Epidemiology and basic study design

Beau B. Bruce, MD, PhD

Disclosures

The statements contained in this talk are solely those of the author and do not necessarily reflect the views or policies of the Centers for Disease Control and Prevention.

No financial or other conflicts of interest relevant to this talk.

If you dig a hole…

… and it's in the wrong place, digging it deeper isn't going to help.

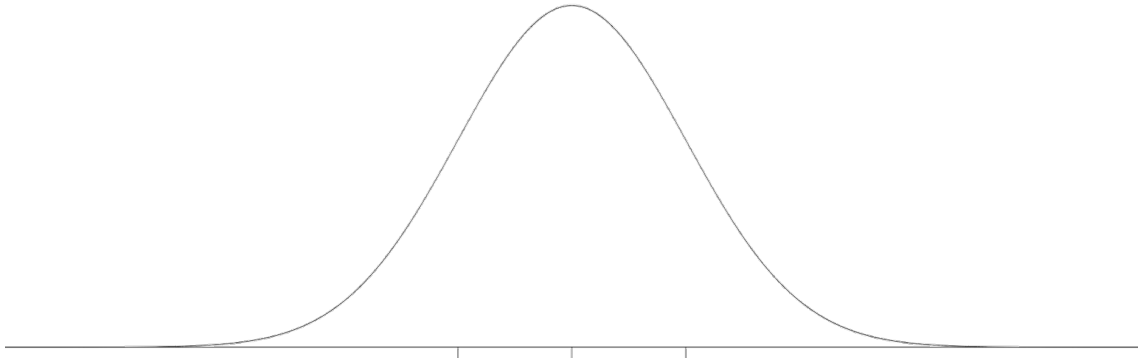
– Seymour Chwast

What is epidemiology?

the basic science of
public health and
clinical medicine

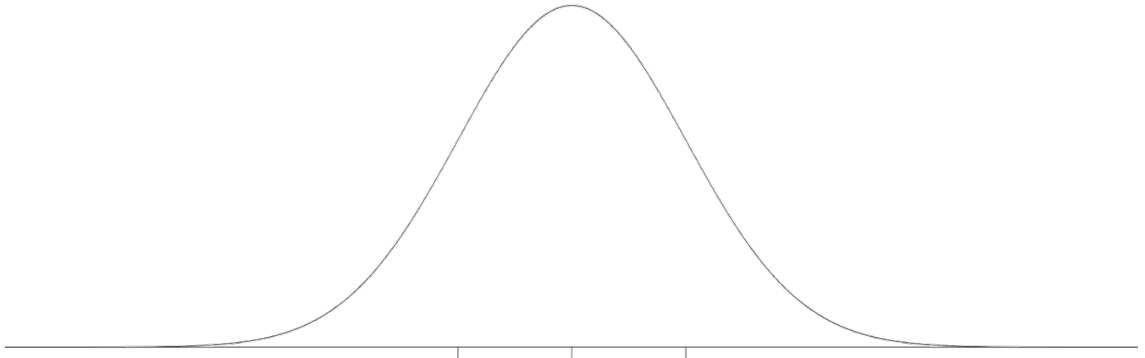
What is epidemiology?

- The science of the “3 D’s” which are the:
  - Determinants of
  - Disease and its
  - Distribution



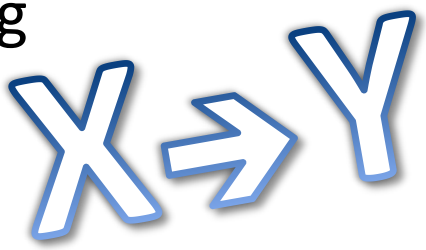
What is epidemiology?

- Many basic and clinical scientists study the *determinants* of *disease*
- The study of *distribution*
  - a distinctive feature of epidemiology
  - explains the close link between epidemiology and statistics



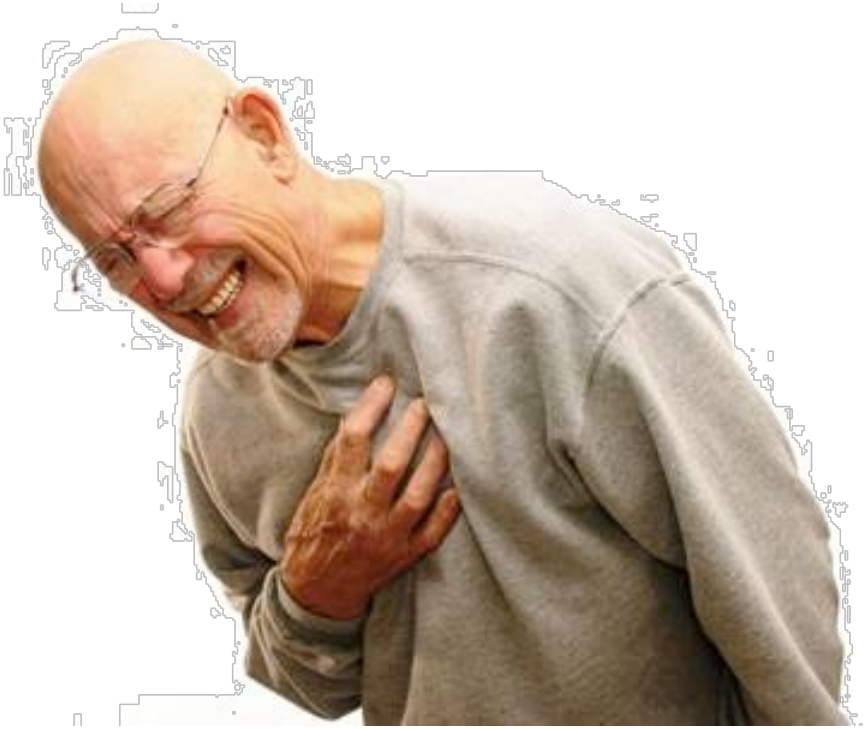
But *determinants* are very important

- Implies that we are studying *causality*
  - We care about causes because when we know the cause of disease we can treat the underlying mechanism
- Science lives in the *covariance*
  - How two or more things change with each other
  - The association between two or more factors
  - The interaction between three or more factors
  - Science is rarely merely about describing something – i.e., it is NOT about "what" but about "HOW"



What is the goal of research?

- Mr. Jones is a 68-year-old smoker weighing 175 pounds who just had a heart attack



What is the goal of research?

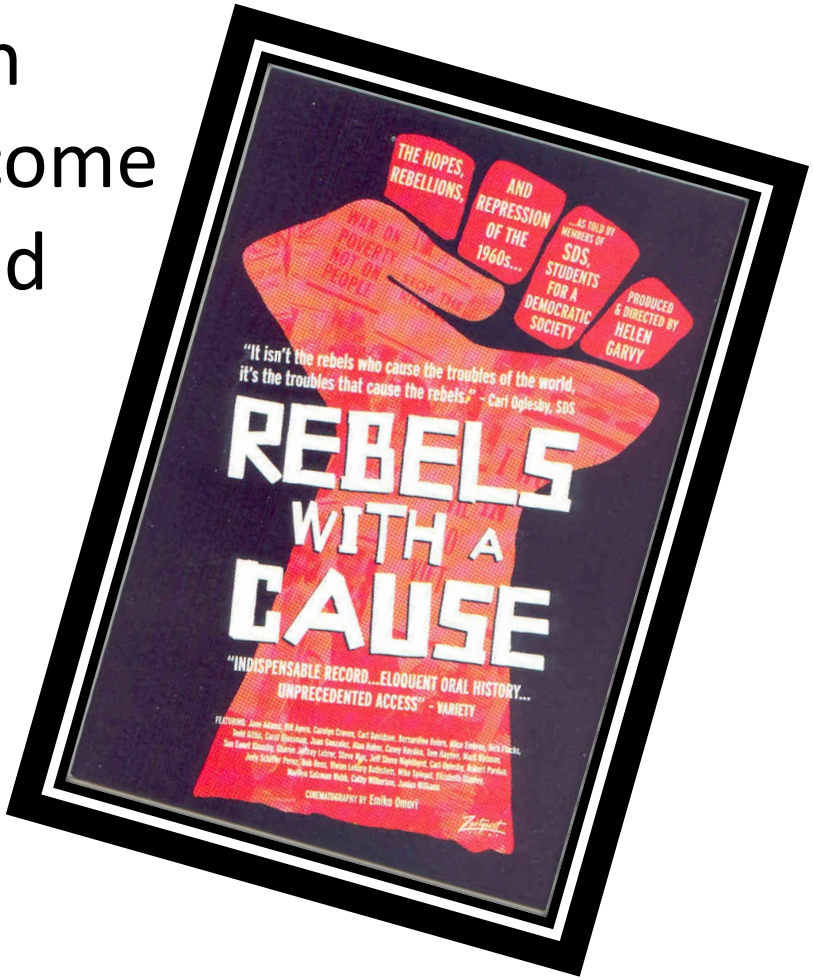
- Mr. Jones is a 68-year-old smoker weighing 175 pounds who just had a heart attack

  - What is the probability that Mr. Jones had a heart attack?
  - What is the probability that a person "similar" (*distribution*) to Mr. Jones will have a heart attack within a certain period of time?

How “Similar”?

What is a cause (*determinant*)?

There is at least one person who experiences an outcome after exposure who would not have experienced the outcome had they been unexposed.



The counterfactual *ideal*

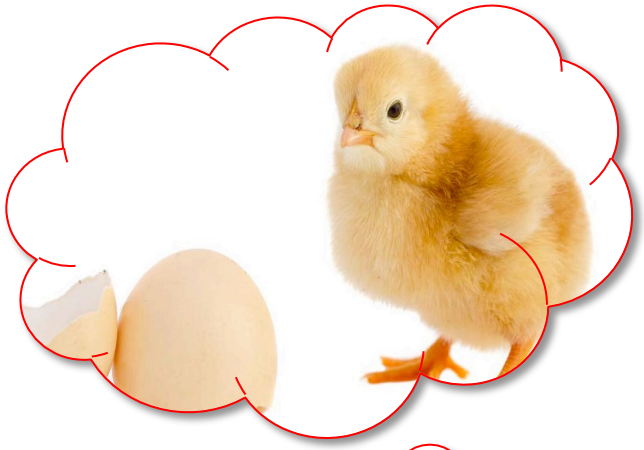
How can we prove that smoking was the cause of Mr. Jones' heart attack?

We could only answer that if we could go back in time *but in his "counterfactual life"* he relives his life as a non-smoker.

That is *unobservable* but is our goal in defining the ideal "similar" group (i.e., comparison or control group) for our study

You observe an association between A and B (*covariance*)

1. A causes B



2. B causes A

3. There was a bias

4. It occurred by chance

Back to digging a hole…

Having a million patients in your study may save you from chance associations (statistics),

Back to digging a hole…

…but it will not

protect you

from the bias

resulting from

poor study design!

(epidemiology)

If you find yourself in a hole…

…the first thing
to do
is stop digging.

– Will Rogers

Bias = Systematic Error

- Random error gets smaller and smaller as you increase sample size
- Systematic error (bias) does not.
- Three broad, but comprehensive sources of bias found in studies:
  - Information bias
  - Selection bias
  - Confounding

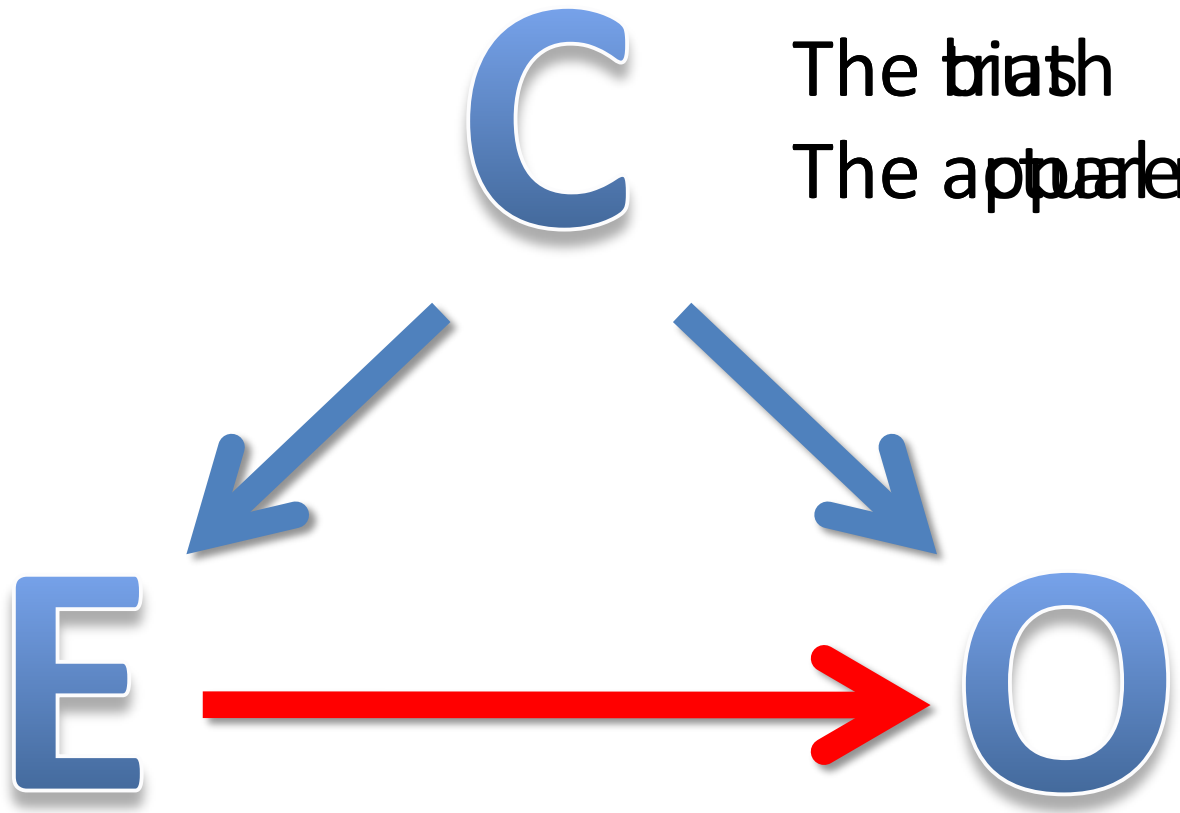
Biases

- Information bias = mismeasurement or misclassification of exposure or disease
  - e.g., use of an imperfect diagnostic test to define disease
- Selection bias = incorrect choice or incorrect sampling from the study source population
  - e.g., use of volunteer study subjects

Confounding

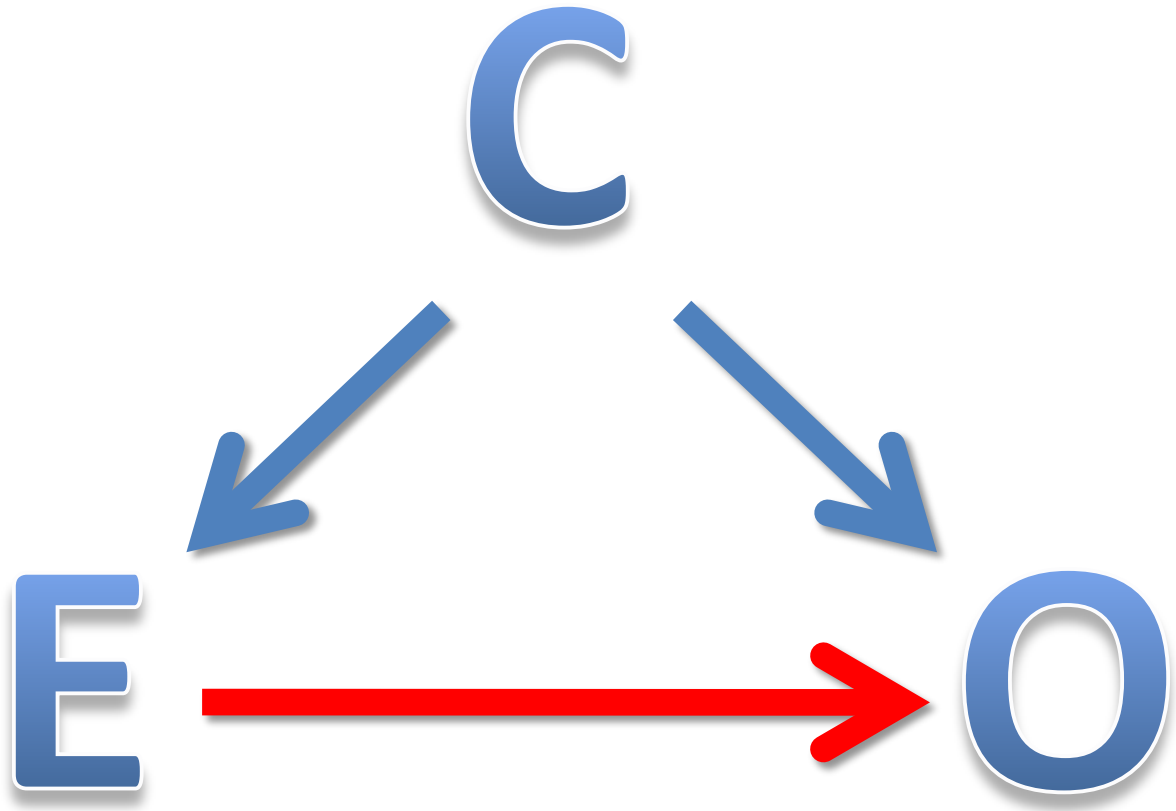
- Easy to confuse with selection bias, but confounding results from a different causal mechanism
- It is the mixing of the effects of a third factor with those of the exposure and outcome so as to distort the association of interest (typically between the exposure and outcome)
- We “control for” confounders

Confounding

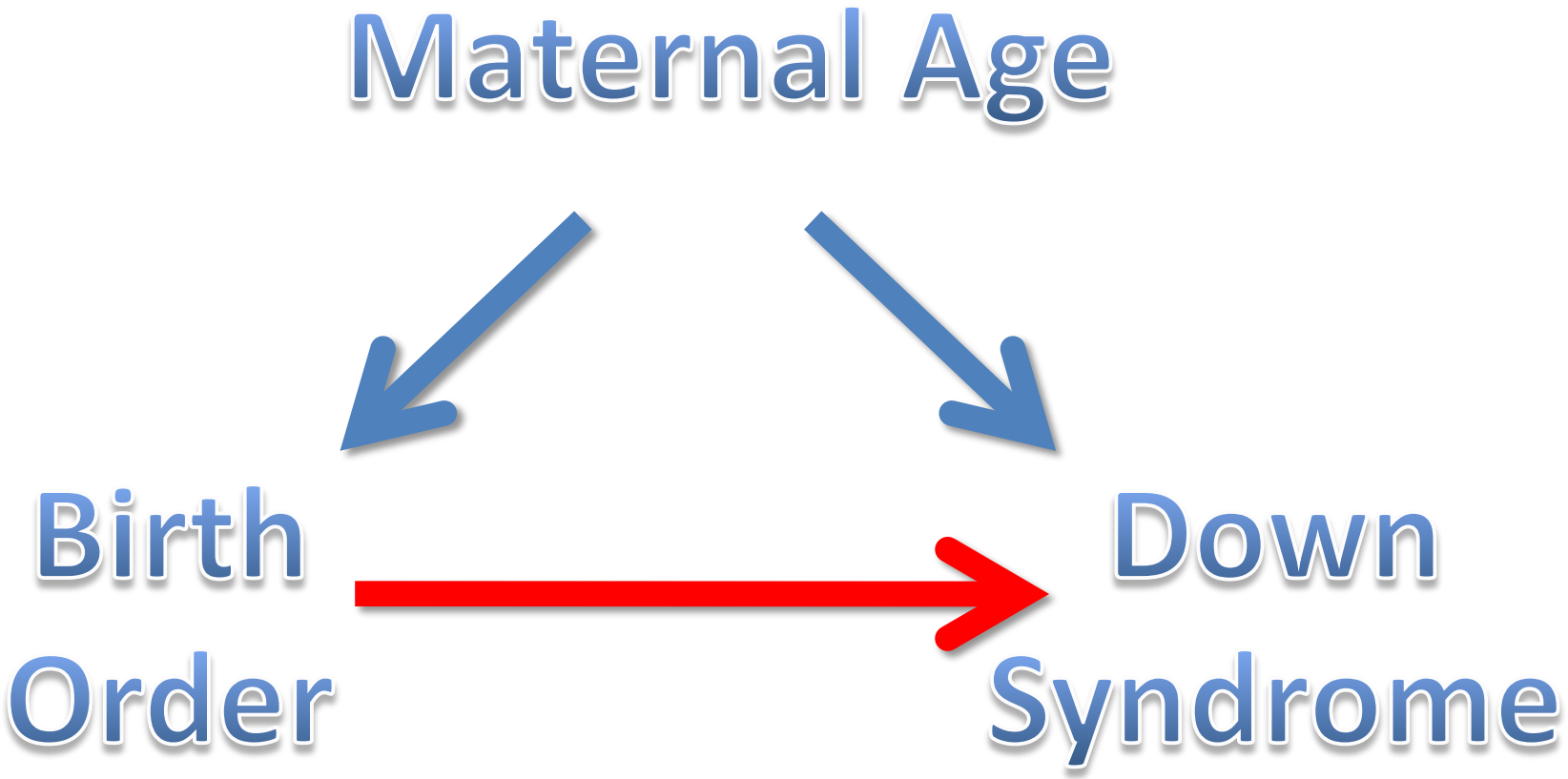


The
The
The

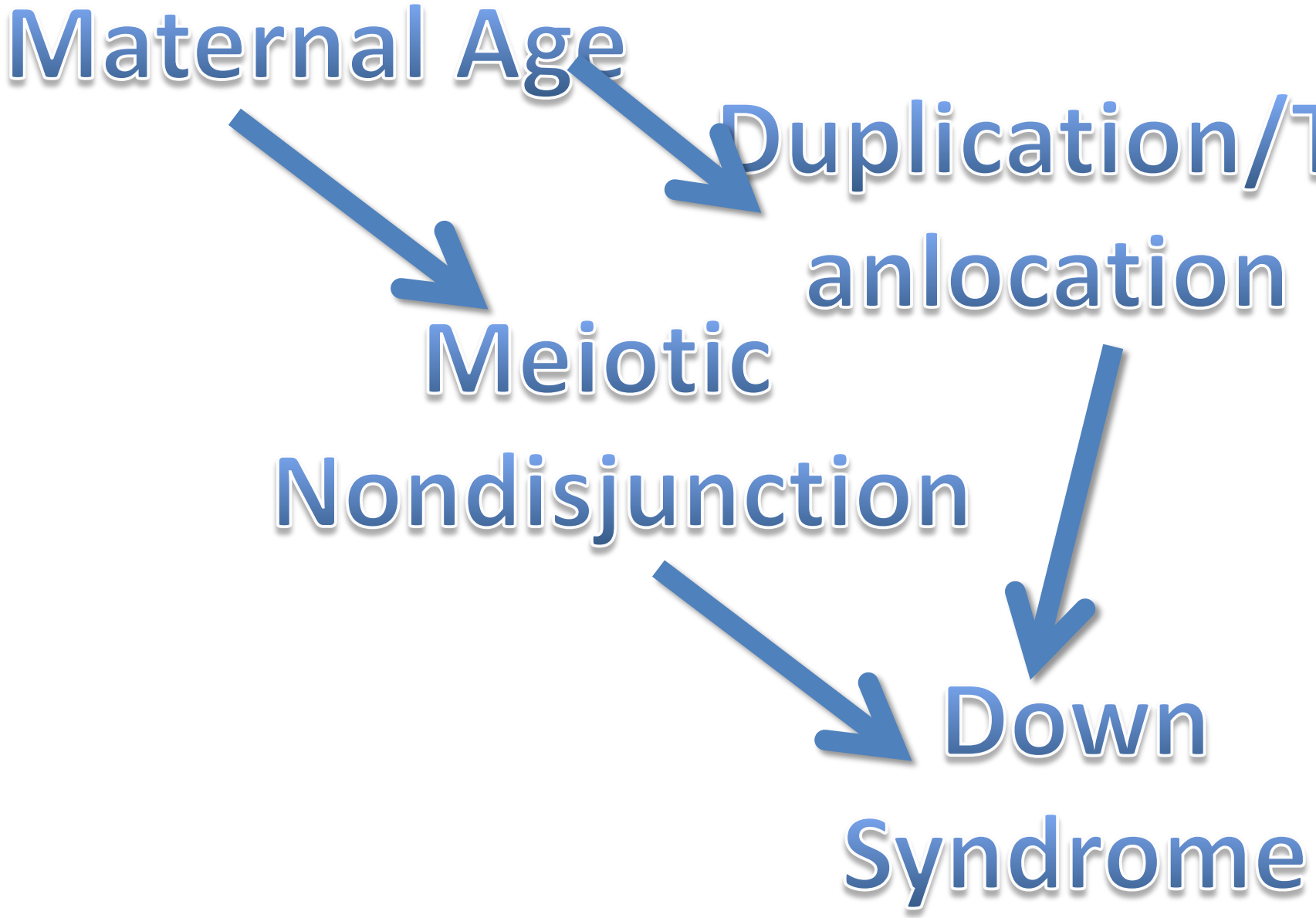
Confounding



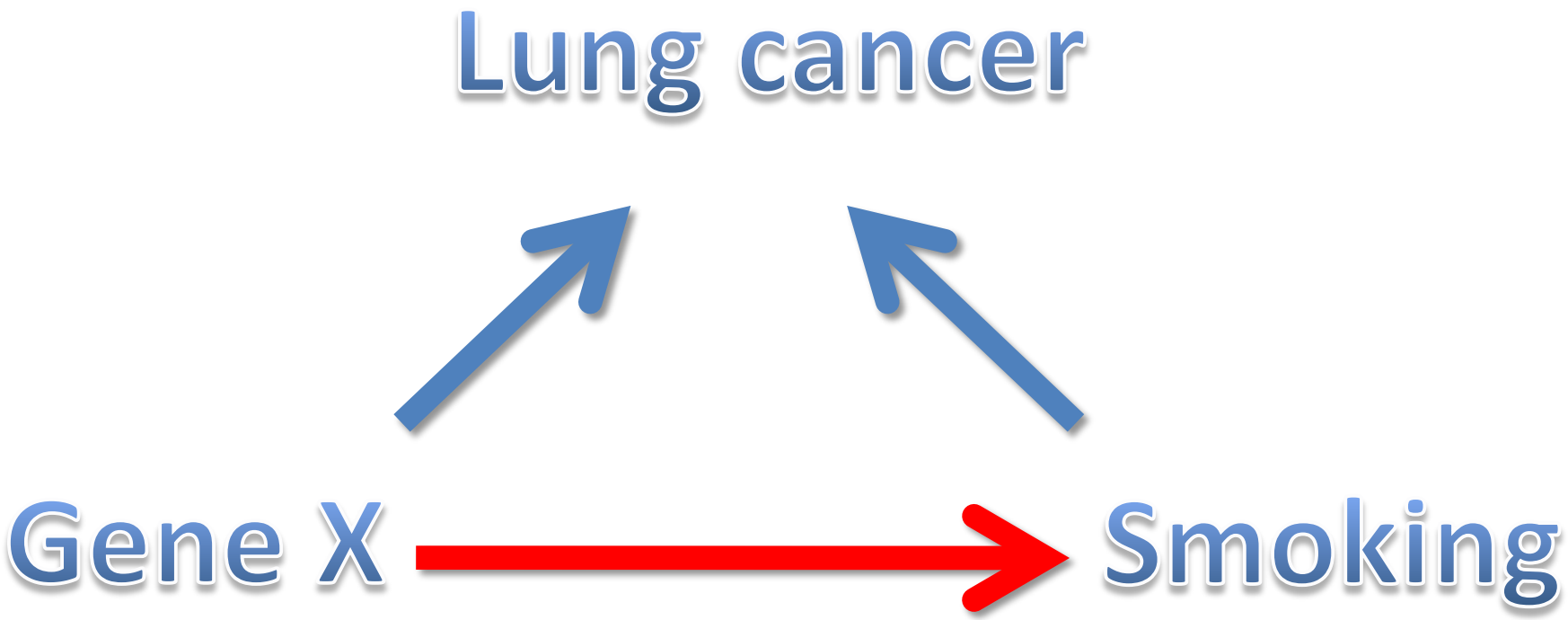
Confounding



Intermediate on causal pathway



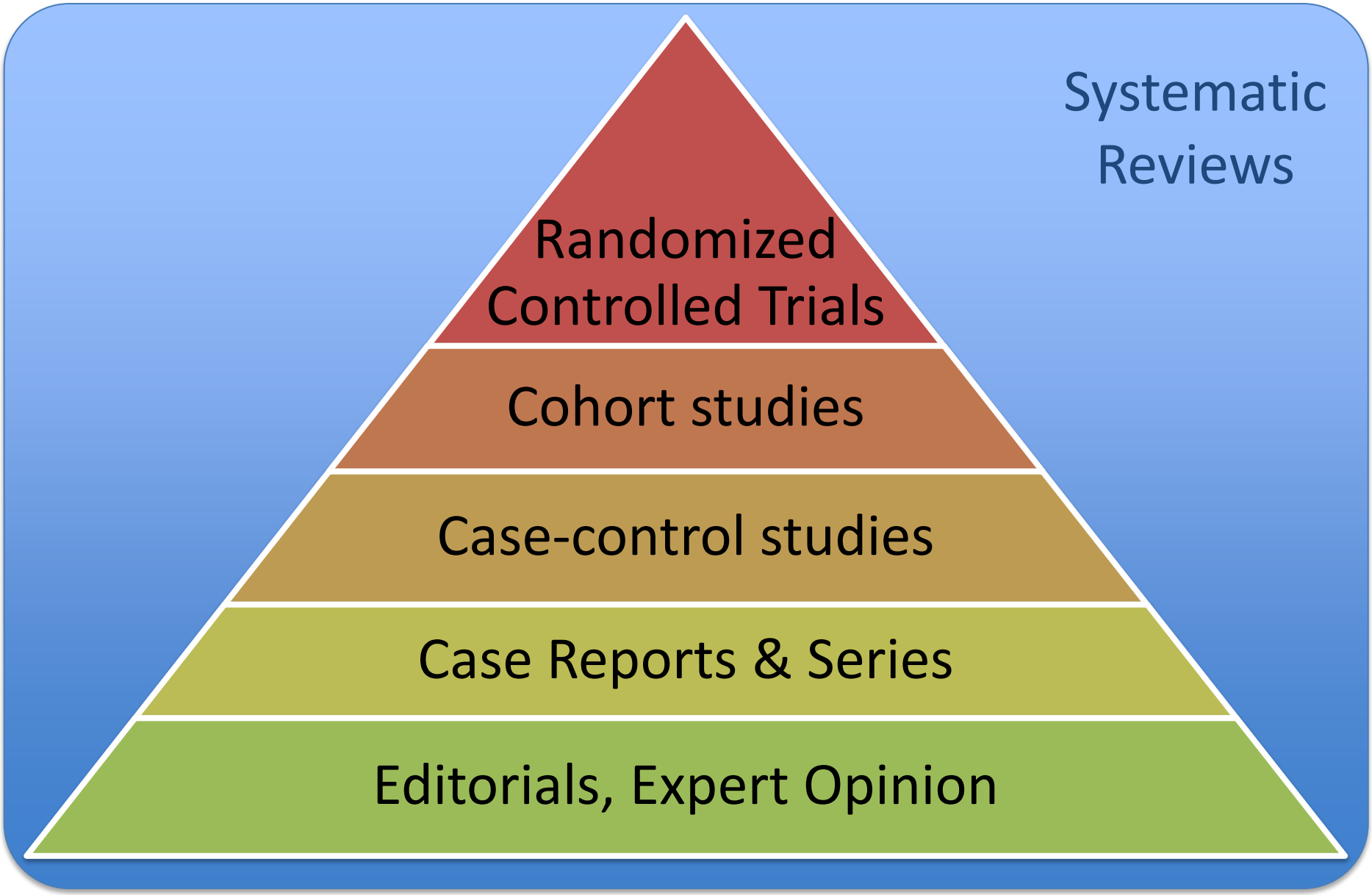
Selection bias: Conditioning on common effect



What to control?

- Control means to keep constant
- Helps us with the counterfactual ideal – everything but the exposure should stay the same (remain constant)
- Control for confounders (common causes of exposure and outcome) but NOT intermediates (variables that mediate the effect of the exposure on outcome) or common effects of exposure and outcome (selection bias)

Levels of Evidence



Levels of Evidence

- Class I – Must Do
- Class II – Should Do
- Class III – Can Do
- Class IV – Voodoo
- Class V – Doodoo

Andrew Lee, MD, Chair of Ophthalmology, The Methodist Hospital, Houston, TX

The big three study designs

- Randomized controlled trial
- Cohort study
- Case control study

Randomized controlled trials

"Life is simple,

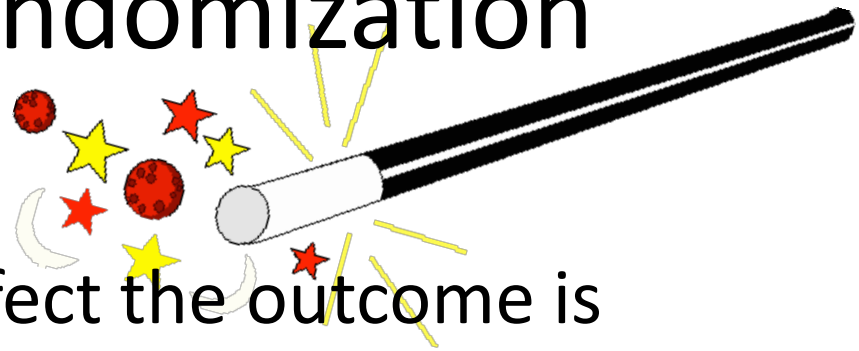
it's just not easy."

-Anon

Randomized controlled trials

That is to say – if you get the randomization right, maintain blinding, and you're lucky you'll get the right answer!

The magic of randomization



- Every factor that might affect the outcome is balanced between the treatment and control groups
- It not only balances the things we know and can control for but also all the things we don't know and thus can't control for (i.e., unmeasured confounders)
- Therefore, any difference in outcome is due to the treatment, i.e., causality is demonstrated!!!

But the magic can be lost…

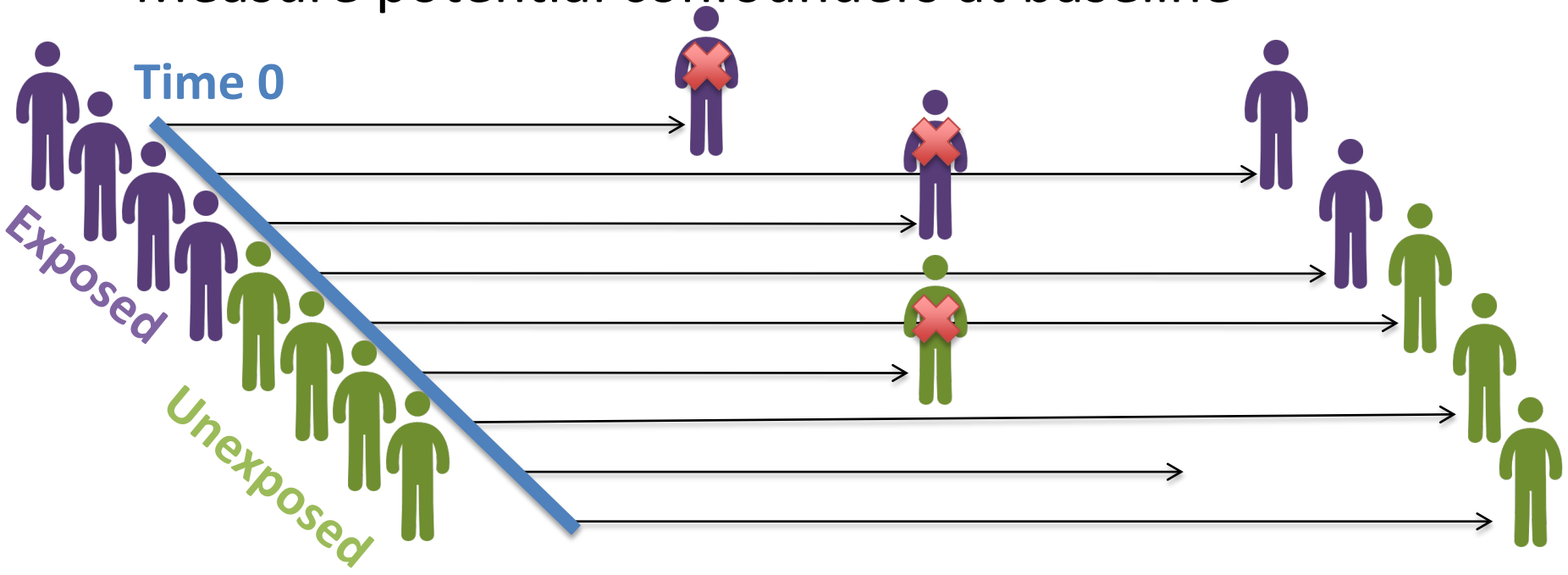
- Failure to blind researchers/subjects
- Failure to follow intention to treat analysis (must keep subjects in randomly assigned categories or you have "broken" the randomization)
- Failure to be lucky

Other issues

- Strict inclusion criteria often limit generalizations
- Typically very expensive
- Difficult for rare diseases
- Bias can occur with differential dropout related to exposure / treatment (selection bias)

Cohort studies

- Take a population of outcome free subjects at time zero with an given exposure distribution and follow for a period of time
- Measure outcome as it happens, compare exposed vs. unexposed
- Measure potential confounders at baseline



Strengths

- Usually the exposure is measured before the disease occurs, so causality is relatively easier to substantiate

- Less subject to recall biases when performed prospectively

Weaknesses

- Still very difficult for rare diseases (have to have lots of people or wait a long time)

- Subject to additional biases:
  - Selection bias in picking the exp/unexp groups
  - Incorrect assumptions about timing of exposure/outcome (particularly a problem with chronic diseases: e.g., when does diabetes start?)
  - Unmeasured/unknown confounders

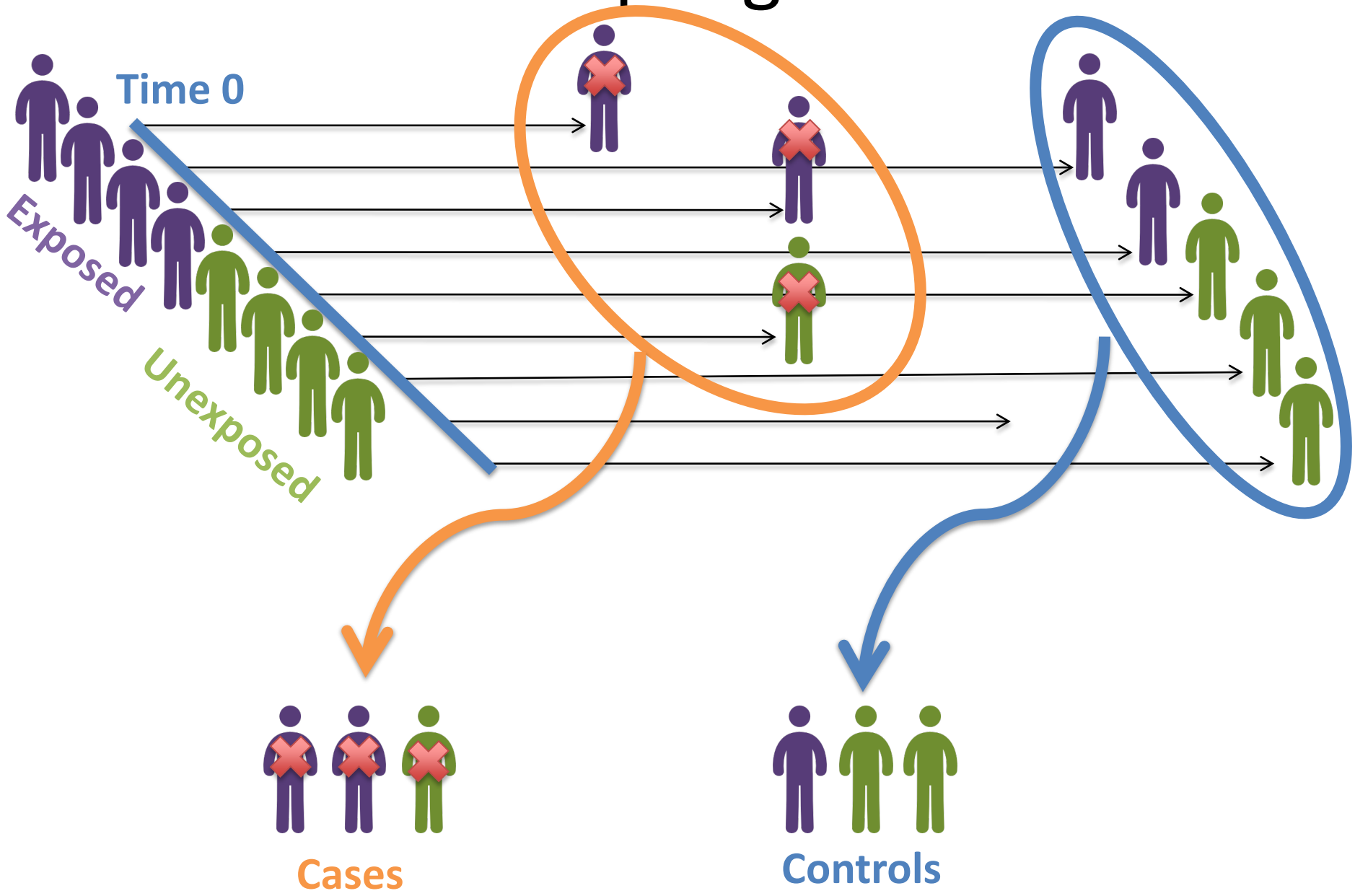
How can you prevent (control for) confounding by design?

- Randomization (via an RCT, only way to deal with unknown confounders)
- Restriction
  - Only enroll patients similar with regards to the potential confounder (i.e., all the same age)
- Matching
  - Choose unexposed subjects to match either the proportion of individuals or individuals themselves with the cohort on the potentially confounding characteristic

Case-control studies

- Disease has already occurred (the cases) and we want to compare them to an appropriate group of controls
- A double sampling from an “underlying cohort”
  1. the cases
  2. the controls
- Perhaps the most important methodological advance of epidemiology

Double sampling of cohort



Advantages

- Can be applied to rare diseases
  (and luckily most diseases are indeed rare!)

- Need relatively few subjects even in the control population – e.g. suppose there are 100 cases of disease in the Atlanta metro area with a population about 5.5 million people – in a cohort would need to collect info on all them!

Disadvantages

- Same as cohort study

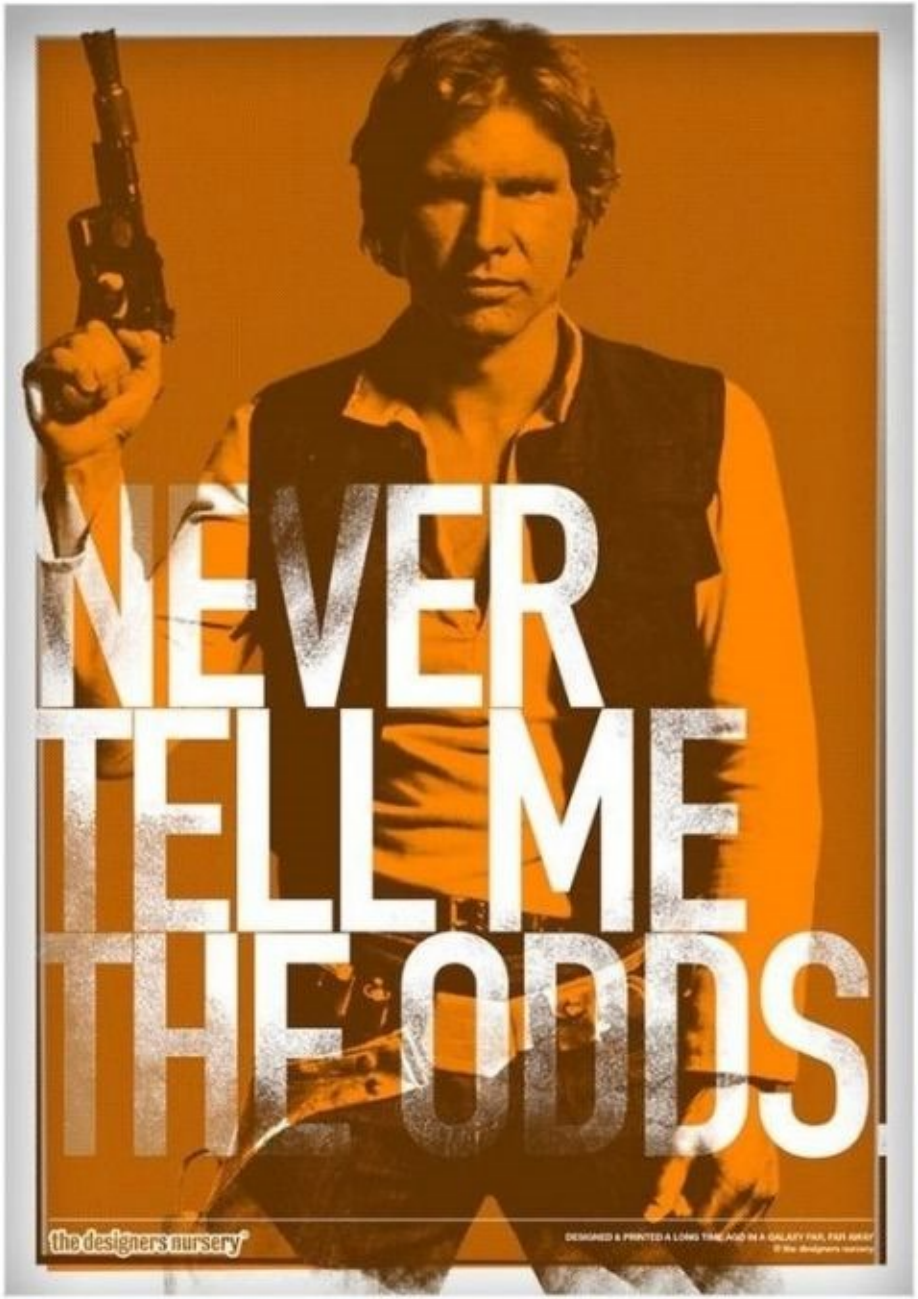
- AND cannot estimate the underlying risk or odds of disease based on exposure

  (e.g., what are the chances I'll get lung cancer if I smoke?)

Disadvantages:
cannot estimate underlying risk

	Exposed	**Unexposed**	
Case	75	25	100
Control	20	80	100
	95	105	200

Is the odds of exposure among the cases = 75/25 = 3?


NEVER
TELL ME
THE ODDS
the designers nursery

What is odds?

- Odds is the number with X divided by the number without X

- Directly maps to probability:

Odds = Prob/1-Prob

Probability range: 0 to 1

Odds range: 0 to infinity

Disadvantages:
cannot estimate underlying risk

	Exposed	**Unexposed**	
Case	75	25	100
Control	20	80	100
	95	105	200

Is the odds of exposure among the cases = 75/25 = 3?

Disadvantages:
cannot estimate underlying risk

	Exposed	**Unexposed**	
Case	75	25	100
Control	20	80	100
	95	105	200

Is the odds of exposure among the cases = 75/25 = 3? YES!

Disadvantages:
cannot estimate underlying risk

	Exposed	Unexposed	
Case	75	25	100
Control	20	80	100
	95	105	200

Is the odds of exposure among the controls = 20/80 = 0.25?

Disadvantages:
cannot estimate underlying risk

	Exposed	Unexposed	
Case	75	25	100
Control	20	80	100
	95	105	200

Is the odds of exposure among the controls = 20/80 = 0.25? YES!

Disadvantages:
cannot estimate underlying risk

	Exposed	**Unexposed**	
Case	75	25	100
Control	40	160	200
	115	185	300

Is the odds of exposure among the controls = 40/160 = 0.25? YES!

Disadvantages:
cannot estimate underlying risk

	Exposed	**Unexposed**	
Case	75	25	100
Control	20	80	100
	95	105	200

Is the odds ratio of exposure between the cases and controls (EOR) = 3/0.25 = 12?

Disadvantages:
cannot estimate underlying risk

	Exposed	**Unexposed**	
Case	75	25	100
Control	20	80	100
	95	105	200

Is the odds ratio of exposure between the cases and controls (EOR) = 3/0.25 = 12? YES!

Disadvantages:
cannot estimate underlying risk

	Exposed	**Unexposed**	
Case	75	25	100
Control	20	80	100
	95	105	200

That's nice, but that was the answer to the question, "what are the chances I smoke if I have lung cancer?" (who cares?)

Disadvantages:
cannot estimate underlying risk

	Exposed	**Unexposed**	
Case	75	25	100
Control	20	80	100
	95	105	200

What we want to know is the odds of disease given exposure, i.e., the chances of lung cancer given I smoke.

Disadvantages:
cannot estimate underlying risk

	Exposed	**Unexposed**	
Case	75	25	100
Control	20	80	100
	95	105	200

Is the odds of disease given exposure 75/20 = 3.75?

Disadvantages:
cannot estimate underlying risk

	Exposed	**Unexposed**	
Case	75	25	100
Control	20	80	100
	95	105	200

Is the odds of disease given exposure
75/20 = 3.75? NO!

Disadvantages:
cannot estimate underlying risk

	Exposed	Unexposed	
Case	75	25	100
Control	20	80	100
	95	105	200

Is the odds of disease given non-exposure 25/80 = 0.3125?

Disadvantages:
cannot estimate underlying risk

	Exposed	**Unexposed**	
Case	75	25	100
Control	20	80	100
	95	105	200

Is the odds of disease given non-exposure
25/80 = 0.3125? NO!

Disadvantages:
cannot estimate underlying risk

	Exposed	Unexposed	
Case	75	25	100
Control	40	160	200
	115	185	300

Is the odds of disease given non-exposure 25/160 = 0.15625? NO! (0.3125≠0.15625!)

Disadvantages:
cannot estimate underlying risk

	Exposed	Unexposed	
Case	75	25	100
Control	20	80	100
	95	105	200

Is the odds ratio of disease between exposed and unexposed (DOR) = 3.75/0.3125 = 12?

Disadvantages:
cannot estimate underlying risk

	Exposed	Unexposed	
Case	75	25	100
Control	40	160	200
	115	185	300

Is the odds ratio of disease between exposed and unexposed (DOR) = 1.875/0.15625 = 12?

Bayes Theorem

- The application of Bayes theorem by those mathematically inclined epidemiologists: $P(B|A) = (P(A|B)P(B))/P(A)$ says that indeed:

  Exposure odds ratio = Disease odds ratio

  !!!

Disadvantages:
cannot estimate underlying risk

	Exposed	Unexposed	
Case	75	25	100
Control	20	80	100
	95	105	200

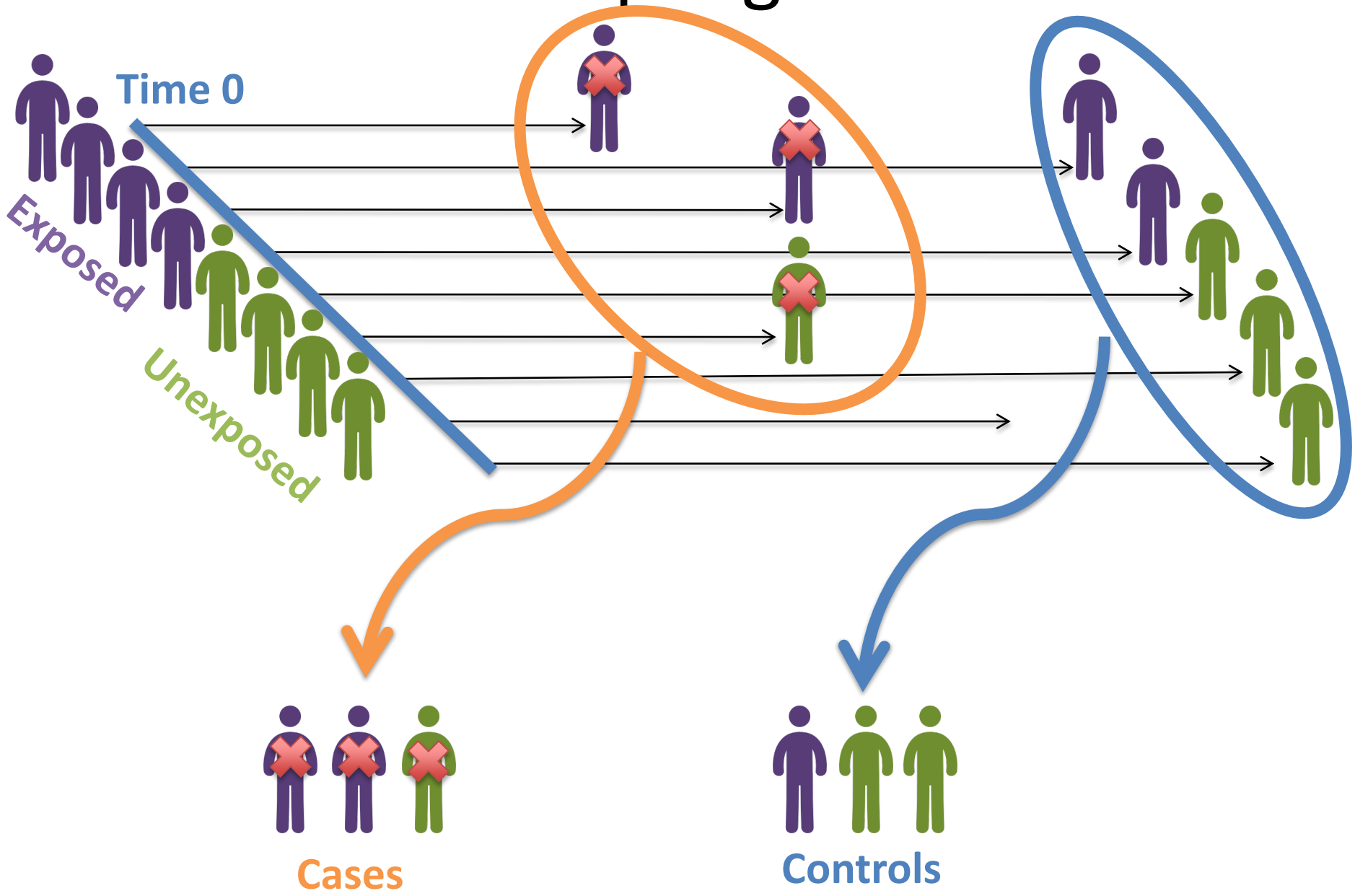
Is the odds ratio of disease between cases and controls (DOR) = 3.75/0.3125 = 12? YES!

Magic of the case-control study

- While you cannot estimate the underlying risk (odds) of disease in each group, you can still compare the two groups (assess the association)

- The control group is used to compare the *distribution of exposure* to the cases

Double sampling of cohort

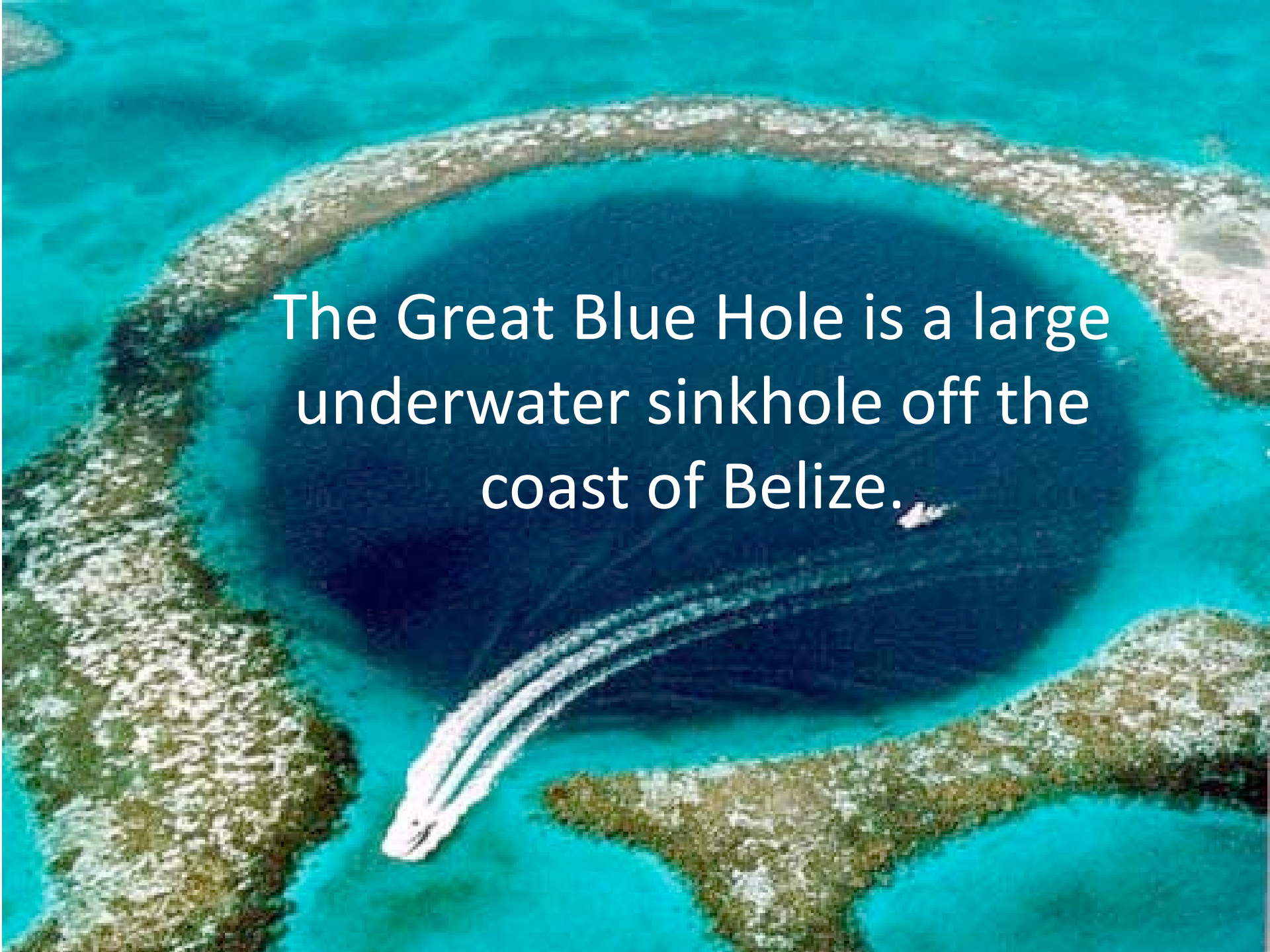


Is there a bias?

- You plan to undertake a case-control study of ovarian cancer and blood type O? Cases are a random sample of women with ovarian cancer. Assume that type O blood is not related to gender. Can you use men for the controls?

Sooooo… Think!

- The most important machine for proper design and performance of a study is between your ears.
- A study and stats program will always give an answer, but if you ignore epidemiologic principles it will likely only give you a wrong one!


The Great Blue Hole is a large underwater sinkhole off the coast of Belize.