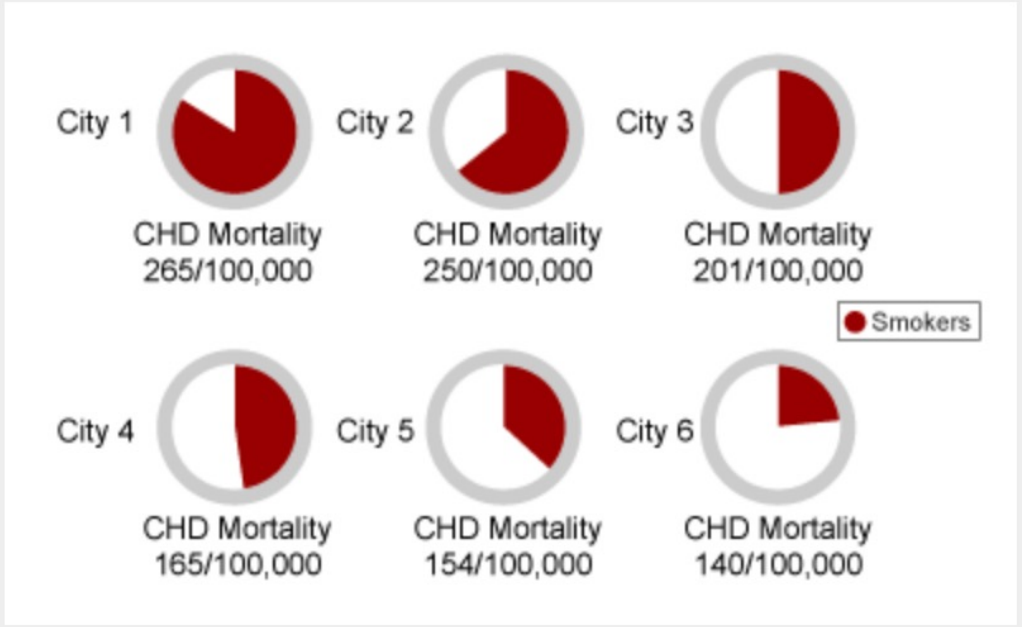
Prospective versus retrospective versus historical

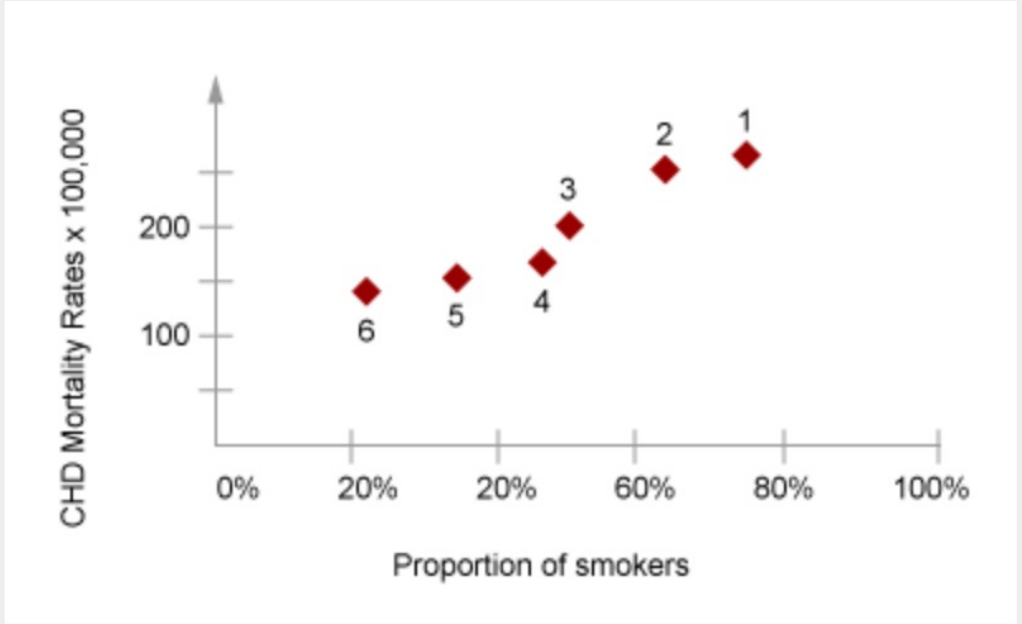
based on when outcomes occurred in relation to the enrollment of the cohort

Group-level data

Ecological studies

- Compare the exposure and outcome status of groups rather than individuals
- Not possible to link the exposure of any particular individual to his or her outcome
- ADV: easy to execute with aggregate publicly-available data; hypothesis generation
- DISADV: *ecological fallacy* – relationship you seen at group-level may not exist at individual-level





MEASURES OF (OUTCOME) OCCURRENCE & MEASURES OF (EXPOSURE) EFFECT

Measures of (outcome) occurrence and (exposure) effect

Type of analytic study	Measure of outcome (disease) occurrence	Measure of (exposure) effect
Ecological	Rate, Risk, Prevalence, Mean or Median	Correlation or Regression Coefficient
Cross-sectional	Prevalence, Odds, Mean or Median	Prevalence Ratio, Prevalence Difference, Odds Ratio, Difference between Means or Medians
Cohort	Rate, Risk, Odds, Mean or Median	Rate Ratio, Risk Ratio, Odds Ratio, Rate Difference, Risk Difference, Vaccine Efficacy, Difference between Means or Medians
Case-control	None*	Odds Ratio,Vaccine Efficacy
Intervention	Rate, Risk, Odds, Mean or Median	Rate Ratio, Risk Ratio, Odds Ratio, Rate Difference, Risk Difference, Vaccine Efficacy, Difference between Means or Median

* Unless the sampling fraction is known for both cases and controls,

- Prevalence
  - The probability of being a case in a population at a given point in time
  - *Unit-less*. State time point with measure

- **Prevalence**

$$= \frac{Number\ cases}{Number\ in\ the\ population}$$

- **Prevalence odds**

$$= \frac{Number\ cases}{Number\ noncases}$$

- Unit-less. Must state the time of observation

- Provides a snapshot
- Numerator is everyone who *does* have the outcome
- Denominator is everyone who *could* have the outcome
- Very helpful to establish burden, coverage, and service need
- Sometimes useful for investigating causes

MEASURES OF OCCURRENCE

- Incidence
  - The probability of **becoming a case** in a population at risk during a specified time period
  - Useful for assessing why people will get disease

- **Incidence risk**

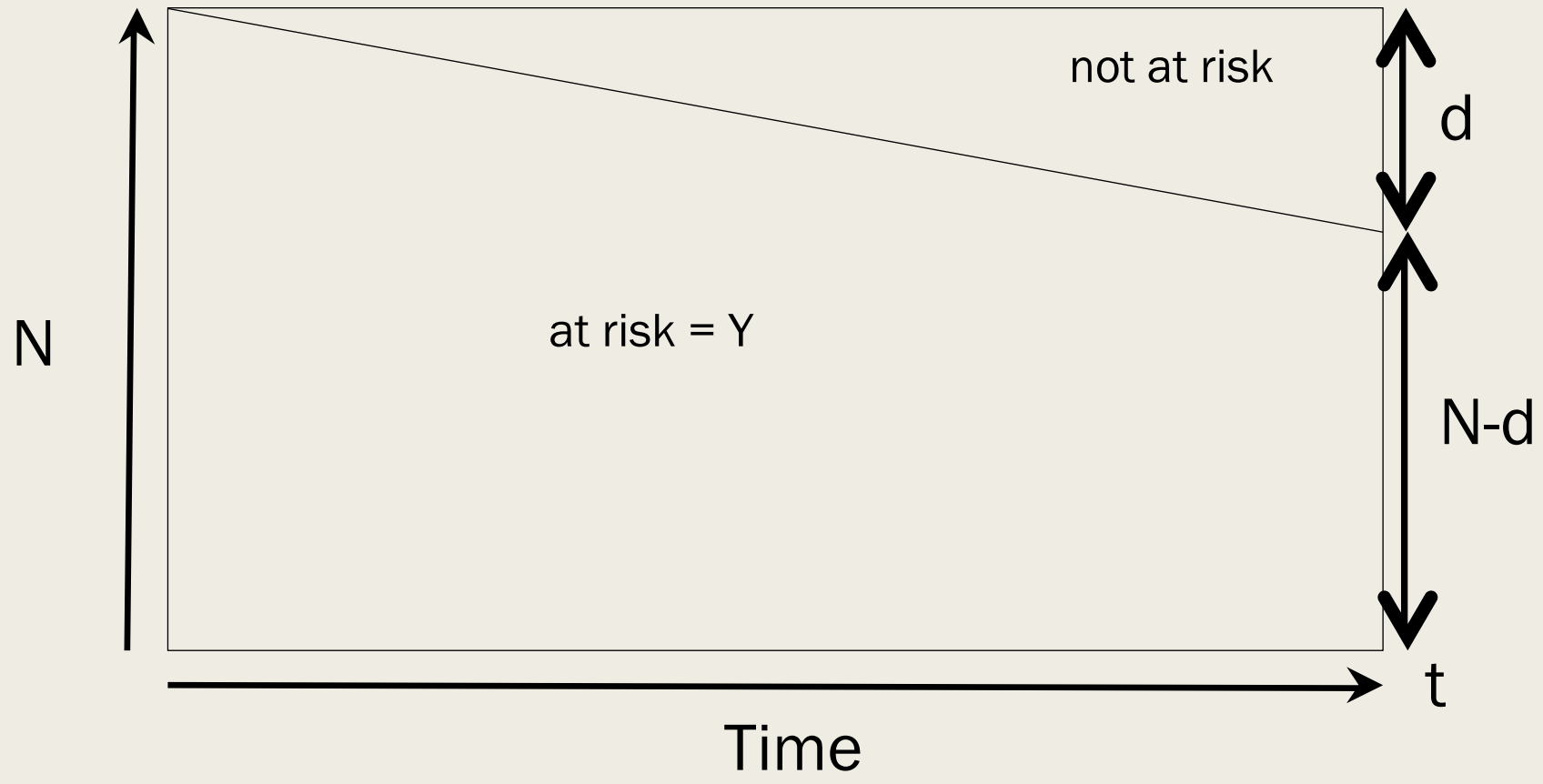
$$= \frac{Number\ of\ new\ cases\ in\ a\ given\ time\ period}{Number\ of\ at\ risk\ at\ the\ start\ of\ the\ time\ period}$$

- **Incidence odds**

$$= \frac{Number\ of\ new\ cases\ in\ a\ given\ time\ period}{Number\ of\ noncases\ at\ the\ end\ of\ the\ time\ period}$$

- Unit-less. Must state the time period with the measurement

STUDY POPULATION



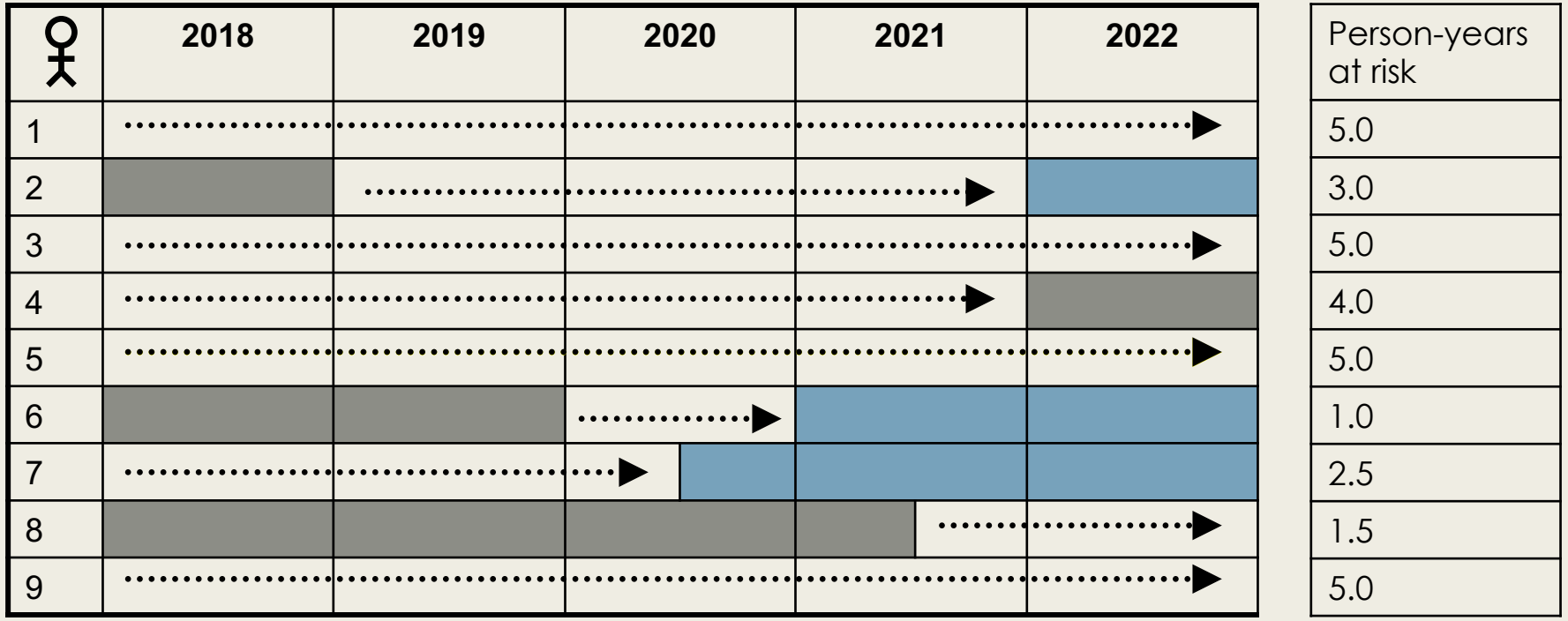
risk over or at time t = d / N
odds over or at time t = d/(N-d)
rate over time t = d/Y

- Incidence risk or incidence odds is a reasonable option for stable populations
  - *Standard start and end time*
  - *Minimal loss to follow-up*

- Not ideal for dynamic populations → Incidence rate

$$\text{Incidence rate} = \frac{New\ cases\ in\ an\ given\ time\ period}{Total\ person\ time\ at\ risk}$$

Cohort study to estimate HIV incidence rate



Person-years at risk
5.0
3.0
5.0
4.0
5.0
1.0
2.5
1.5
5.0

Key: At risk, under observation

Not in study

HIV seroconversion

Total Person-Years at risk = 32.0

New cases during follow-up = 3

Person-time stops accumulating when:

A. study ends, or

B. person develops the condition of interest, or

C. person is lost to follow-up or dies

- Incidence figures require follow up over time

- Incidence figures start with people who are at risk of an outcome
  - Incidence risk uses people at risk in denominator
  - Incidence rate uses person-time at risk in denominator
  - Same numerator: new cases

- Incidence rates are more appropriate when follow-up time varies

- Incidence measures are more appropriate for identifying causes than prevalence measures

Traditionally used measures

- **Case fatality rate**
  - Proportion of cases of a specified cause that die in a specified period of time (%) ~ risk not rate
- **Cause-specific mortality rate**
  - Number of deaths from a specified cause in a defined population in a defined period of time ~ incidence risk not a rate
- **Infant mortality rate (IMR)**
  - *Number of deaths in children under-one / number of live births in the same period in a specified population*
  - *Is this is a risk or a rate or neither?*

MEASURES OF (EXPOSURE) EFFECT

Teenage pregnancy

Girls exposed to 'electronic babies' more likely to become pregnant, study finds

More girls in Australian study who used the dolls – designed to prevent teenage pregnancy – became pregnant than those who did not

Teenagers in London with the electronic dolls, which have been used in 89 countries. Girls taking part in an Australian study were found to be more likely to become pregnant if they had been exposed to the dolls. Photograph: Graham Turner for the Guardian

"...that 17% of girls who used the dolls had become pregnant by the age of 20, compared with 11% of those who did not."

2 x 2 contingency table

	Disease	No disease	Row total
Exposed	a	b	a + b
Unexposed	c	d	c + d
Column total	a + c	b + d	a + b + c + d=N

Risk ratio= R_1/R_0

R_1= a / (a+b)

R_0= c / (c+d)

Odds ratio= $Odds_{exp}/Odds_{unexp}$

$$OR = \frac{(a/b)}{(c/d)} = \frac{a \times d}{b \times c}$$

	Pregnant	Not pregnant	Total
Doll user	210	1057	1267
Non-user	168	1399	1567

Baseline group

Risk of pregnancy among doll users = 210/1267 = 0.166 = 17%

Risk of pregnancy among non-users = 168/1567 = 0.107 = 11%

$$Incidence\ risk\ ratio = \frac{210/1267}{168/1567} = 1.55$$

	Pregnant	Not pregnant	Total
Doll user	210	1057	1267
Non-user	168	1399	1567

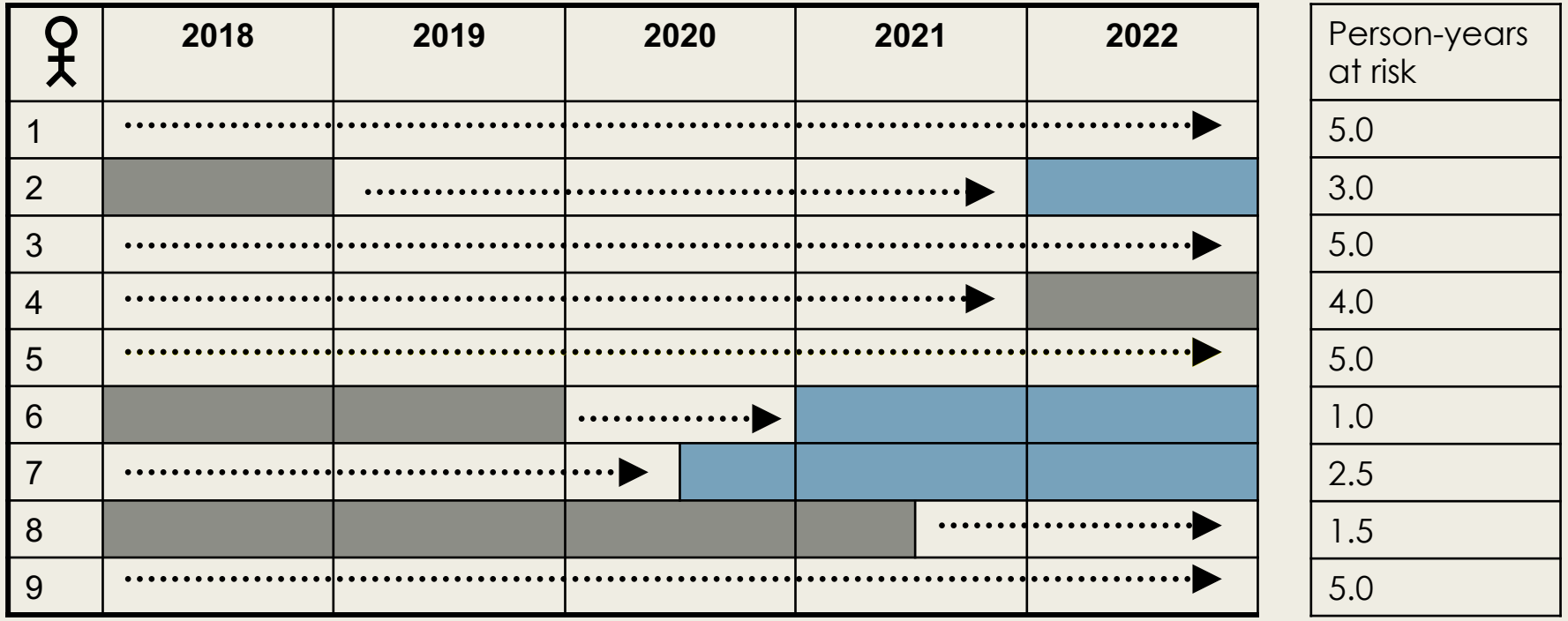
Baseline group

Risk of pregnancy among non-users= 168/1567 = 0.107 = 11%

Risk of pregnancy among doll users = 210/1267 = 0.166 = 17%

$$Incidence\ risk\ ratio = \frac{168/1567}{210/1267} = 0.65$$

Cohort study to estimate HIV incidence rate



Person-years at risk
5.0
3.0
5.0
4.0
5.0
1.0
2.5
1.5
5.0

Key: At risk, under observation

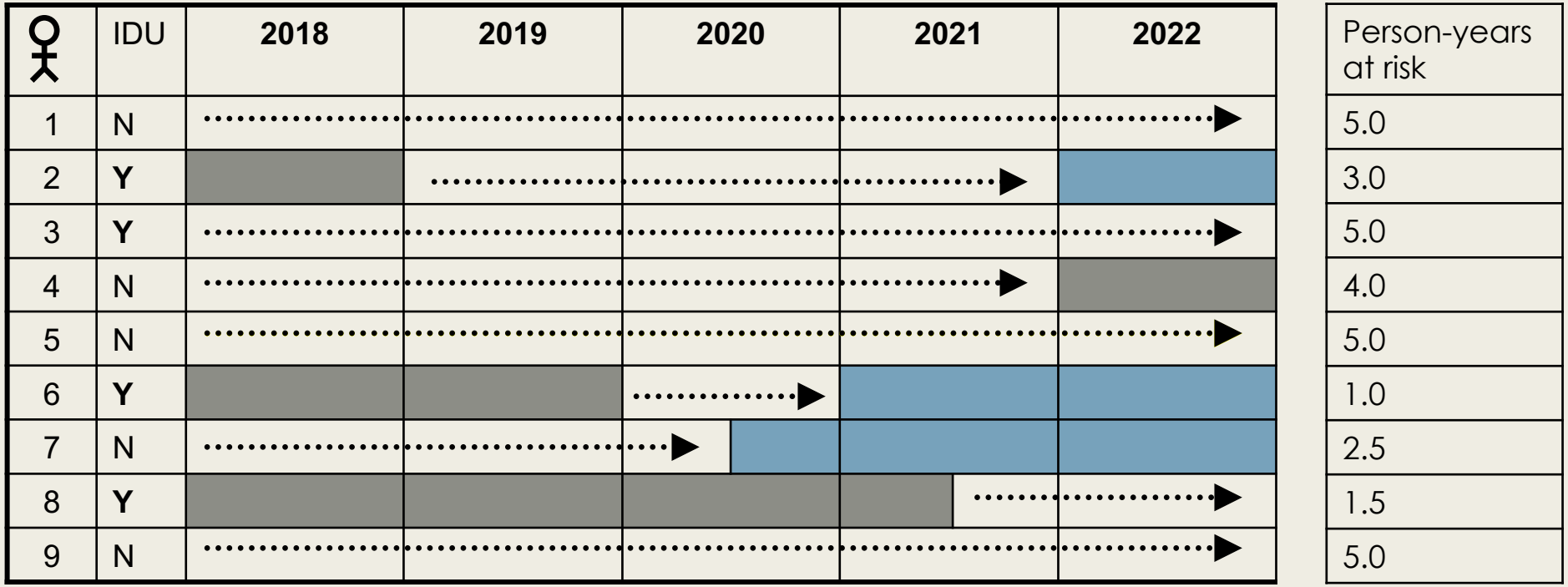
Not in study

HIV seroconversion

Total Person-Years at risk = 32.0

New cases during follow-up = 3

Cohort study to estimate HIV incidence rate by whether intravenous drug use (IDU) status



Person-years at risk
5.0
3.0
5.0
4.0
5.0
1.0
2.5
1.5
5.0

IDUs have 10.5 person-years at risk and 2 HIV infections

Non users have 21.5 person-years at risk and 1 HIV infection

	HIV infection	Person-years
IDU	2	10.5
Non-user	1	21.5

Rate of HIV among IDUs
= 2 / 10.5 = 0.19
= 19 per 100 person-years

Rate of HIV among non-users
= 1 / 21.5 = 0.046
= 4.6 per 100 person-years

$$Incidence\ rate\ ratio = \frac{2/10.5}{1/21.5} = 4.09$$

Risk (or rate) difference (RD)

- Absolute (actual) difference between two risks (or rates)
- Subtract the risk (or rate) in unexposed (r_0) from the risk (or rate) in the exposed (r_1)

$$\text{Risk (or rate) difference} = (r_1 - r_0)$$

Provides and absolute measure of public health burden

Not all associations are causal or independent

- Random error (chance)
- Bias
- Confounding
- Real association

"Be as sceptical as possible as to reason for an association"

CAUSALITY

Causal effects?

Key premise:

- Diseases have causes
- Some causes can be partly or fully eliminated→ prevent some cases of disease

Problem

- Most epidemiological studies are observational and measure associations

Association ≠ Causation

Establishing causality

- Notions of causes of disease are inextricably linked to:
  - Current levels of scientific knowledge
  - Current scientific paradigms
    - Guiding theoretic concepts of a science at a specific point in time

Bradford Hill (1965): aspects to consider

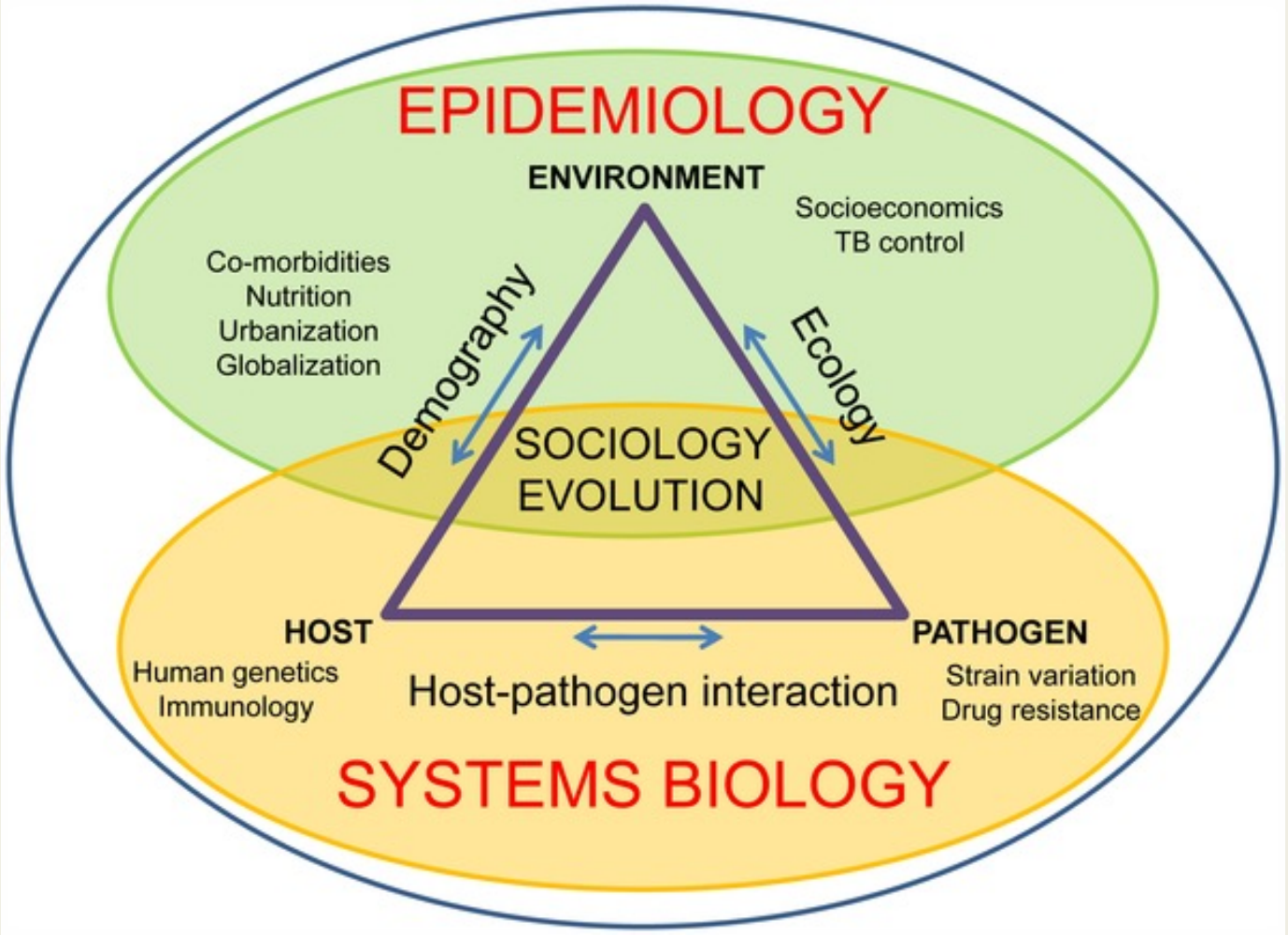
- **Temporality** Does the exposure happen before the disease?
- **Strength of association** Is the association (e.g. RR, OR) strong?
- **Biological gradient (dose-response)** Does more exposure lead to more diseases?
- **Experiment (reversibility)** Would removing the exposure reduce the burden of disease?
- **Consistency** Is the association consistent with previous findings?
- **Biological plausibility** Is the association biologically plausible?
- **Coherence** Does the association conflict with our understanding of the disease?
- **Specificity (Koch-Henle)** Is the association specific to this exposure/disease?
- **Analogy** Is the association analogous to other cause-effect relationships?

Epidemiological triad

"Systems epidemiology approach"

- **Pathogen (exposure)**
- Host
- Environment

- Epidemiology addresses the burden of the disease and the social, economic, and ecological causes of its frequency and distribution
- Systems biology integrates approaches that address the host, the pathogen, and interactions between the two



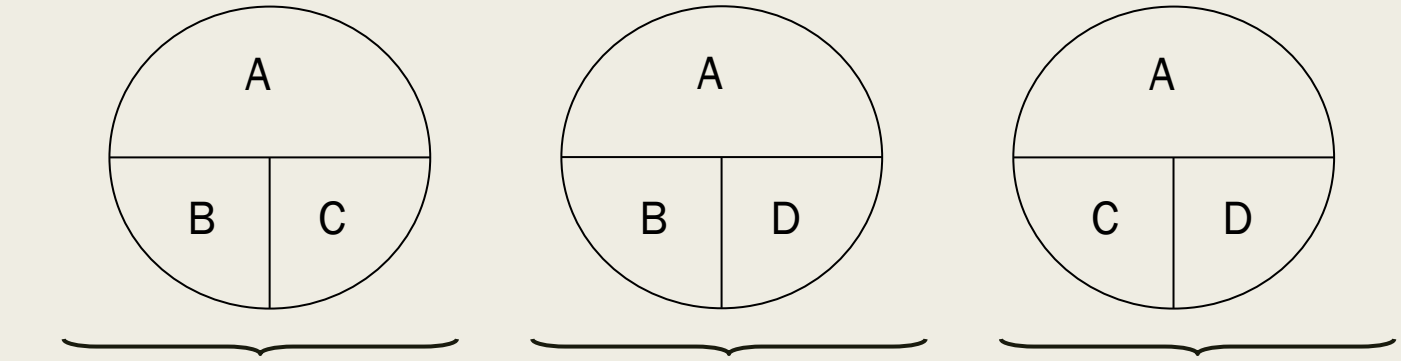
Comas I, Gagneux S (2009) The Past and Future of Tuberculosis Research. PLOS Pathogens 5(10): e1000600. https://doi.org/10.1371/journal.ppat.1000600

Necessary but not sufficient

- Growing recognition that while the exposure, e.g. an infectious agent may be necessary for disease to occur, it is far from sufficient
- Epidemiologists such as Wade Hampton Frost and others devoted increasing attention to the broader social and environmental context of disease occurrence
- Appreciation of complexity of the interaction of genes (host) and environment

Multi-cause paradigm

- **Agent-host-environment** model did not work for many non-infectious diseases
- Rothman's Causal Pies (1976)
- Multiple causes act together to produce disease → disease can result from more than one causal mechanism



Causal mechanism

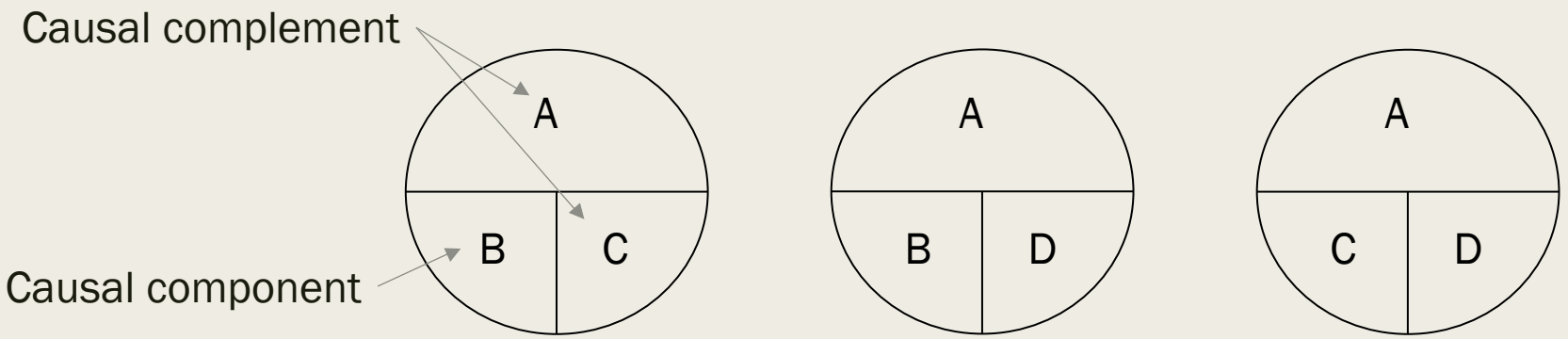
A necessary but not sufficient

B on its own is not necessary or sufficient:

- some individuals with B will get disease, but so will some individuals without B (e.g. *hepatitis B infection and liver cancer)*

Causal pies

- Strength of a causal component effect on disease occurrence in a particular population depends on the prevalence of its causal complement
  - *a factor may be an important cause of a disease in one population, but not in another population*



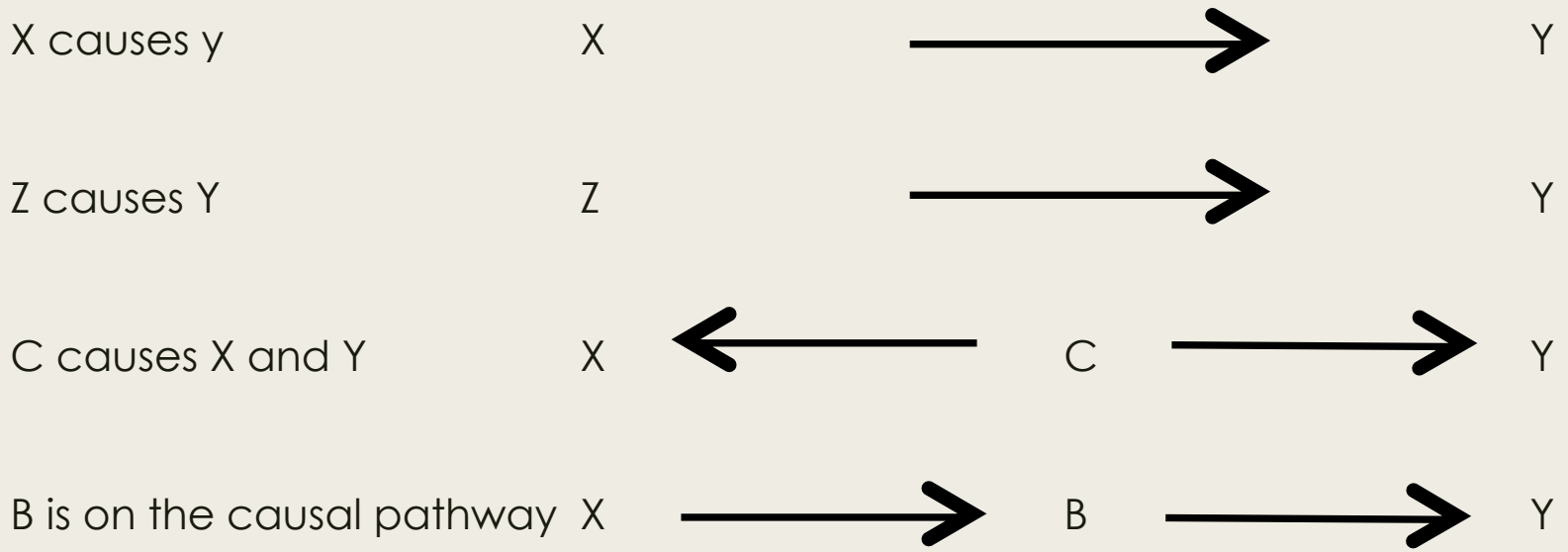
Conceptual frameworks

- Define your research question: exposure and outcome
- Set down clearly assumptions about the inter-relationships between variables (literature review; domain expertise)
- Consider causes of diseases operating at many levels (hierarchical)
- Describe these relationships pictorially
- Given these assumptions, we can:
  - *Summarise what is currently known*
  - *Define the data to be collected*
  - *Guide the statistical analysis and interpretation of findings*

Causal diagrams

- Causal diagrams explicitly consider the interrelationships between different components in a causal pathway

- For a specific study question, they help identify:
  - potential confounders
  - mediating factors (intermediate steps in causal pathway)



Reading material on approaches to multi-causality

- Conceptual frameworks:
  - Victora et al. The role of conceptual frameworks in epidemiological analysis: a hierarchical approach. Int J Epidemiol 1997;26:224-27

- Causal diagrams:
  - Greenland et al. Causal diagrams for epidemiologic research *Epidemiology,* 1999; 10: 37-48

INTRO TO DAGS

Directed acyclic graphs

- Directed acyclic graphs are a common form of '**causal diagram**'
- They are an aid to planning and conducting observational data analysis where we seek to estimate **causal effects**
- **Causal effects** are interesting, because they symbolise the (potentially modifiable) effect of an **exposure** on an **outcome**
- Most epidemiology and quantitative social science is ultimately interested in estimating causal effects

DAGs are directed because causality is directed (over time); a cause cannot occur after a consequence

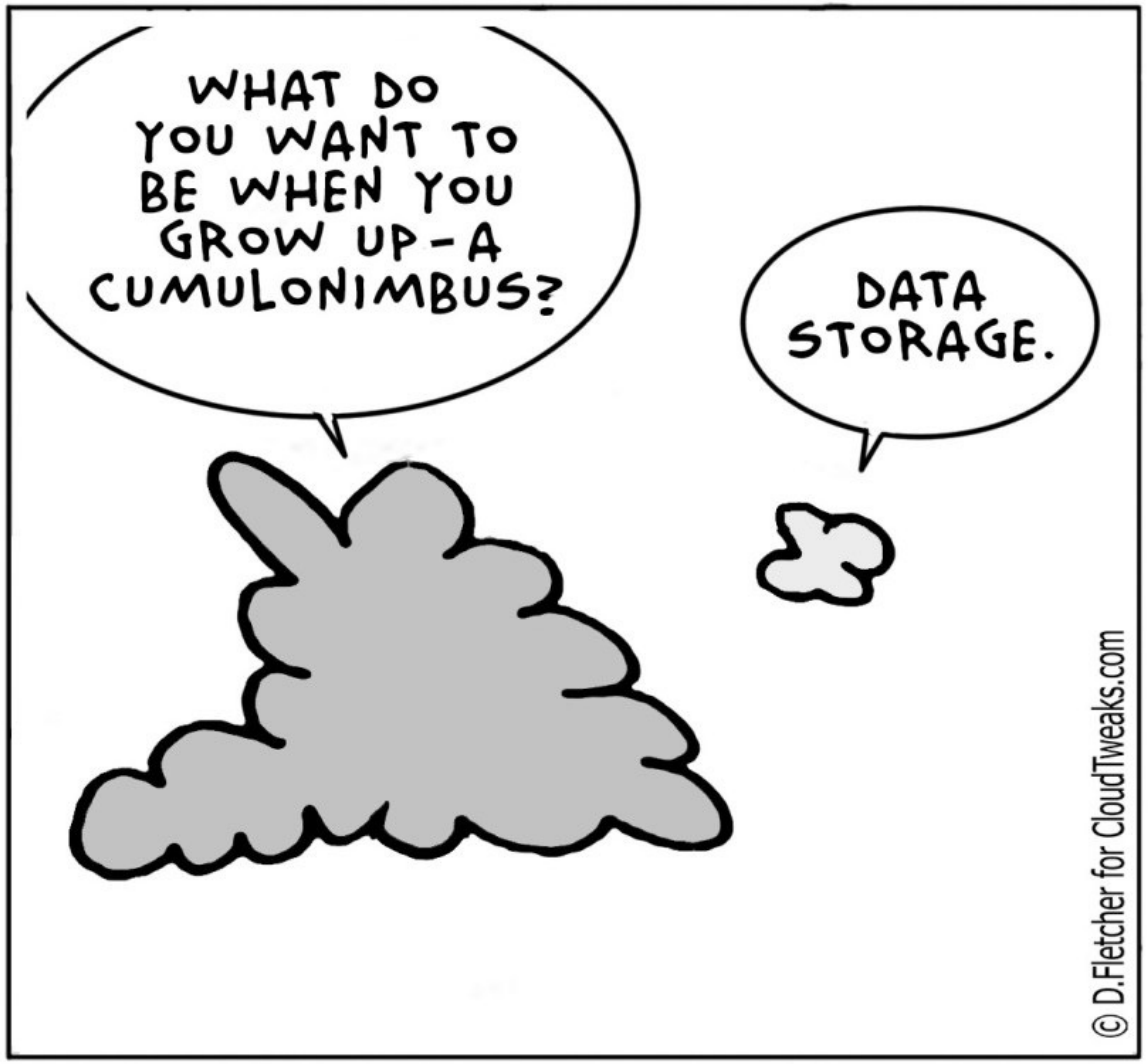
When we draw an arc between X and Y, we state that we believe:

- Changing X modifies the probability of Y (probabilistic reasoning)
- If X had been different, Y would have been different (counterfactual reasoning)
  - Y 'listens' to X
  - If we could wiggle X, it would wiggle Y

"we must emphasize that no approach solves the central … problem of inferring causation from non-experimental data.

….. all causal inference is based on assumptions that cannot be derived from observations alone."

Greenland et al. Epidemiology, 1999; 10: 37-48.


WHAT DO YOU WANT TO BE WHEN YOU GROW UP - A CUMULONIMBUS?
DATA STORAGE.
© D.Fletcher for CloudTweaks.com

DAGitty – draw and analyse causal diagrams

- DAGITTY

10 steps to drawing a DAG like a pro!

1. Develop and state a clear research question
2. Consider and state your context
3. Draw your DAG(s) as early as possible
4. Get help - don't draw it alone
5. Include all relevant variables
6. Draw your DAG(s) in temporal order
7. Draw forward arcs, unless confident otherwise
8. Check & update your DAG(s) against your data
9. Use your DAG(s) to inform and interpret your model
10. Share & publish your DAG(s)