Workshop on measles and rubella modelling,
IIT Bombay, Mumbai, 5-9 February 2024

Introduction to epidemiological thinking

Mark Jit
London School of Hygiene & Tropical Medicine

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Overview

I. Measurement
II. Bias and confounding
III. Epidemiological studies
IV. Causation

Illustration: “Death's Dispensary” (John Pinwell). Drawn at 1866, at about same time that John Snow found a link between drinking contaminated water and cholera. The water supply of London, like that of other major European capitals, was untreated river water.

I.
Measurement

What is epidemiology?

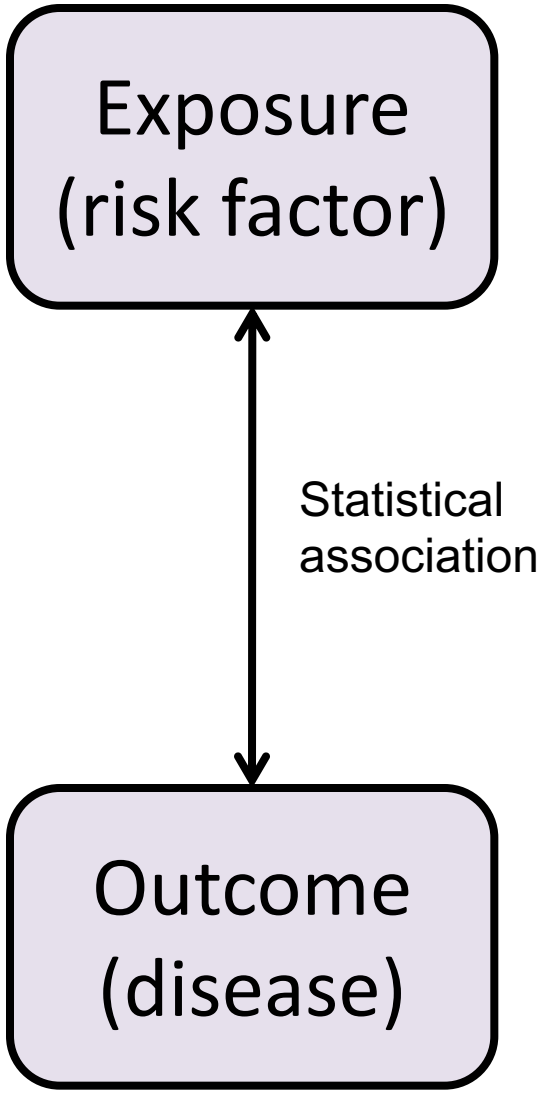
EPIDEMIOLOGY

epi (Gk.) =	*demos* (Gk.) =	*logos* (Gk.) =
upon	population	word, study

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

Last, J. M. (1995) *A Dictionary of Epidemiology* (3rd edition)

Aims of an epidemiological study



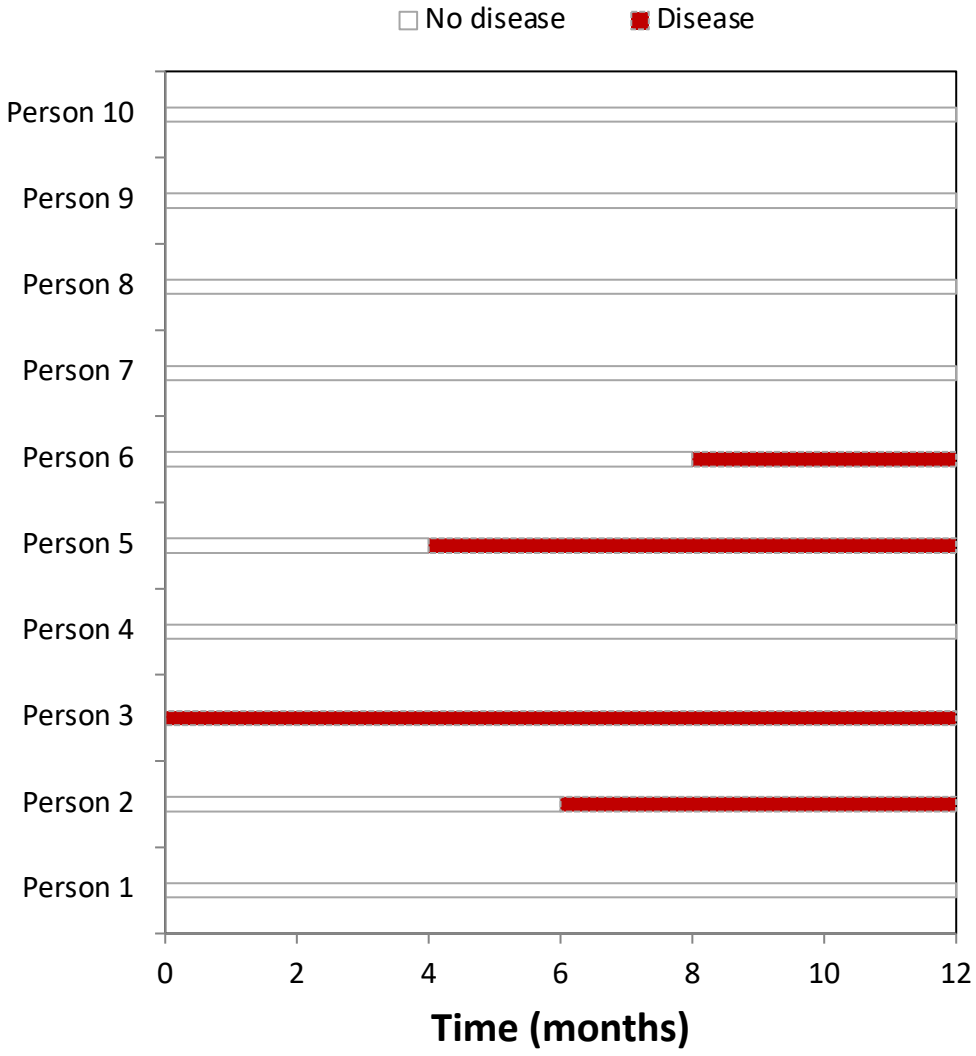
Measuring states: Prevalence and incidence

Prevalence
Proportion of people in a health (or disease) state at a particular time.

At time 12 months, the prevalence of the disease is 4/10 = 0.25

Incidence
Proportion of people who newly acquire a health (or disease) state over a period of time.

Over 12 months, the incidence of the disease is 3/10 = 0.33



Incidence measures: risk, rate and odds

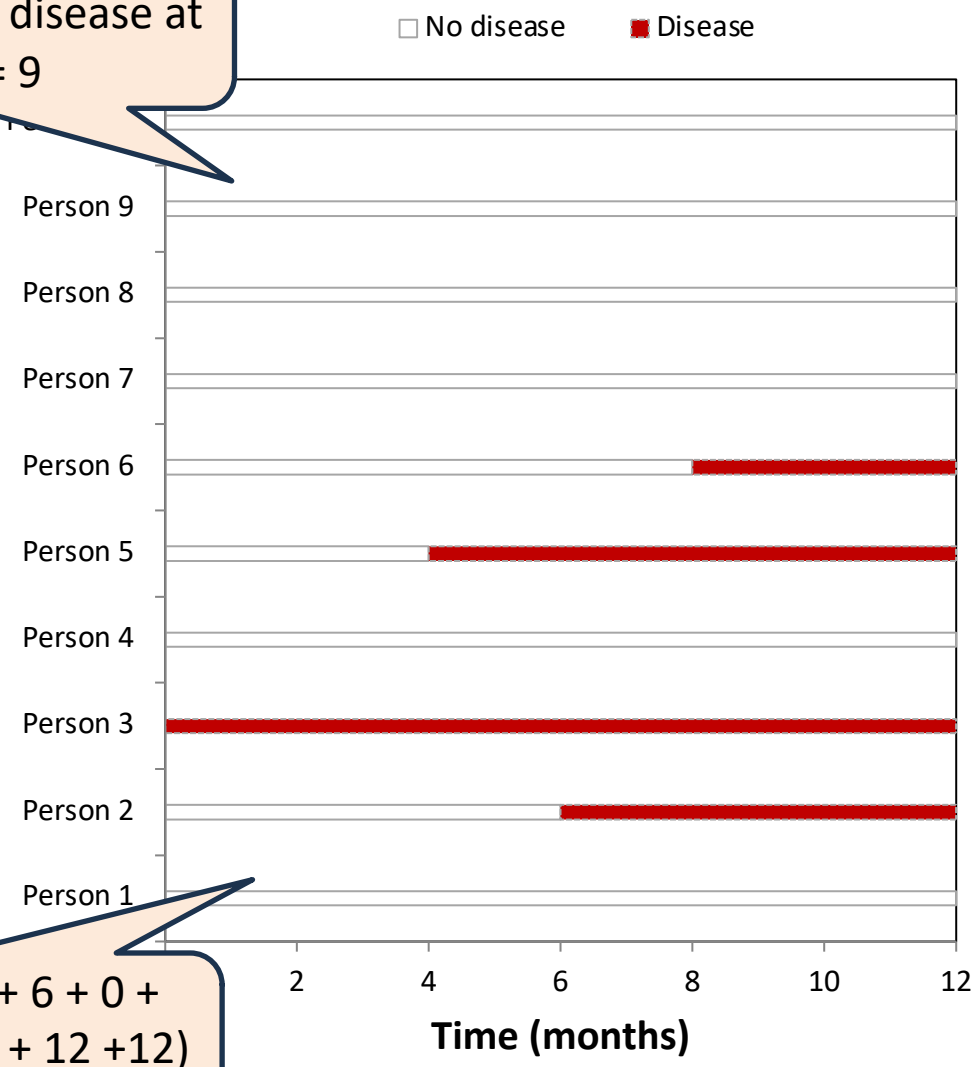
Risk
Ratio of new cases to number of people at risk

Over 12 months, the risk of the disease is 3/9 = 0.33 per person

Rate
Ratio of new cases to total time at risk

Over 12 months, the rate of the disease is (3 cases)/(90 person-months) = 0.4 cases per person-year

People at risk = 10 – 1 (1 person already has disease at the start) = 9



Time at risk = (12 + 6 + 0 + 12 + 4 + 8 + 12 + 12 + 12 +12) = 90 months

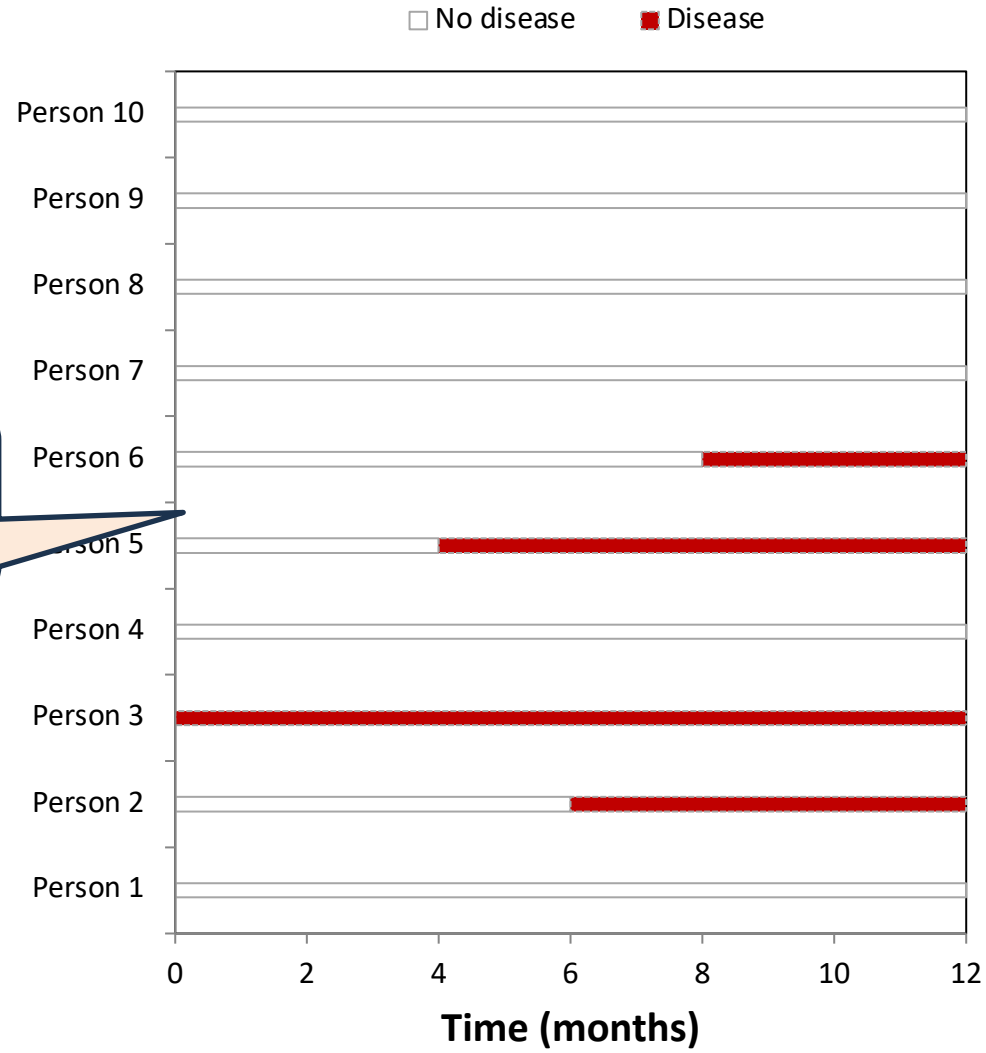
Incidence measures: risk, rate and odds

Odds

Ratio of new cases to non-cases

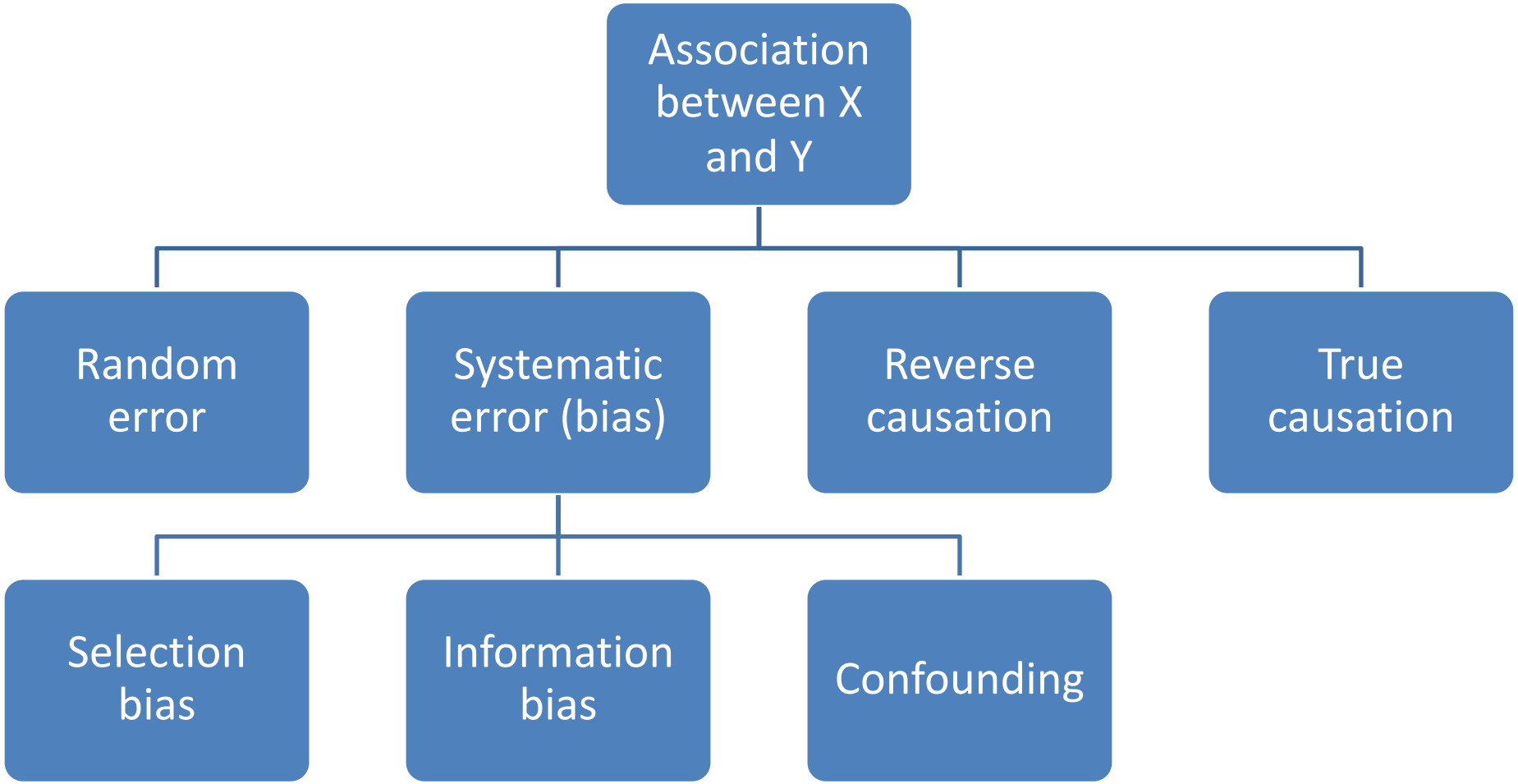
Over 12 months, the odds of the disease are 3/6 = 0.5

3 people become new cases
6 people are not cases

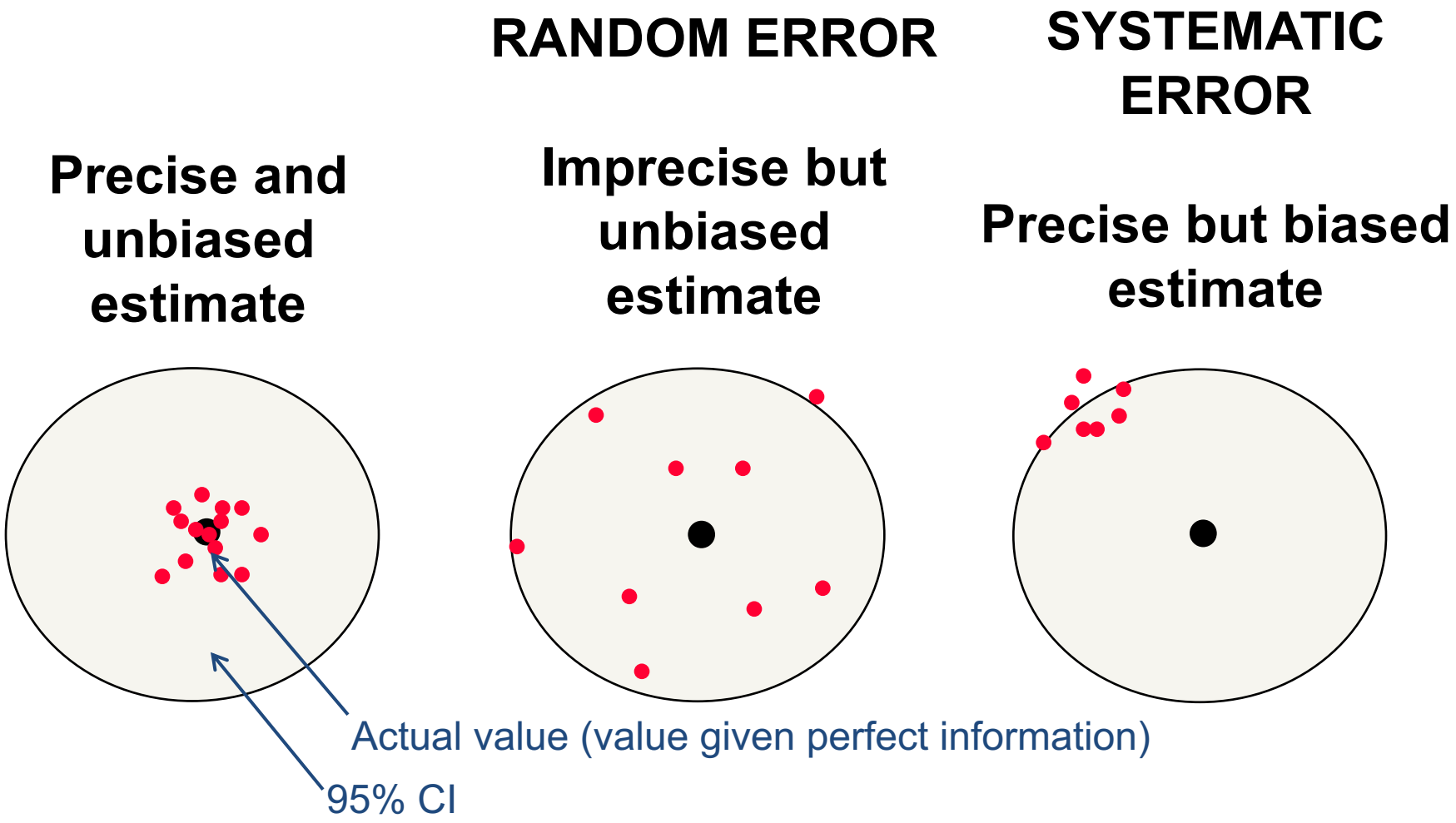


II.
Bias and confounding

Association and causation

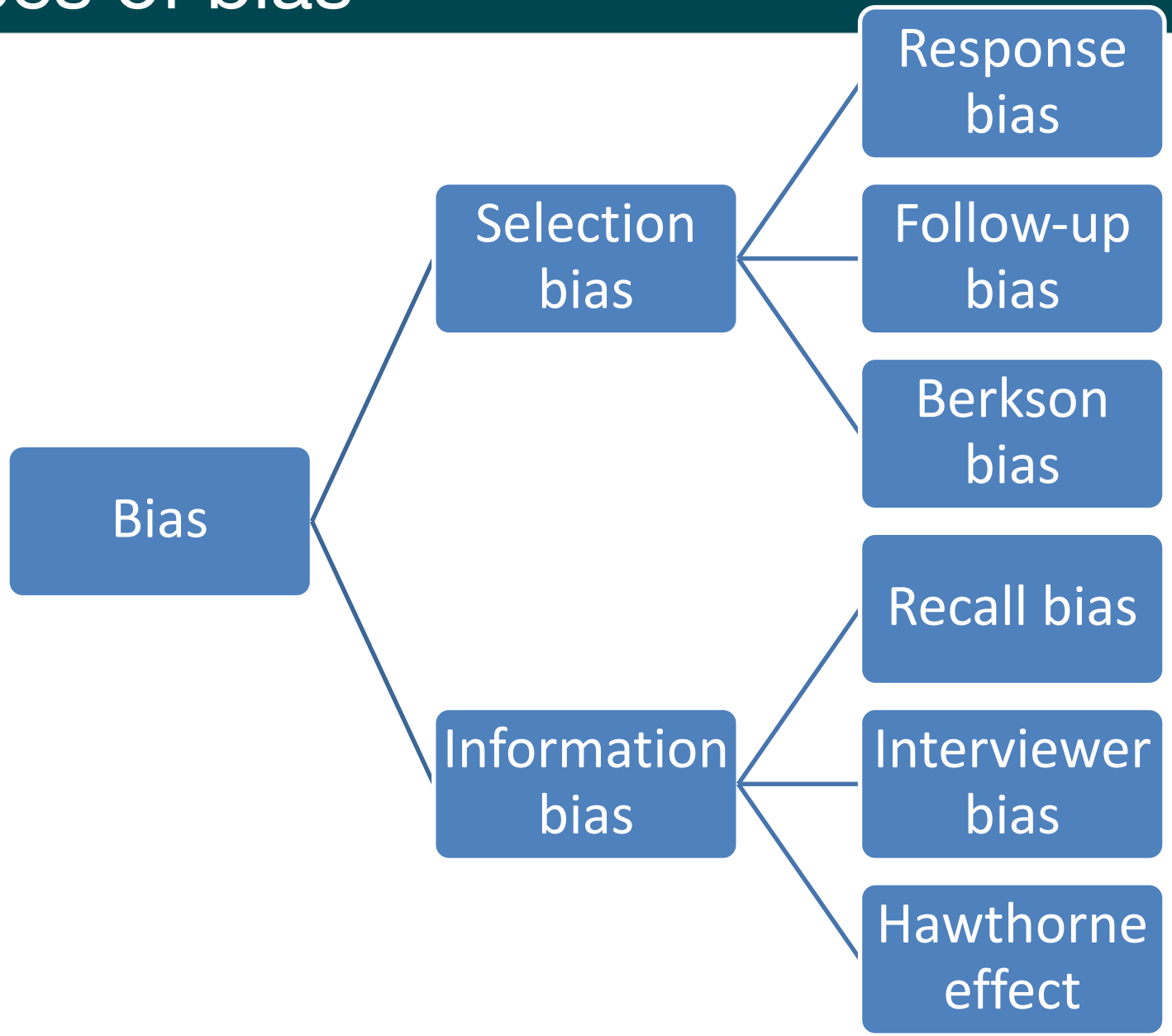


Precision and validity

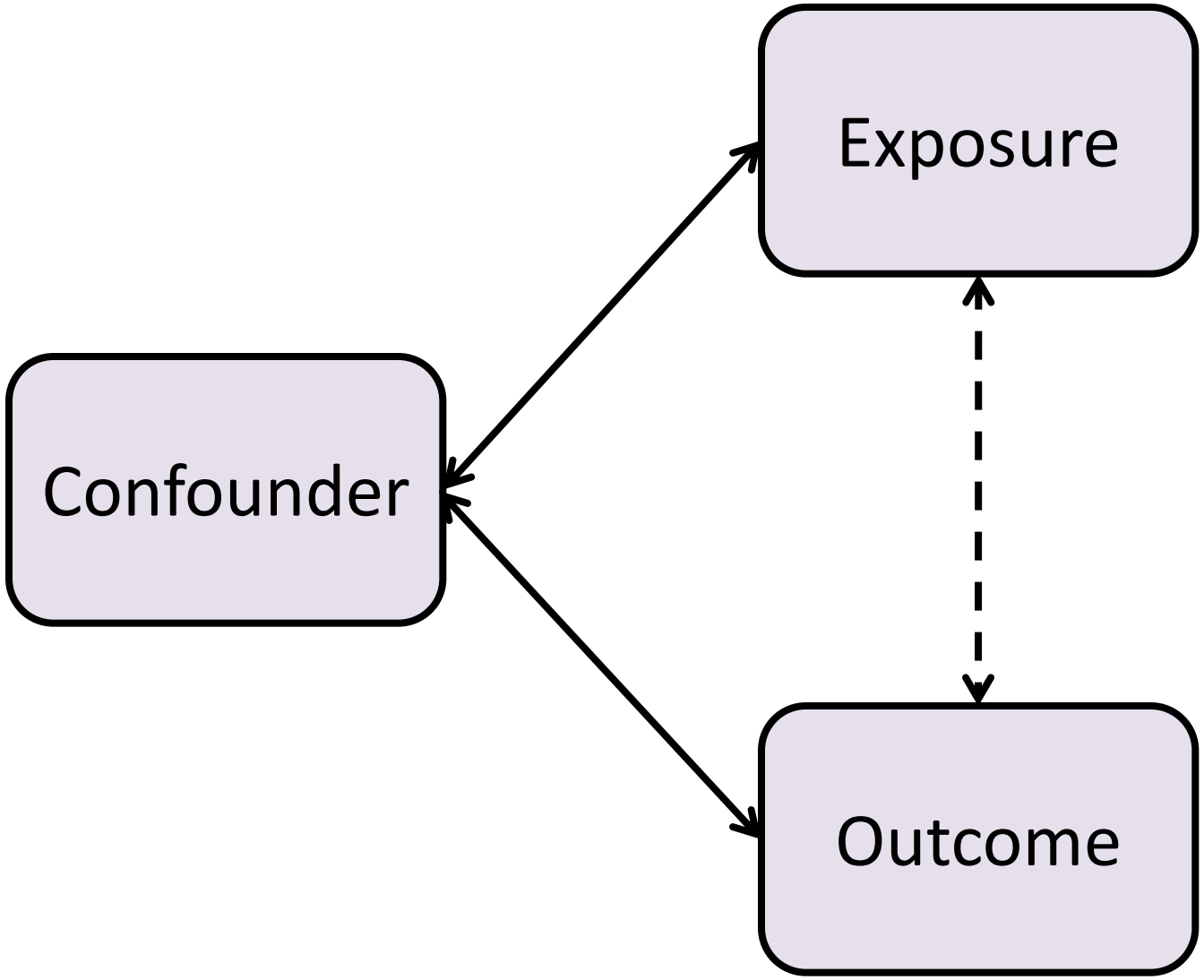


p<0.05 means the result is unlikely to have occurred due to random error. It does not guarantee protection against a result due to systematic error!

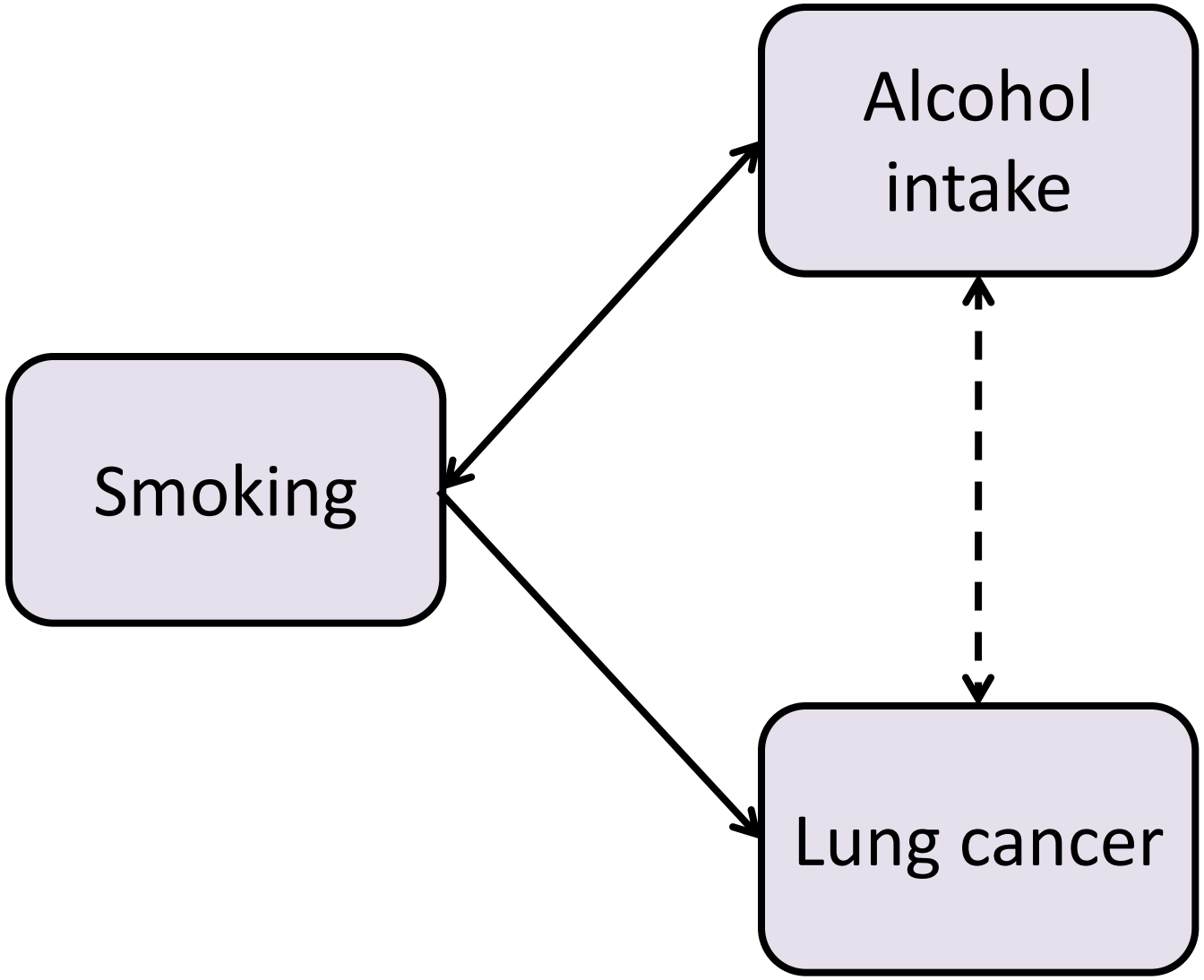
Some sources of bias



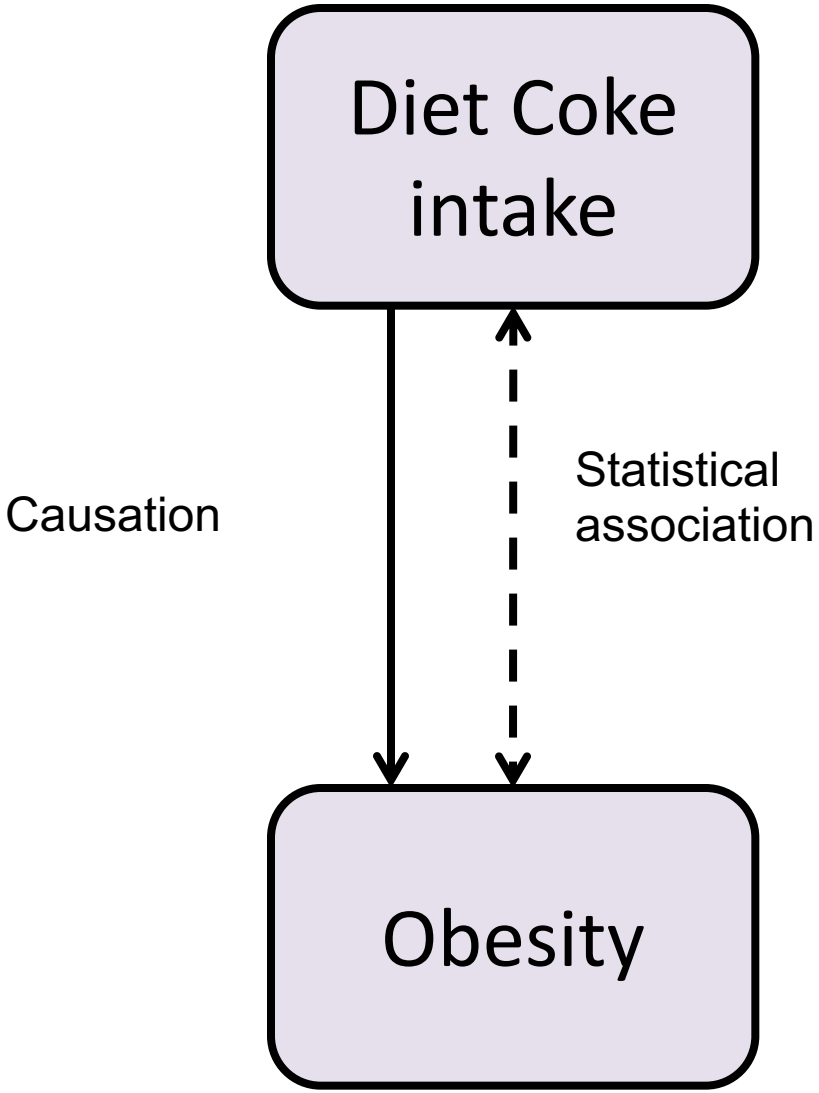
Confounding



Confounding

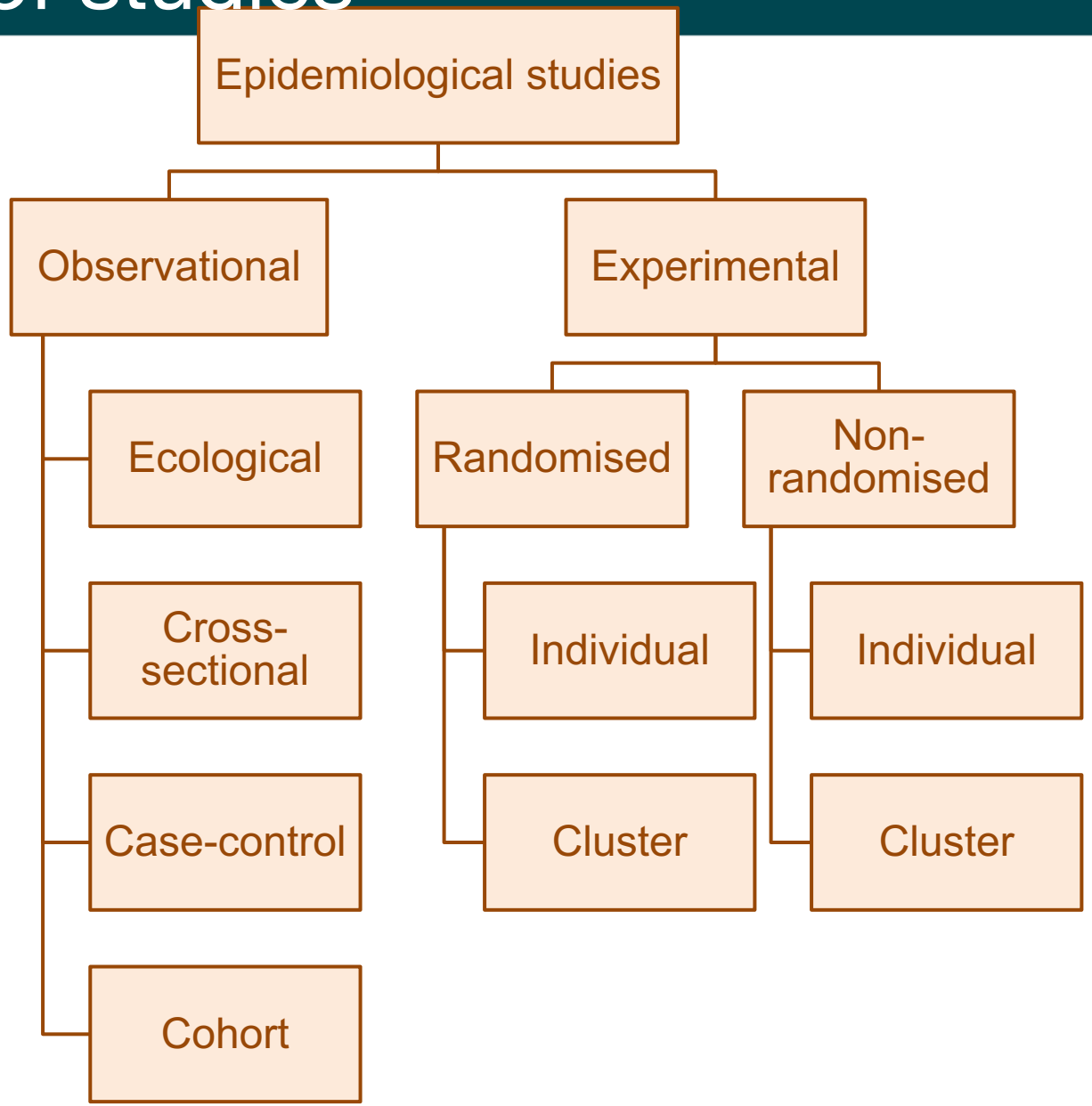


Reverse causation



III. Epidemiological studies

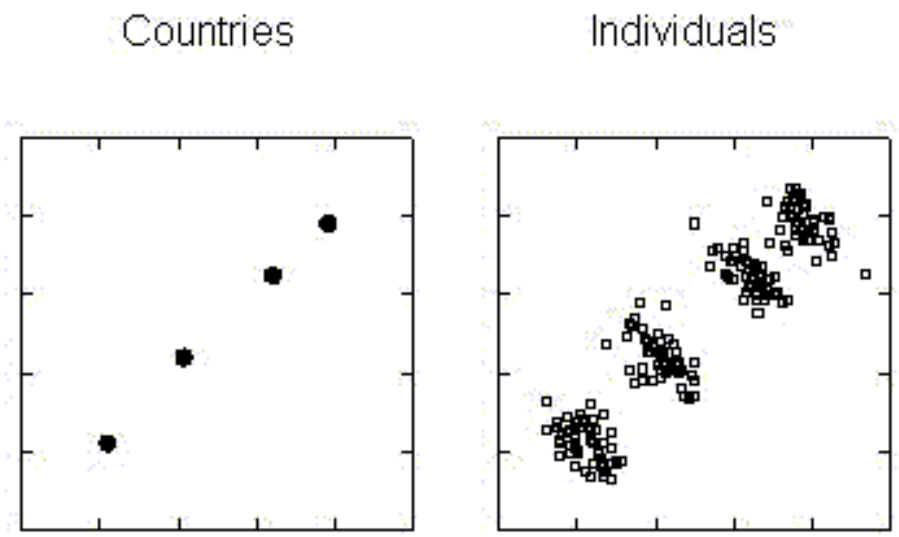
Classification of studies



Ecological studies

The ecological fallacy

Unit of assignment & analysis in ecological studies = GROUP



Variation within groups eliminated

Strength of association (including effect of confounders) overestimated

Association at GROUP level may not be replicated at INDIVIDUAL level

Cross-sectional studies

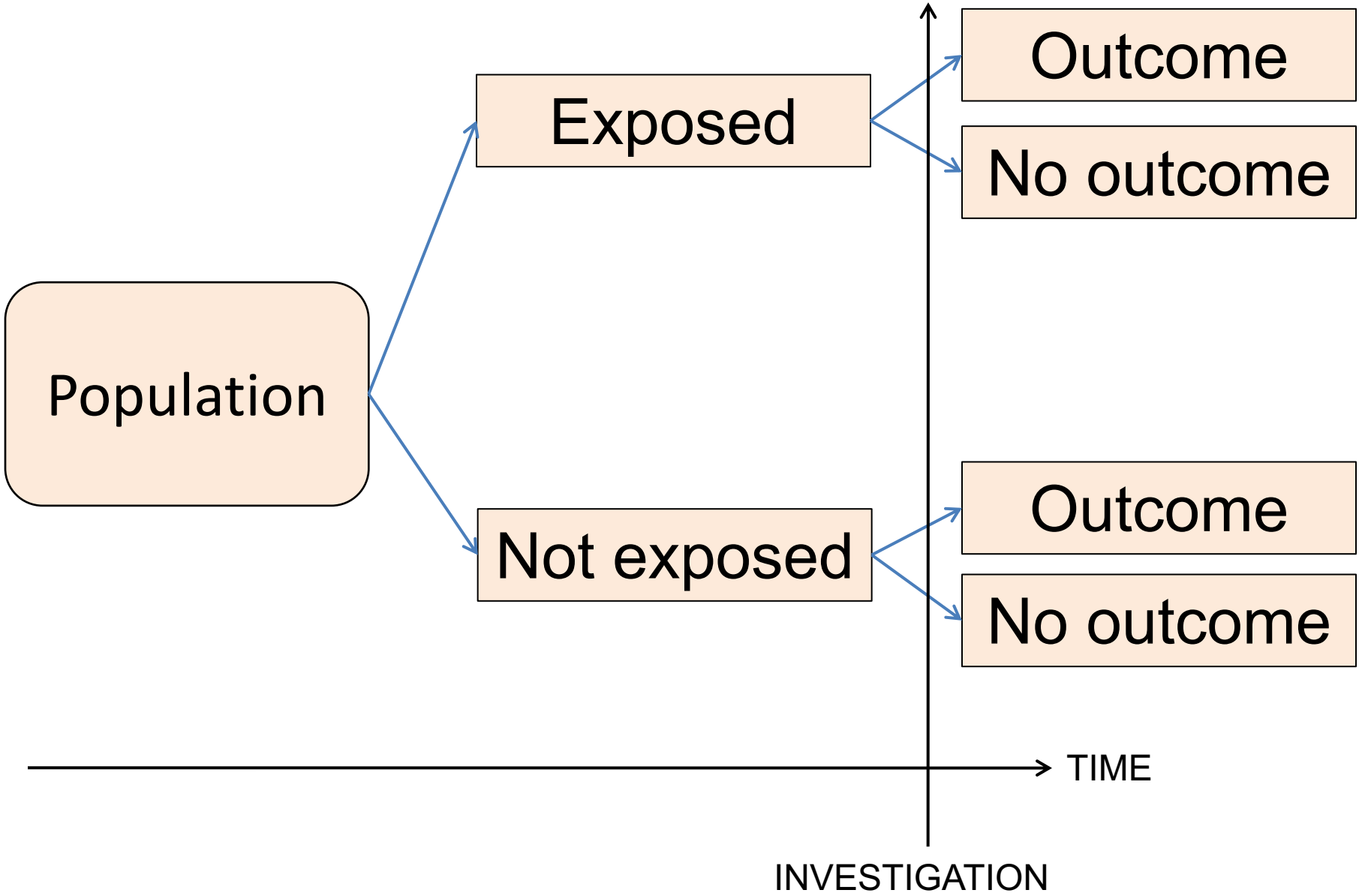
Choose the source population.

↓

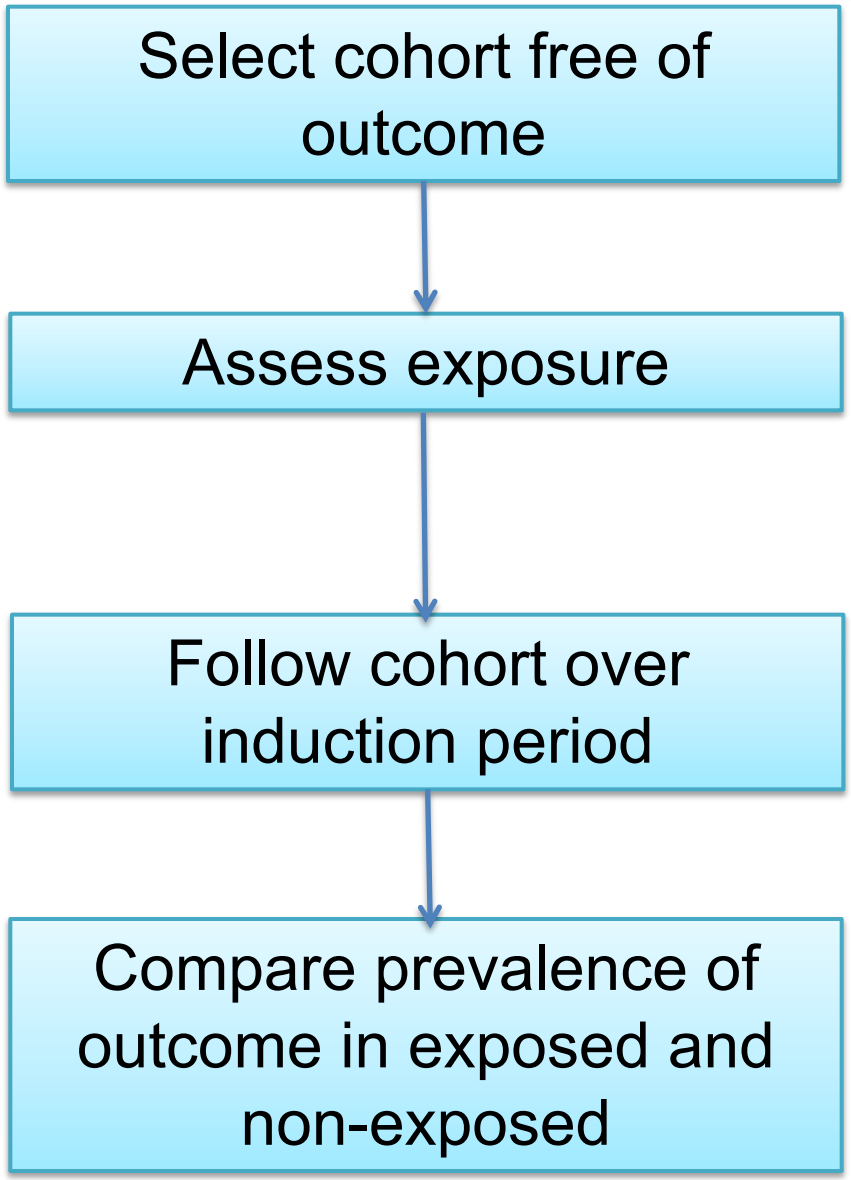
Measure exposures and outcomes in the population at a point in time.

ADVANTAGES	DISADVANTAGES
Can collect data on a range of outcome and exposure variables at the same time.	No control over exposure of interest or confounders.
Useful when exposure is fixed (eg sex, ethnicity, genetic makeup).	Cannot show temporality.

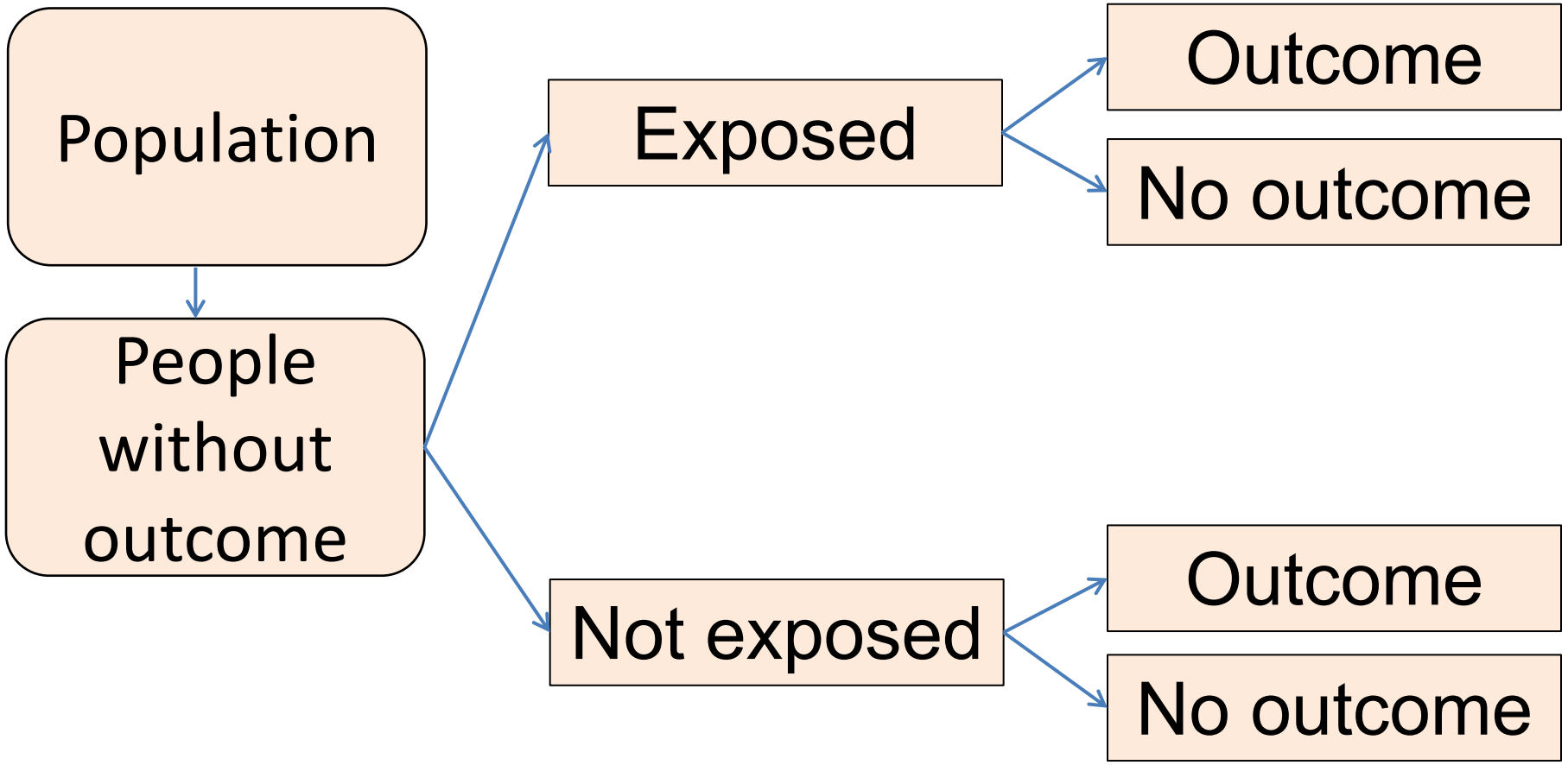
Cross-sectional studies



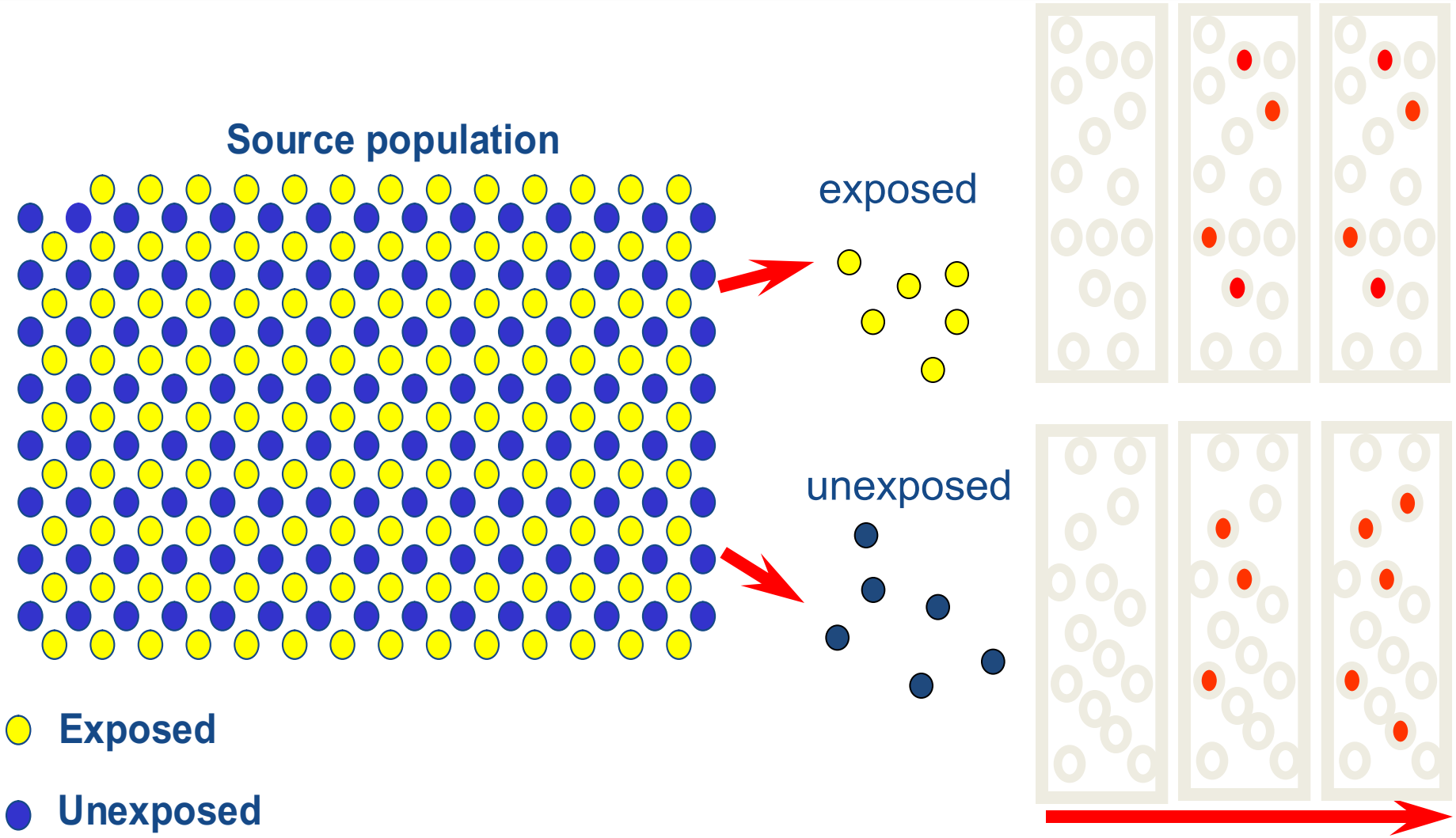
Cohort studies



Cohort studies



Cohort studies



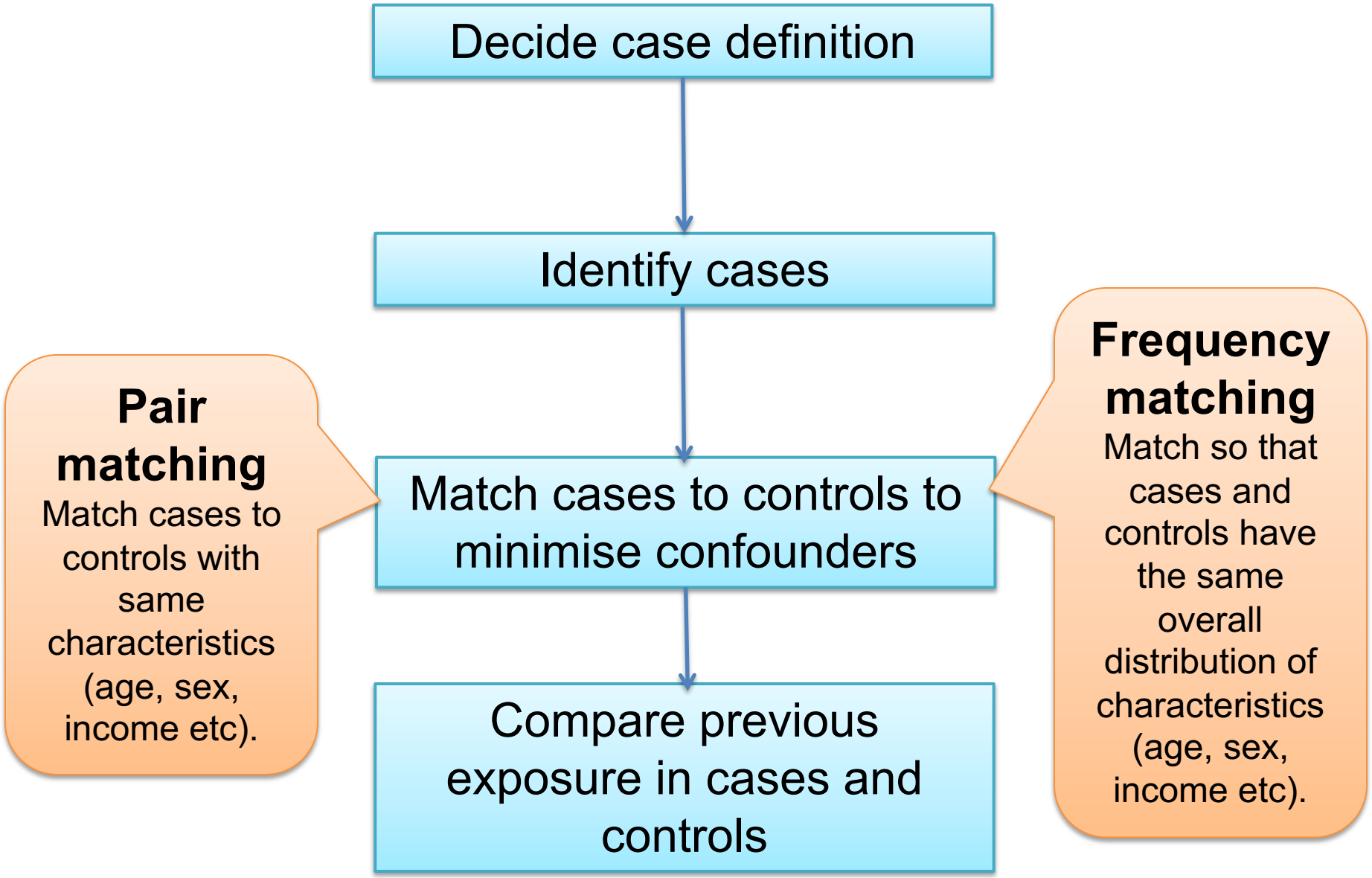
Cohort studies

This is called a 2×2 table

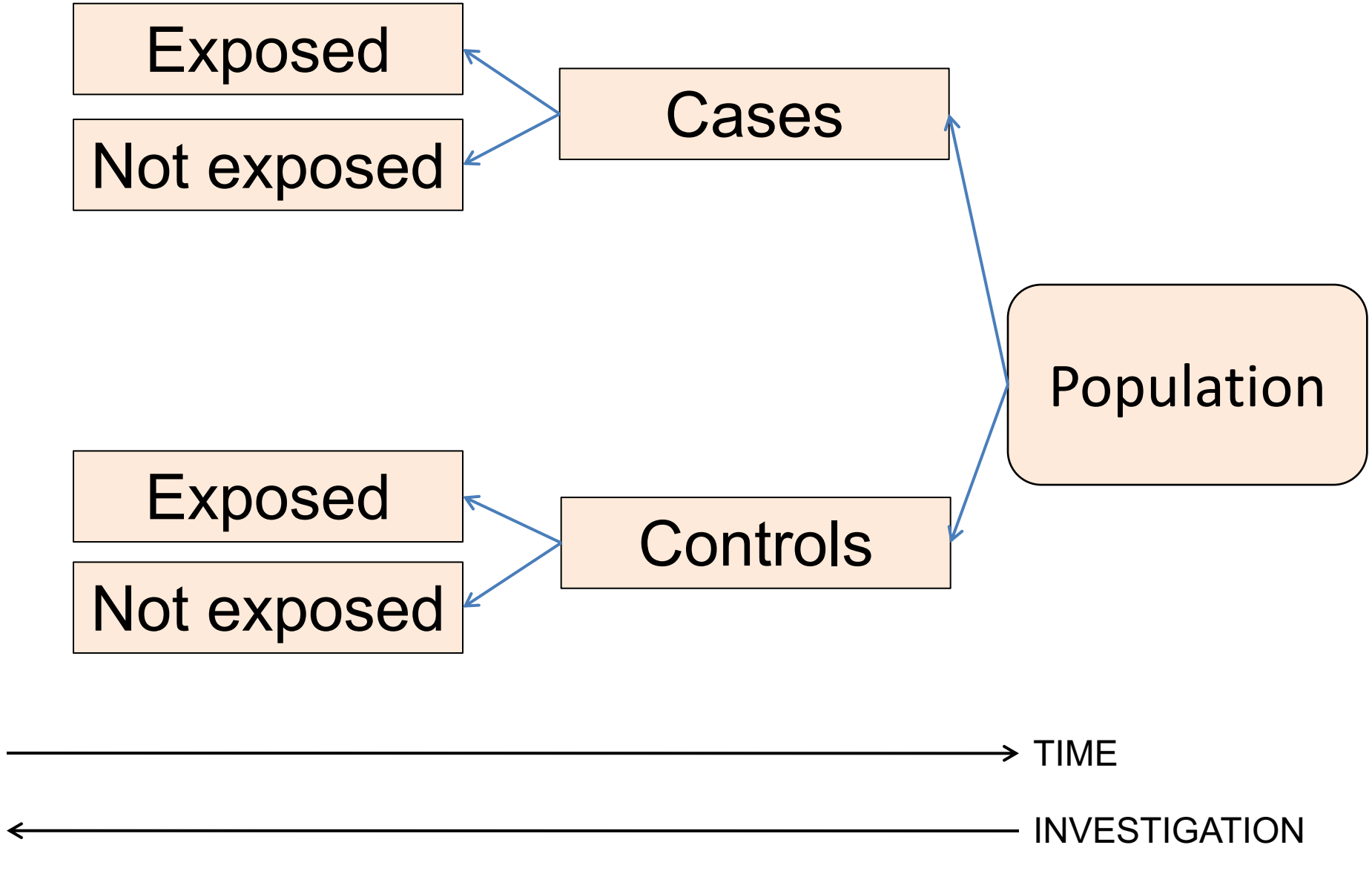
	Outcome	No outcome
Exposed	a	b
Unexposed	c	d

Risk in exposed (RE)	a / (a+b)
Risk in unexposed (RU)	c / (c+d)
Relative risk (RR)	RE / RU

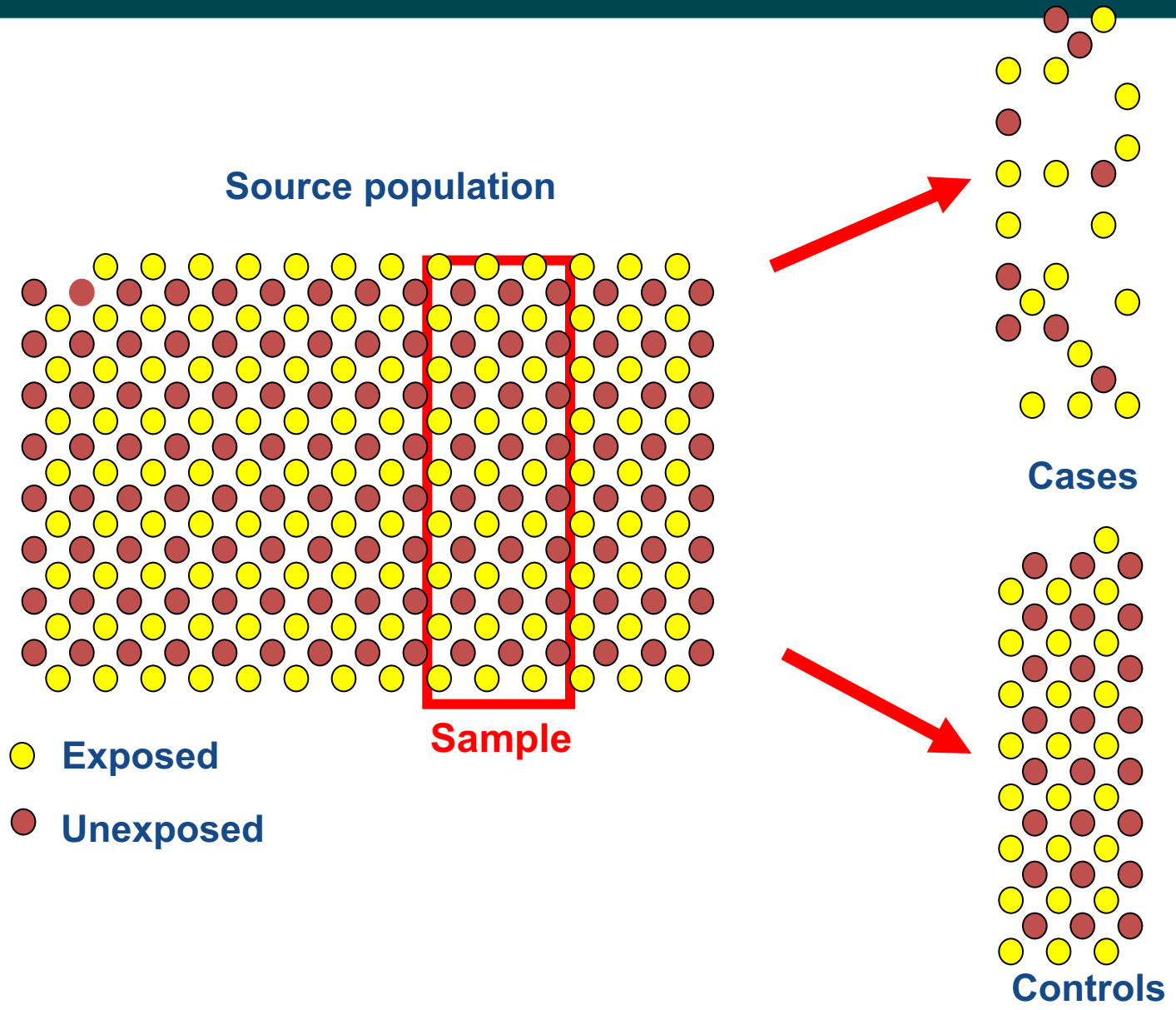
Case-control studies



Case-control studies



Case-control studies



Case-control studies

Prevalence of outcome in source population not known → Cannot calculate relative risk!

	Outcome (cases)	**No outcome (controls)**
Exposed	a	b
Unexposed	c	d

Odds in exposed (OE) a / b

Odds in unexposed (OU) c / d

Odds ratio (OR) OE / OU = ad / bc

Epidemiological studies

Cohort or case-control?

Case-control studies		Cohort studies	
↑	**Speed** Diseases with long latency periods can be investigated quickly.	↓	**Speed** Follow up can take years for diseases with long latency periods.
↑	**Rare outcomes** Easy to investigate precisely (because cases can be chosen).	↓	**Rare outcomes** Difficult to investigate precisely (because selected cohort may include few or none of them).
↓	**Rare exposures** Difficult to investigate precisely.	↑	**Rare exposures** Can be investigated precisely by selective sampling.
↓	**Temporality** Retrospective assessment of exposure, so does not establish temporality. Hence hard to prove causality.	↑	**Temporality** Prospective assessment of exposure, so establishes temporality. Hence better evidence for causation ("gold standard" in observational studies).
↓	**Population attributable risks** Difficult to estimate because risk of outcome in study may not be same as in population.	↑	**Population attributable risks** Can be estimated using risk of outcome in cohort.
↓	**Sources of selection bias** Poor matching of cases to controls.	↓	**Sources of selection bias** Poor choice of cohort, poor response rate, loss to follow up.
↓	**Sources of information bias** Recall bias, interviewer bias, non-random measurement error.	↑	**Sources of information bias** Less bias because exposure is measured prospectively, but can still happen due to non-random measurement error.

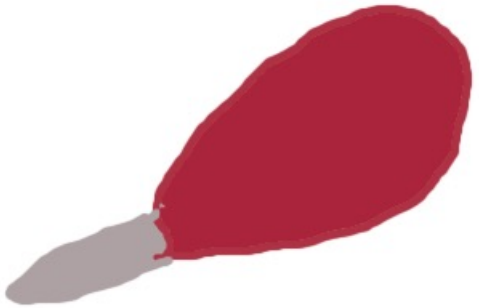
Exercise

A party ends tragically when several attendees have food poisoning. The local health protection team asks 50 party attendees what they ate, and they all answer all questions.

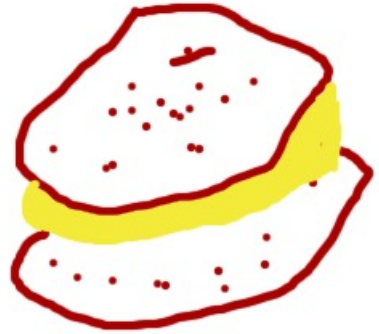
(i) What kind of study is this?

(ii) Based on the responses below, what is the most likely cause of the outbreak?

- 5/35 who ate **salad** were sick
- 7/15 who didn't eat **salad** were sick



- 4/20 who ate **chicken** were sick
- 4/30 who didn't eat **chicken** were sick



- 3/20 who ate **cheeseburger** were sick
- 6/30 who didn't **cheeseburger** were sick

Exercise

This is a case-control study since we start with people who got sick, and match them to people who didn't. (The 50 respondents may not be everyone who attended the party!)

Question	Total number not sick	Total number sick
Ate the salad	30	5
Did not eat the salad	8	7
Ate the chicken	16	4
Did not eat the chicken	26	4
Ate the cheeseburger	17	3
Did not eat the cheeseburger	24	6

Odds ratio of getting sick if ate

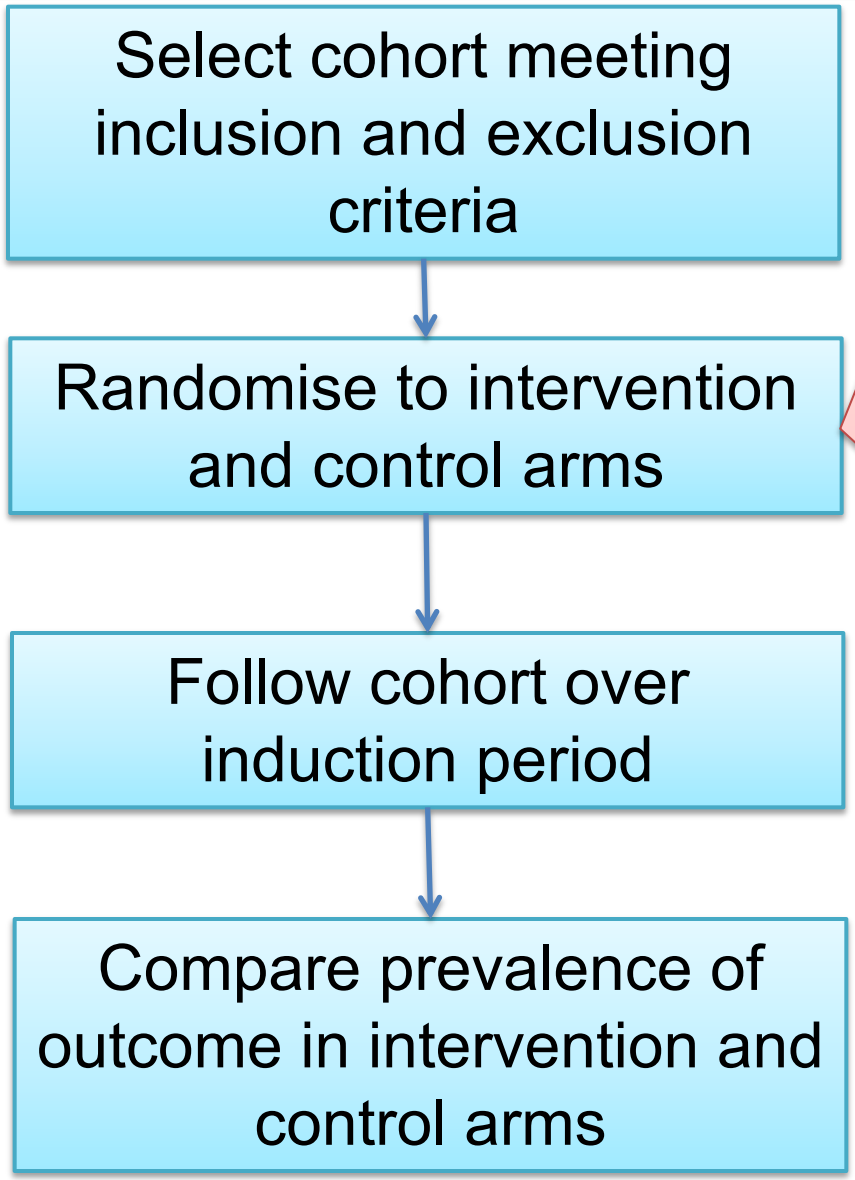
… the salad = (5/30) / (7/8) = 0.19

… the chicken = (4/16) / (4/26) = 1.6

… the cheeseburger = (3/17) / (6/24) = 0.71

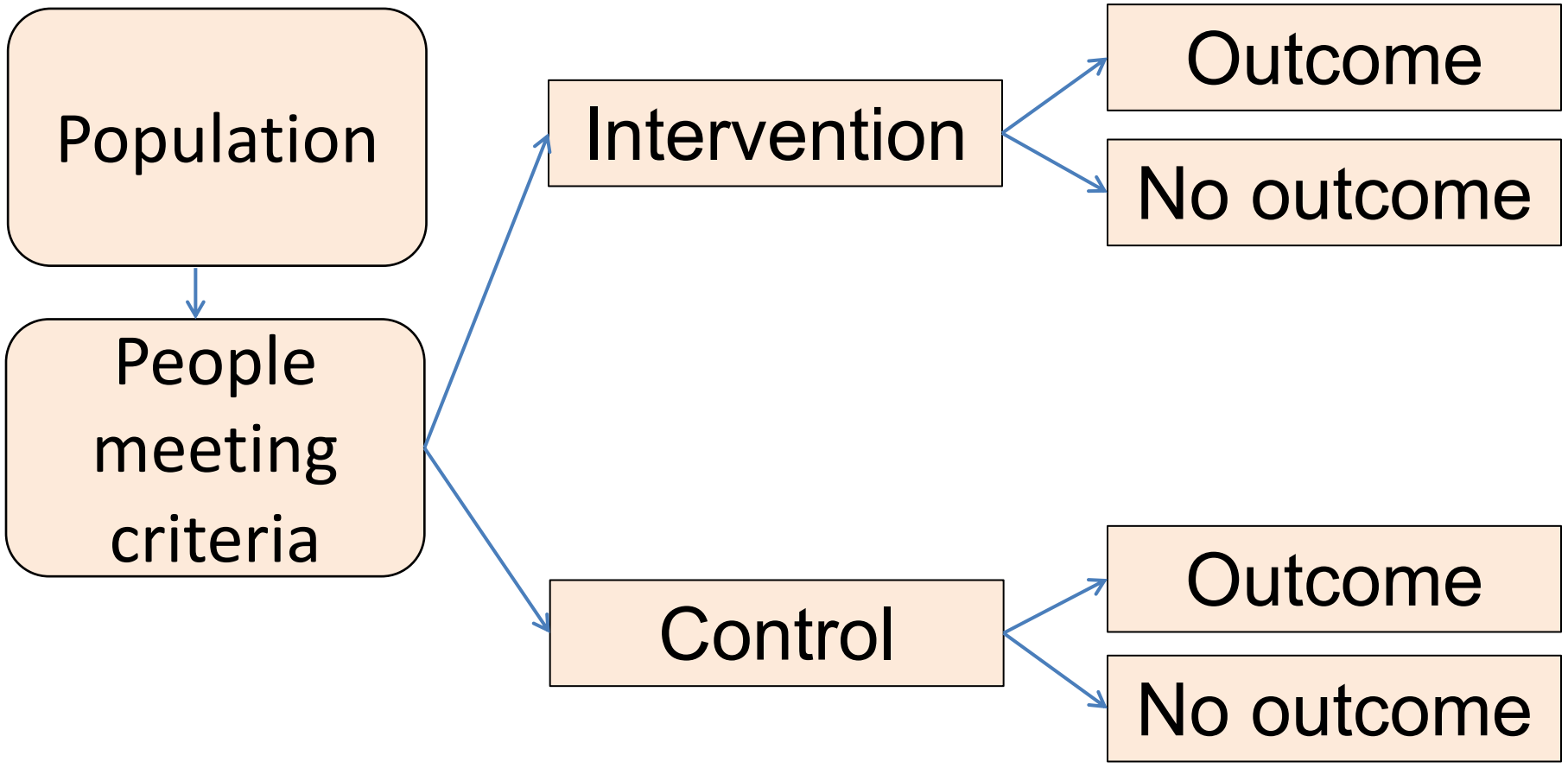
So the chicken is the **most likely** risk factor.

Experimental studies (trials)



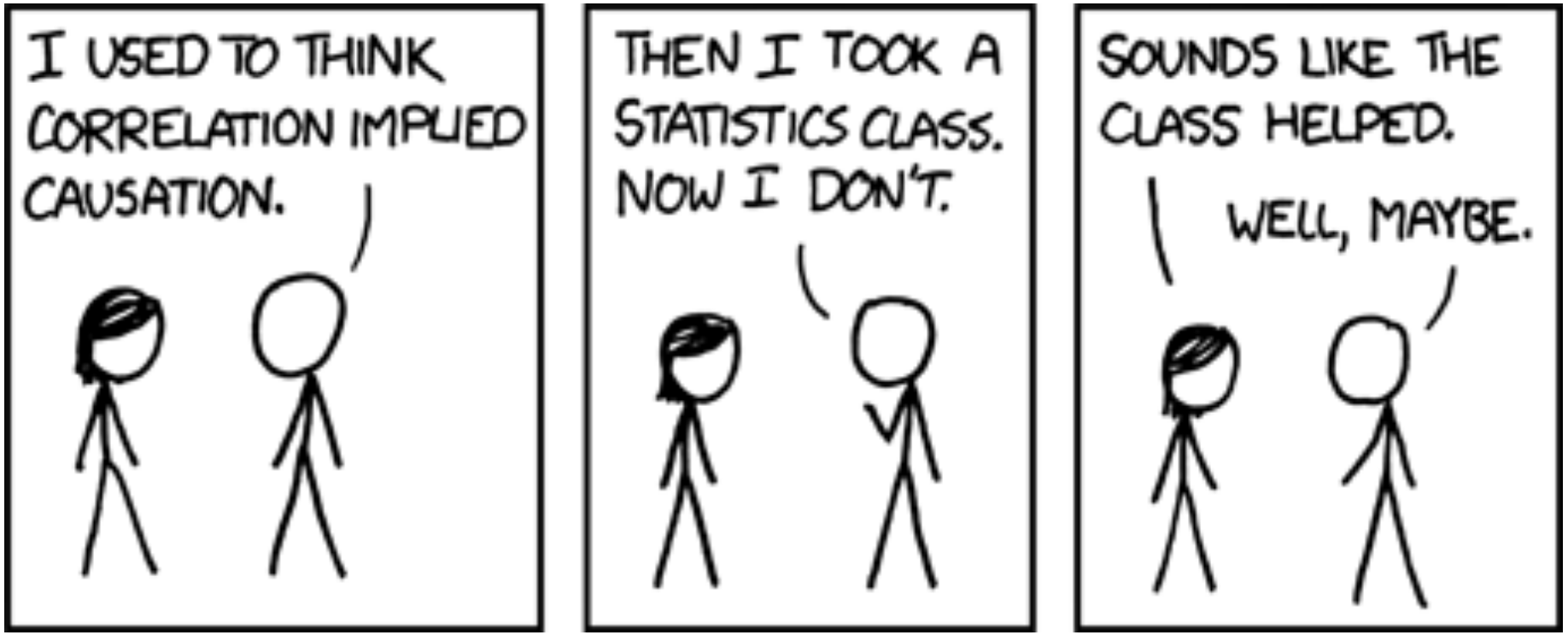
Key benefit: Allows investigator to control for ALL (known and unknown) risk factors in allocation.

Experimental studies (trials)



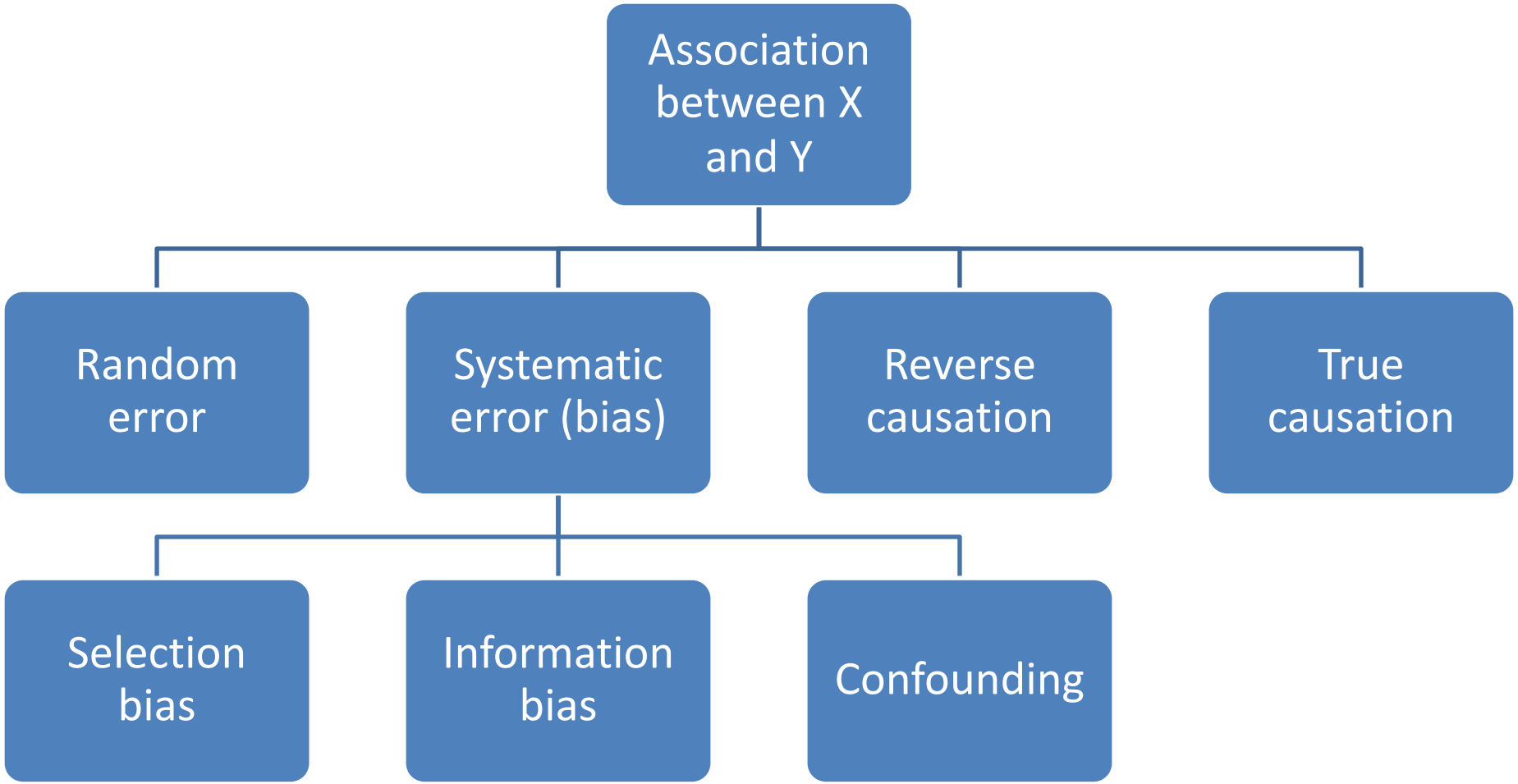
III.
Causation

Correlation ≠ causation



Source: xkcd (http://imgs.xkcd.com/comics/correlation.png)

Reasons for associations



What can you do to reduce each of these errors?

Mitigating errors in causal inference

Random error	• Increase sample size • Take more precise measurements
Bias	• Identify biases in study design
Confounding	• Randomise • Match cases and controls by confounding variables • Stratify by confounding variables • Multivariable regression
Reverse causation	• Use a prospective study design

The Bradford-Hill criteria

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

The Bradford Hill criteria

Hill (1965) *Proc R Soc Med* 58:295

Does the association between smoking and lung cancer imply causation?

Criteria	Meaning
Consistency	Several different studies with similar results – they are less likely to be all wrong.
Strength	Strong association – unlikely to be entirely due to random error, bias and confounding.
Temporality	Exposure must precede outcome in time.
Dose-response	Increasing exposure increases risk and/or intensity of outcome.
Specificity	One cause for one effect.
Coherency	Consistent with existing knowledge of disease.
Plausibility	Can be explained by a mechanistic biological model.
Analogy	Similar to other known associations.
Experimental evidence	Association shown in a randomised controlled trial in humans or animals.