STAT24630

Jingshu Wang

Causal Inference Methods and Case Studies

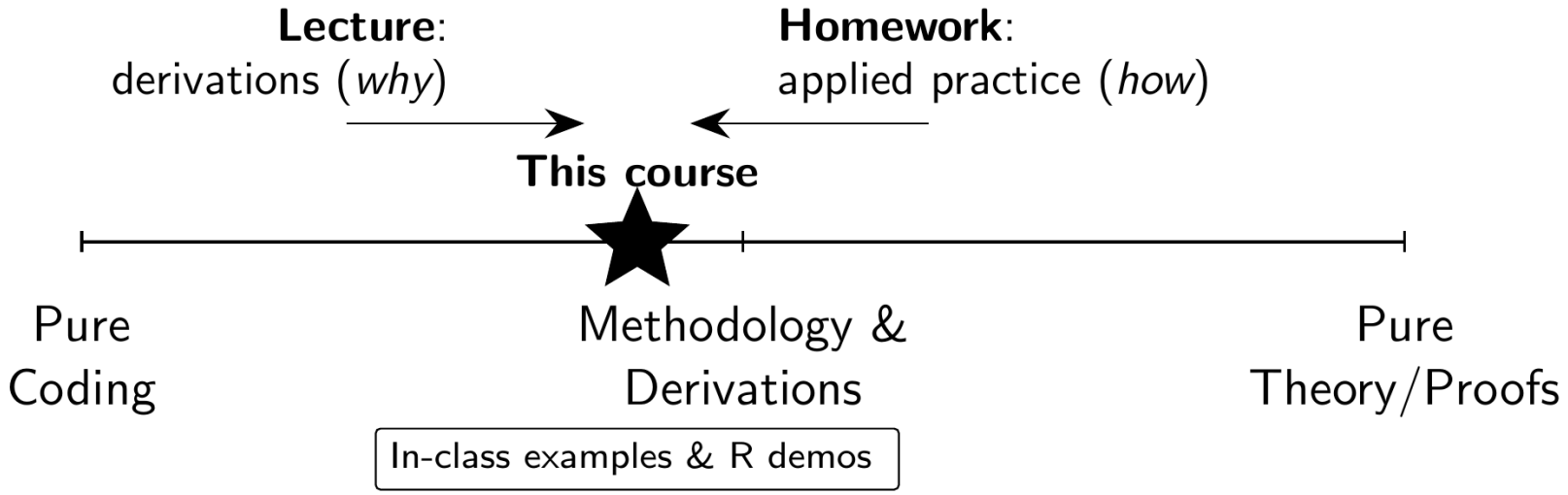
Lecture 1
Introduction with four examples

Outline

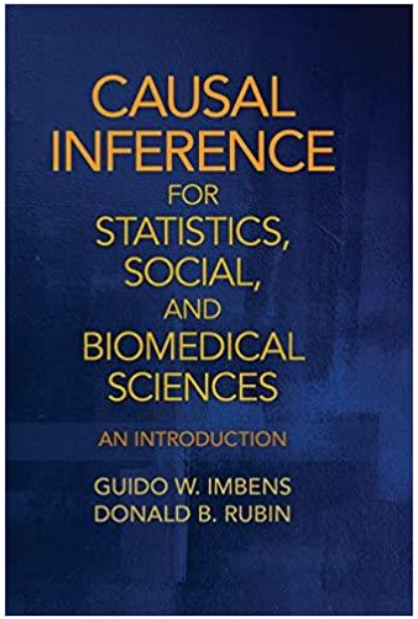
- Course logistics
- Introduction with four examples
  - Would high HDL cholesterol level be protective against heart disease?
  - Does maternal smoking have a beneficial effect to reduce infant mortality?
  - Phase 3 randomized trial for the COVID-19 vaccine
  - Effect of compulsory school attendance on schooling and earnings

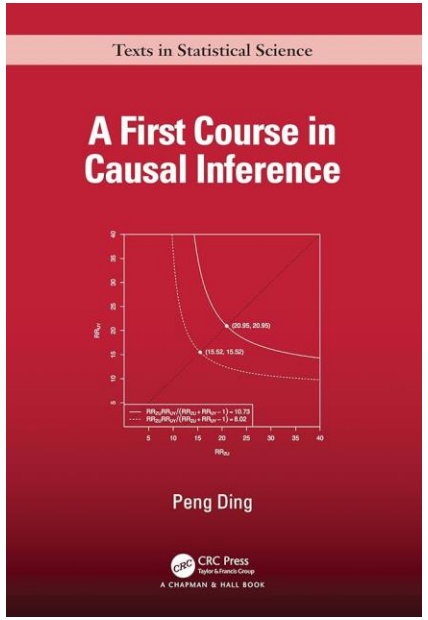
What to expect from this course

- Focus: *causal inference methodology and assumptions*
- We use **R regularly** for applications: but this is not an R programming class (R basics expected)
- Lectures include **math derivations**
  - not full proofs, but enough to understand *why* methods work
- Homework is **more applied**, emphasizing practice with real data and coding
- Goal: by the end, you should be able to
  - Frame a causal question
  - Choose and apply a method in R
  - Defend the assumptions in plain language



Textbook





- *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* by Guido W. Imbens and Donald B. Rubin
  - E-source freely accessible from the UChicago account: *https://doi-org.proxy.uchicago.edu/10.1017/CBO9781139025751*
- Some lectures will also be based on *A first Course in Causal Inference* by Peng Ding
  - E-source available at https://arxiv.org/abs/2305.18793.

Assignments and Exams

- Four homework assignments (40%): submit at the end of week 2/4/6/8
  - HW1 will be due on 10/12 11:59pm
- Homework are submitted via Gradescope

- Two online quizzes (15%) 10/24 and 11/21
  (40 minutes each with flexible time window on the quiz day)

- Final (45%):
  - Final group course project (35%): report + presentation
    - Group size: 3-4, presentation time: last week 10-12 minutes per group
    - Dataset and causal question finalized by week 6
  - Final individual self-assessment (10%)
    - 1-2 pages, 3 questions
  - Please read the syllabus carefully for instructions

Course participation

- This course is more engaging when everyone contributes.

- Participation means:
  - Asking questions in class (no question is too basic!)
  - Sharing thoughts during short pair/group activities
  - Offering examples or clarifications during case studies
  - Helping peers on Piazza discussions

- You are not graded directly on speaking up every day, but consistent contributions (in class or online) will be considered when making rounding decisions for students close to a grade boundary.

- Top contributors on Piazza will also receive small bonus credit (up to 5 points).

Course policies

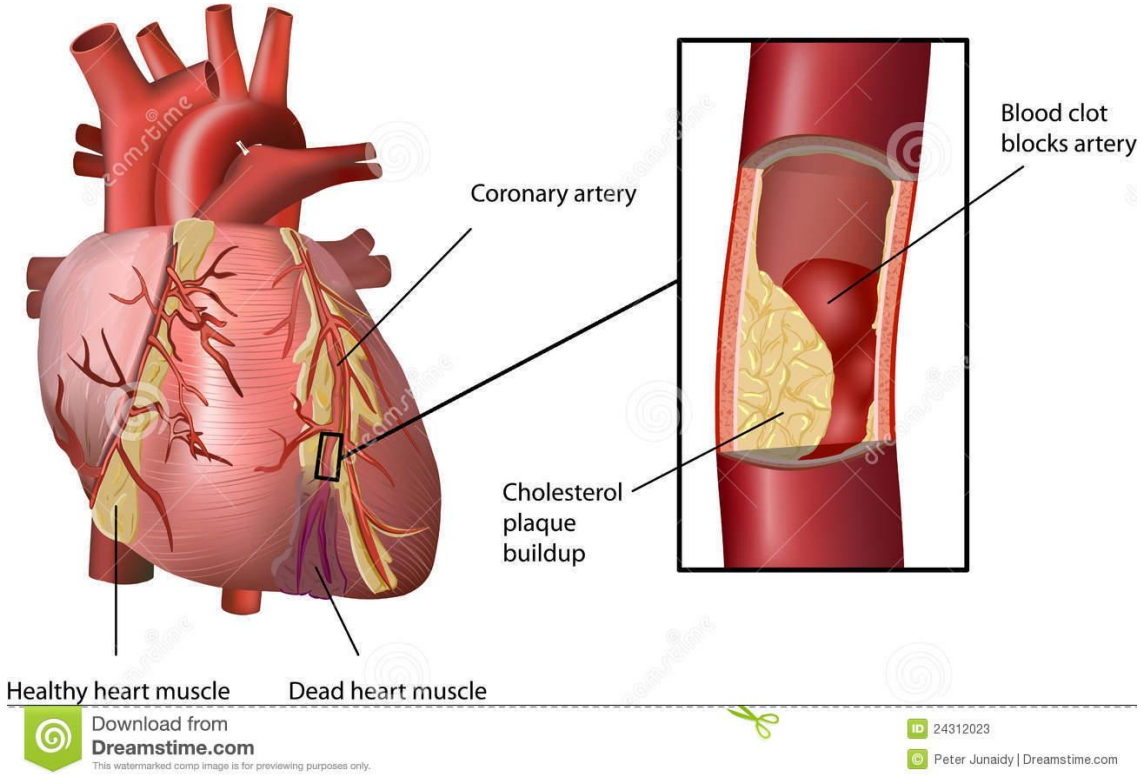
- We allow 2-hour grace period for homework submission
- Late homework/exams without pre-approval will NOT be accepted or graded.

- Homework Collaboration
  - You are encouraged to discuss course material and homework with other students
  - But with restrictions (see details on syllabus)

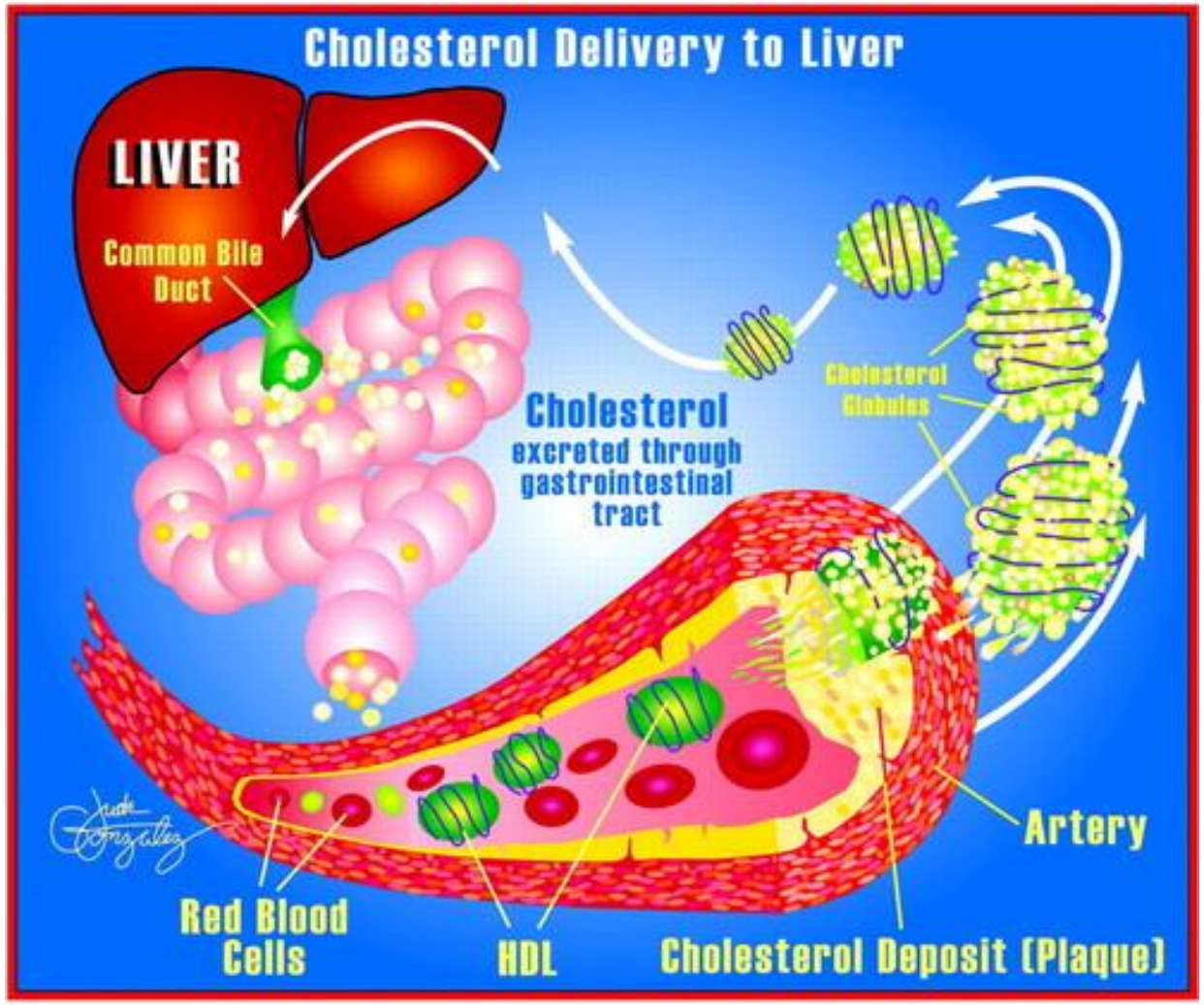
- AI policy
  - Homework and final project
    - Allowed:
      - use AI to help with R coding and debugging
      - polish grammar and language in your write-ups
    - Not allowed:
      - Use AI to generate explanations, causal reasoning, or complete answers
    - Must note the usage in homework / report
  - Quiz
    - AI is not permitted

Example 1

Would high HDL cholesterol level be protective against heart disease?



Beneficial effect of high HDL cholesterol levels?



- Scientists believed that HDL cholesterols (HDL-C) are "good cholesterols" and have several beneficial effects

- Most important ability is to drive a process called "**reverse cholesterol transport**"
- HDL is a "mop" that helps to extract excess cholesterol deposited in blood vessel walls and deliver it back to the liver

- If this is true, we can design a drug to increase HDL-C to help reducing the risk of heart disease

- Can we find empirical evidence to support this hypothesis?

Toth, P. P. (2005). The "good cholesterol" high-density lipoprotein. *Circulation*, *111*(5), e89-e91.

Beneficial effect of high HDL cholesterol levels?

An empirical study to evaluate relationship between HDL-C and the risk of vascular disease [Major lipids, apolipoproteins, and risk of vascular disease. *JAMA,* 2009.]

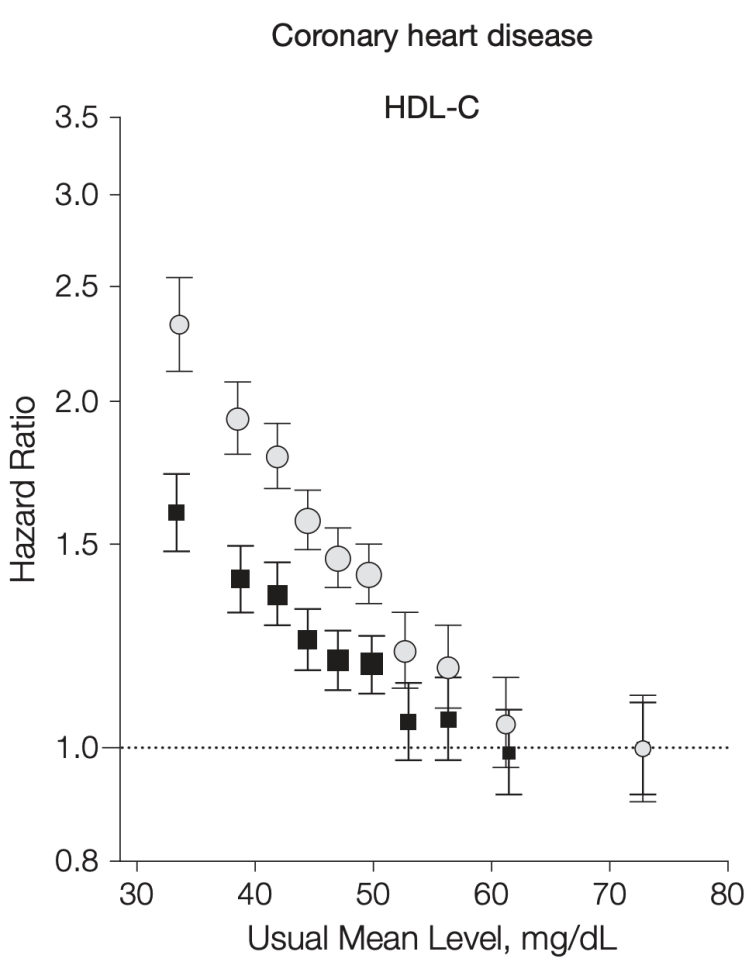
- Individual records were supplied on 302,430 people without initial vascular disease from 68 long-term prospective studies, mostly in Europe and North America.
- Researchers in total observed 8857 nonfatal myocardial infarctions, 3928 coronary heart disease [CHD] deaths, 2534 ischemic strokes, 513 hemorrhagic strokes, and 2536 unclassified strokes

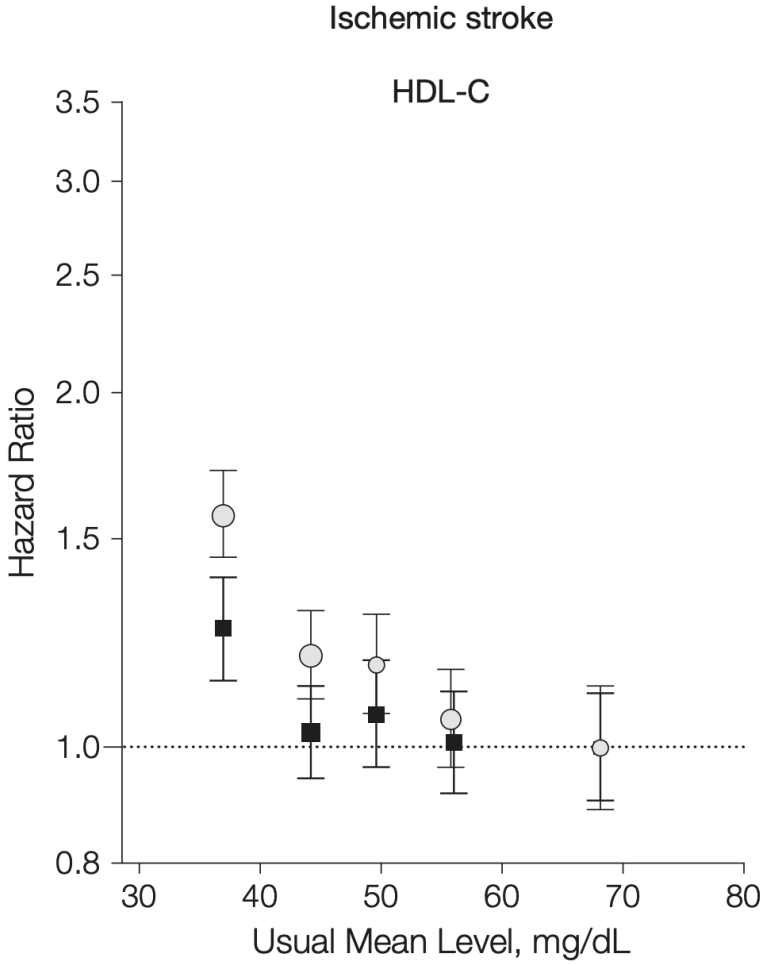
- Researchers compared the risk of vascular disease (measured by the hazard rate, a higher hazard rate corresponds to a higher risk of getting the disease) across individuals having different HDL-C levels.

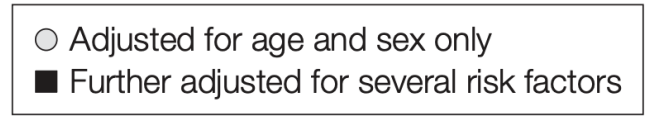
- They used a regression analysis to adjust for confounding factors including age, sex, systolic blood pressure, smoking status, history of diabetes, body mass index, and lipid measures

- **Hazard ratio**: ratio of the hazard rate between two different groups of individuals

Beneficial effect of high HDL cholesterol levels?







- The paper suggested **a strong negative association** between the HDL-C level and risk of vascular disease
- After adjusting for confounding factors, the negative association is weaker, though it's still significant
- Does this suggest the beneficial effect of HDL-C?

Beneficial effect of high HDL cholesterol levels?

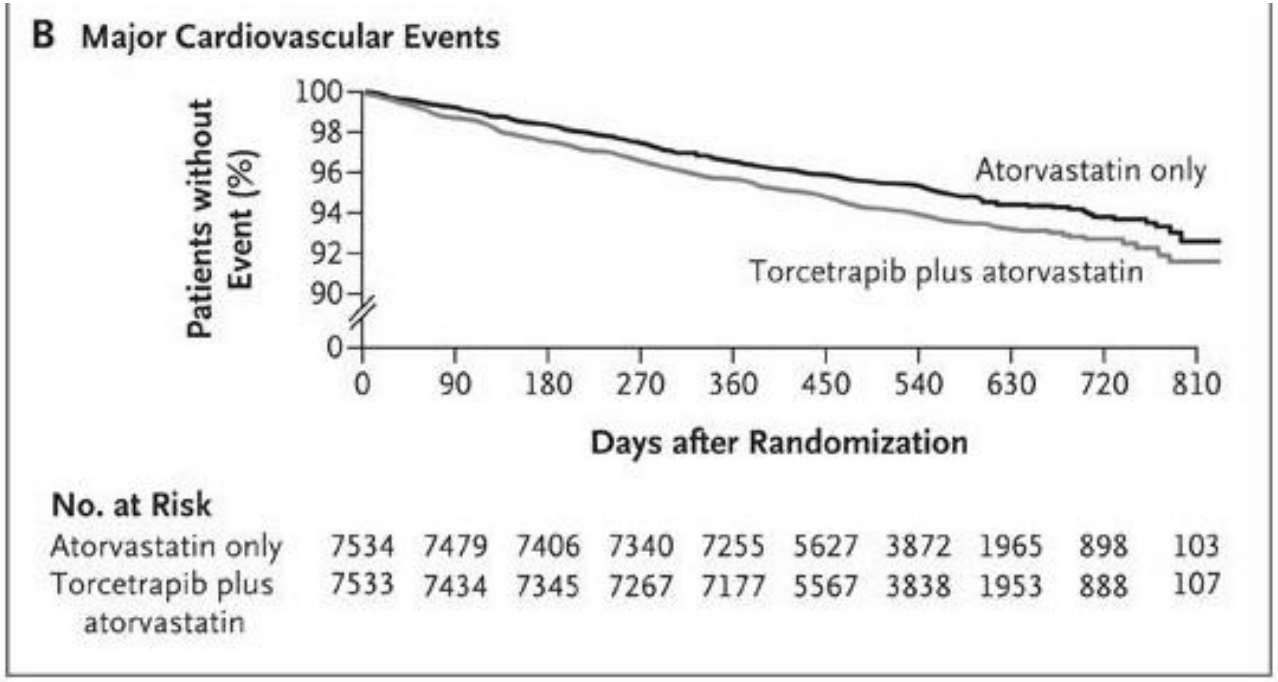
A randomized double-blind study for a drug increasing HDL-C levels
[Effects of torcetrapib in patients at high risk for coronary events. *New England journal of medicine*, 2007]

- The drug torcetrapib: a potent CETP inhibitor that can increase HDL-C levels
- Scientists conducted a randomized, double-blind study involving 15,067 patients at high cardiovascular risk. The patients received either torcetrapib plus atorvastatin or atorvastatin alone. (atorvastatin: an FDA approved drug to treat heart disease)
- The primary outcome was the time to the first major cardiovascular event, time to death from coronary heart disease, nonfatal myocardial infarction, stroke, or hospitalization for unstable angina

Beneficial effect of high HDL cholesterol levels?

Table 2. Changes from Baseline at 3 Months and 12 Months in Selected Measures.*

Variable	Change at 3 Months			Change at 12 Months		
	Atorvastatin Only	Torcetrapib plus Atorvastatin	P Value	Atorvastatin Only	Torcetrapib plus Atorvastatin	P Value
Lipids (absolute change) — mg/dl						
Cholesterol						
Total	+1.6±20.5	+5.1±23.9	<0.001	+2.1±22.4	+9.3±26.3	<0.001
High-density lipoprotein	+0.5±6.2	+29.0±14.4	<0.001	+0.5±6.8	+34.2±17.0	<0.001
Low-density lipoprotein	+0.6±15.8	−20.5±20.8	<0.001	+0.9±17.1	−21.5±22.7	<0.001



- Torcetrapib does greatly increase HDL-C levels
- The survival probability for the group with torcetrapib decreases even a bit faster
- The randomized trial suggests the failure of the drug
- Why is there a contradiction?
  - We may have not adjusted for enough confounding factors
  - The drug may have other effects

Example 2

Birth weight paradox: Does maternal smoking have a beneficial effect to reduce infant mortality?

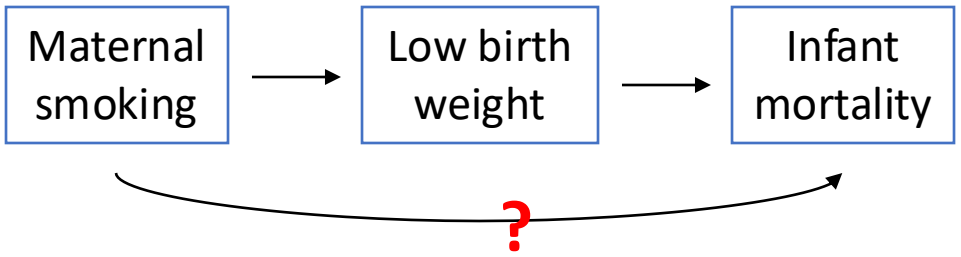
https://consultqd.clevelandclinic.org/smoking-even-one-cigarette-while-pregnant-doubles-the-risk-of-sudden-unexpected-infant-death-study-finds/

Maternal smoking, birth weight and perinatal mortality

How does a smoking mother affect her baby?

- Researchers have observed that women who smoke have smaller infants since long time ago (Simpson, 1957)
- Birth weight is a strong predictor of neonatal and infant mortality
- Low birthweight (babies who are born weighing less than 2,500 grams, average newborn weights about 8 pounds) account for 60–80% of all neonatal deaths [4 million neonatal deaths: When? Where? Why? *Lancet,* 2005]

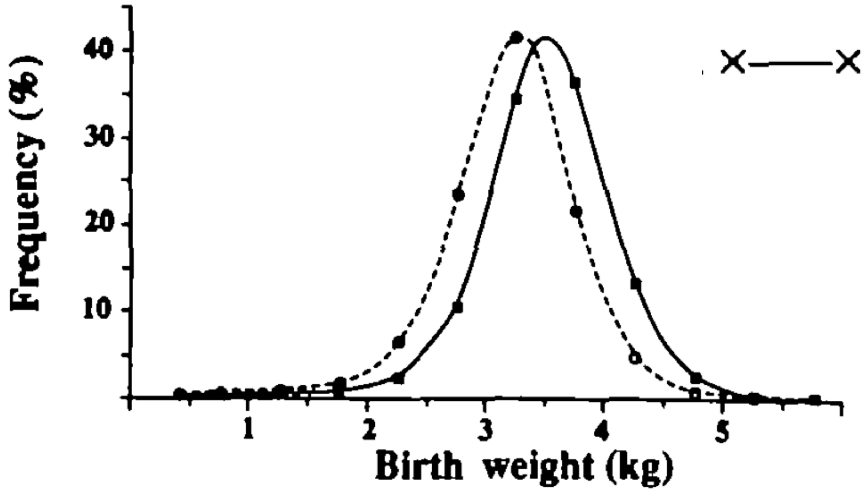
- Question: besides reducing the babies' birthweight, are their other effects of maternal smoking on infant health?



Maternal smoking, birth weight and perinatal mortality

An analysis by Wilcox
[Birth weight and perinatal mortality: the effect of maternal smoking. *American journal of epidemiology*, 1993.]

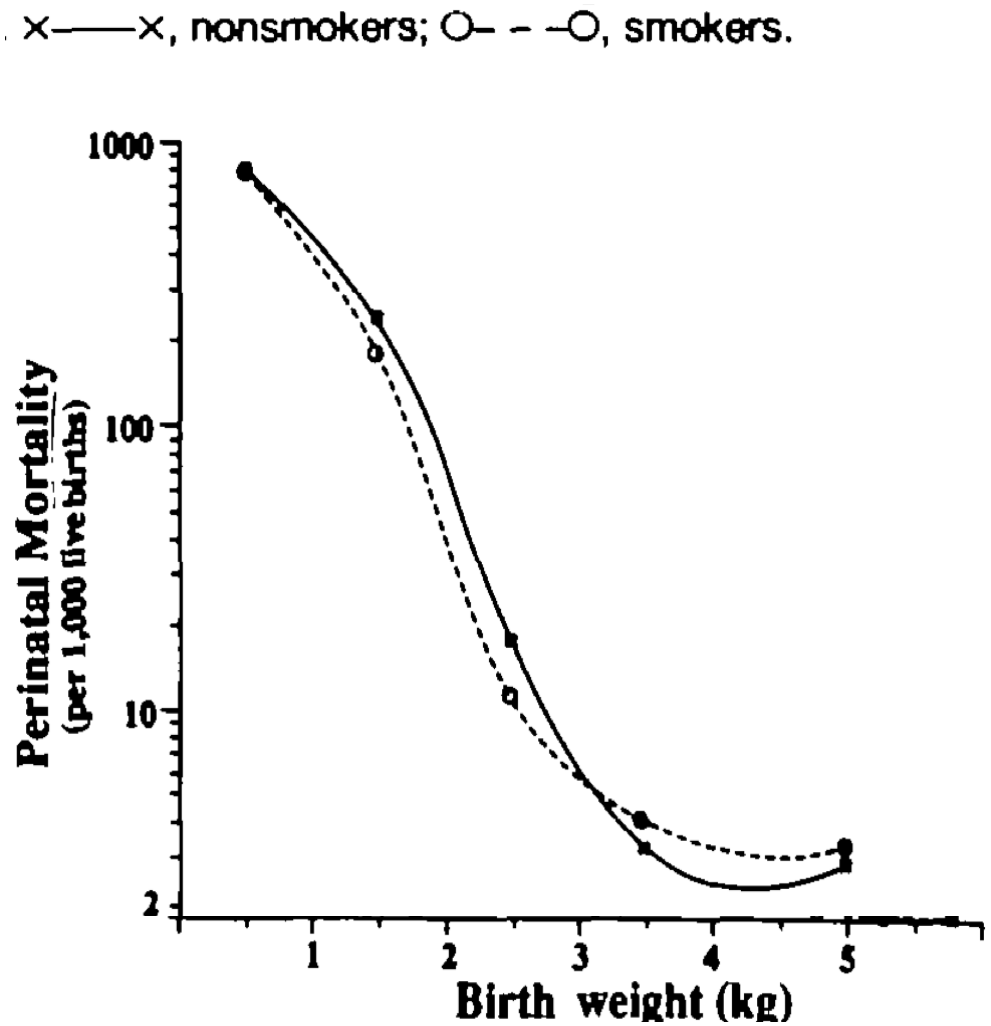
- Data source: a file of Missouri vital statistics records for 1980-1984, assembled as part of the National Institute of Child Health and Human Development Multinational Study of Birth Weight-specific Perinatal Mortality Rates
- perinatal mortality: stillbirths plus deaths in the first 28 days
- Two groups of samples: mothers who had reported no smoking during pregnancy and those who reported smoking at least one pack of cigarettes a day
- 215,428 babies in the first group (unexposed group) and 42,270 babies in the second group (exposed group)



×——×, nonsmokers; O– – –O, smokers.

- Infants of smoking group were, on average, 320 g lighter than unexposed infants (3,180 g compared with 3,500 g).
- Perinatal mortality in Missouri is 14.5/ 1,000 infants born to smokers, compared with 10.4 for unexposed infants.
- What about birthweight-adjusted mortality rate?

Maternal smoking, birth weight and perinatal mortality



- Surprisingly, among infants less than 3 kg, weight-specific mortality rates are lower for exposed infants than unexposed.

Birth-weight paradox

- Would this suggest a beneficial direct effect of maternal smoking towards infant mortality?

Maternal smoking, birth weight and perinatal mortality

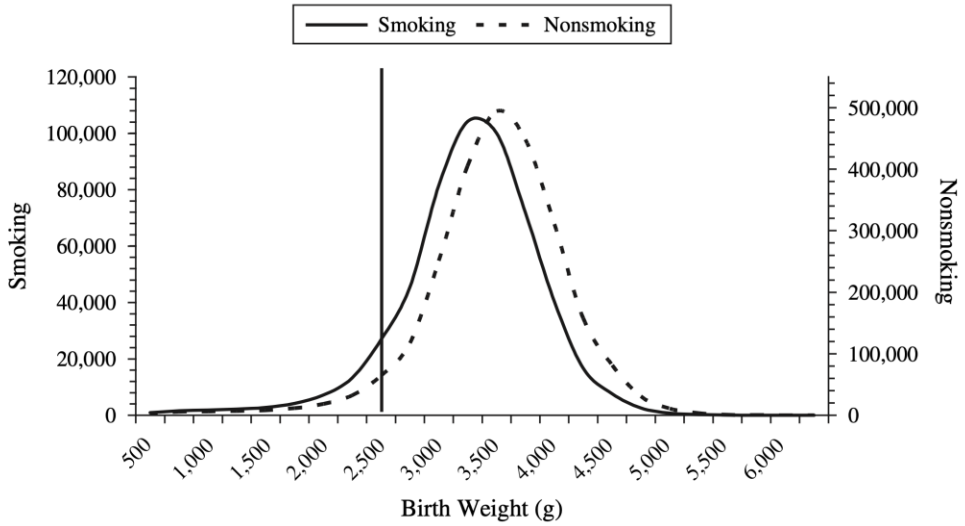
Another data analysis

[The birth weight "paradox" uncovered?. *American journal of epidemiology*, 2006]

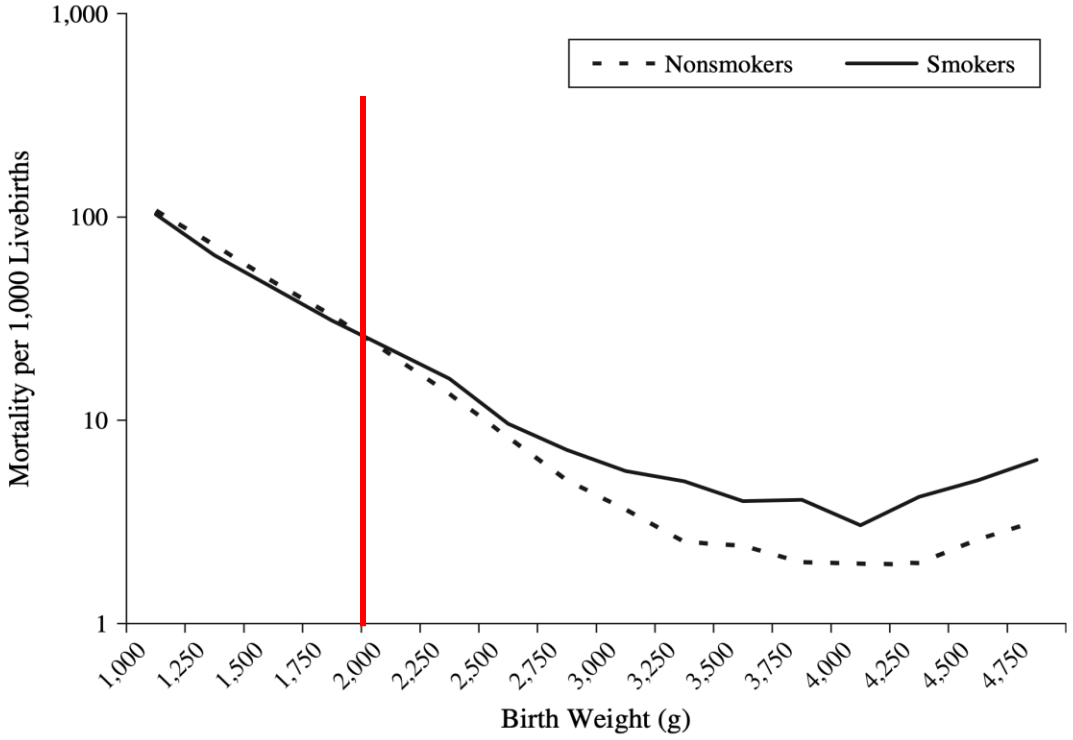
- Data source: all infants born alive in the United States in 1991 through the national linked birth/infant- death data sets assembled by the National Center for Health Statistics (about 3 million babies for analysis)

- birth-weight-specific infant mortality is calculated by stratify the babies into 250g categories, and calculate the birth mortality within each category
- Researchers also adjust of other potential confounders: maternal age, gravidity, education, marital status, race/ethnicity, and prenatal care via logistic regression

- We observe a similar weight reduction for the smoking group (3,145g v.s. 3,370g)

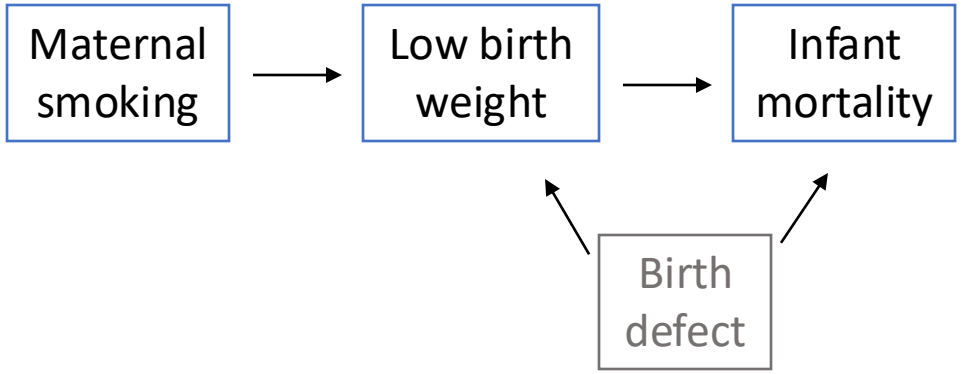


Maternal smoking, birth weight and perinatal mortality



- Low birthweight infant: infant mortality rate ratio for exposed versus nonexposed infants was 0.79 (95% CI: 0.76, 0.82)
- infant mortality rate ratio is 1.80 (95 percent CI: 1.72, 1.88) among infants with higher birth weights.

- A possible explanation of the birthweight paradox



- Both maternal smoking and birth defects (or malnutrition) can cause low birth weight

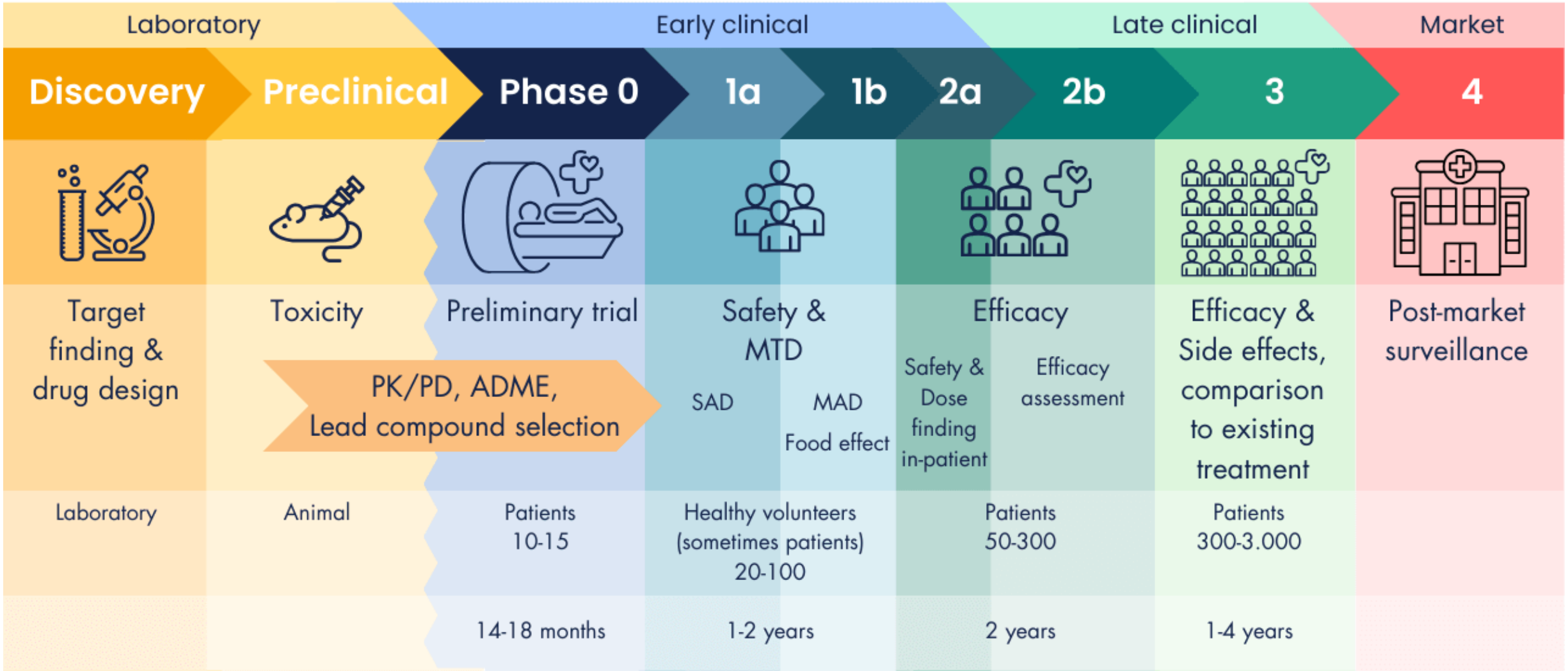
- For the unexposed group baby with low birth weight, they are more likely to have birth defects than the exposed group, which can directly increase infant mortality

Example 3

Randomized trials for drug development

https://www.usatoday.com/story/news/health/2020/12/15/safe-and-effective-moderna-covid-vaccine-poised-fda-authorization/3901783001/

Phases of drug development



PK PharmacoKinetics
PD PharmacoDynamics

ADME Absorption, Distribution, Metabolism and Excretion
MTD Maximum Tolerated Dose

TRACER

Simplified representation of phases in drug development. Study content and dates and numbers given may vary between studies. No rights can be derived from this figure.

www.tracercro.com

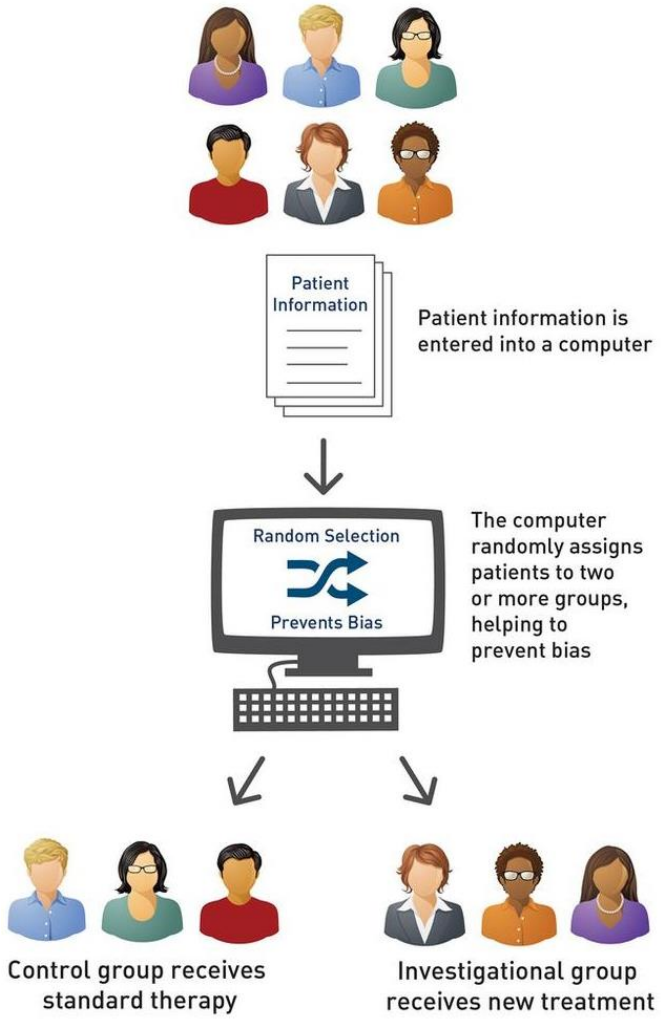
Clinical Trial Phases

- Phase 0 (not randomized): Testing a low dose of the treatment to check it isn't harmful
- Phase 1 (not randomized): Finding out about side effects, and what happens to the treatment in the body
- Phase 2 (sometimes randomized): Finding out more about side effects and looking at how well the treatment works
- Phase 3 (randomized): Comparing the new treatment to the standard treatment
  - Gold standard for FDA approval
- Phase 4 (not randomized): Finding out more about long term benefits and side effects

https://www.ipmglobal.org/our-work/research/clinical-trial/clinical-trials-approach

Phase 3 randomized trial

- Randomized, placebo-controlled trials is the gold standard for FDA approval
- On December 18, 2020, FDA approved an emergency use authorization (EUA) for the Moderna vaccine against COVID-19
- The EUA is based a rigorous evaluation of the safety, effectiveness and manufacturing quality of the vaccine
  - Why EUA?



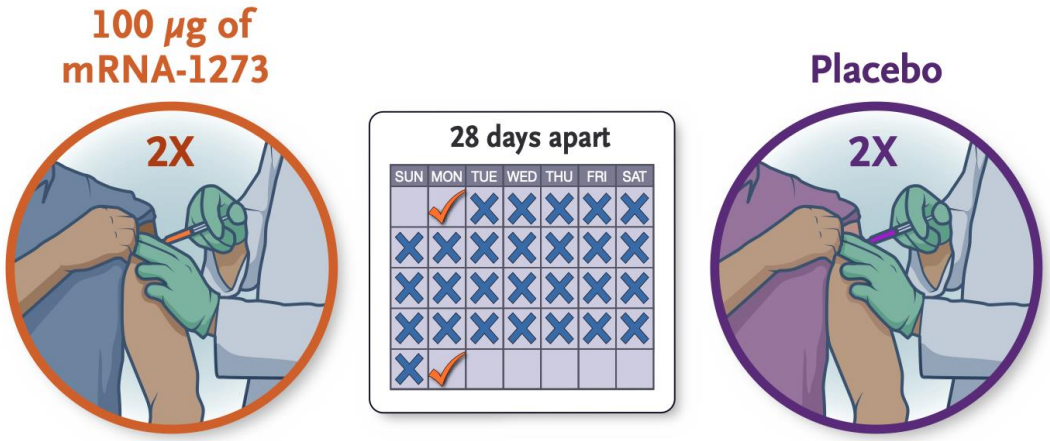
https://www.ipmglobal.org/our-work/research/clinical-trial/clinical-trials-approach

Phase 3 randomized trial for the Moderna vaccine

Phase 3 randomized trial for the Moderna vaccine
[Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *New England journal of medicine*, 2020.]

- The phase 3 randomized, observer-blinded, placebo-controlled trial was conducted at 99 centers across the United States

- The trial enrolled 30,420 volunteers who were randomly assigned in a 1:1 ratio to receive either vaccine or placebo (15,210 participants in each group).
- Patients receive two intramuscular injections of mRNA-1273 (100 µg) or placebo 28 days apart.



Phase 3 randomized trial for the Moderna vaccine

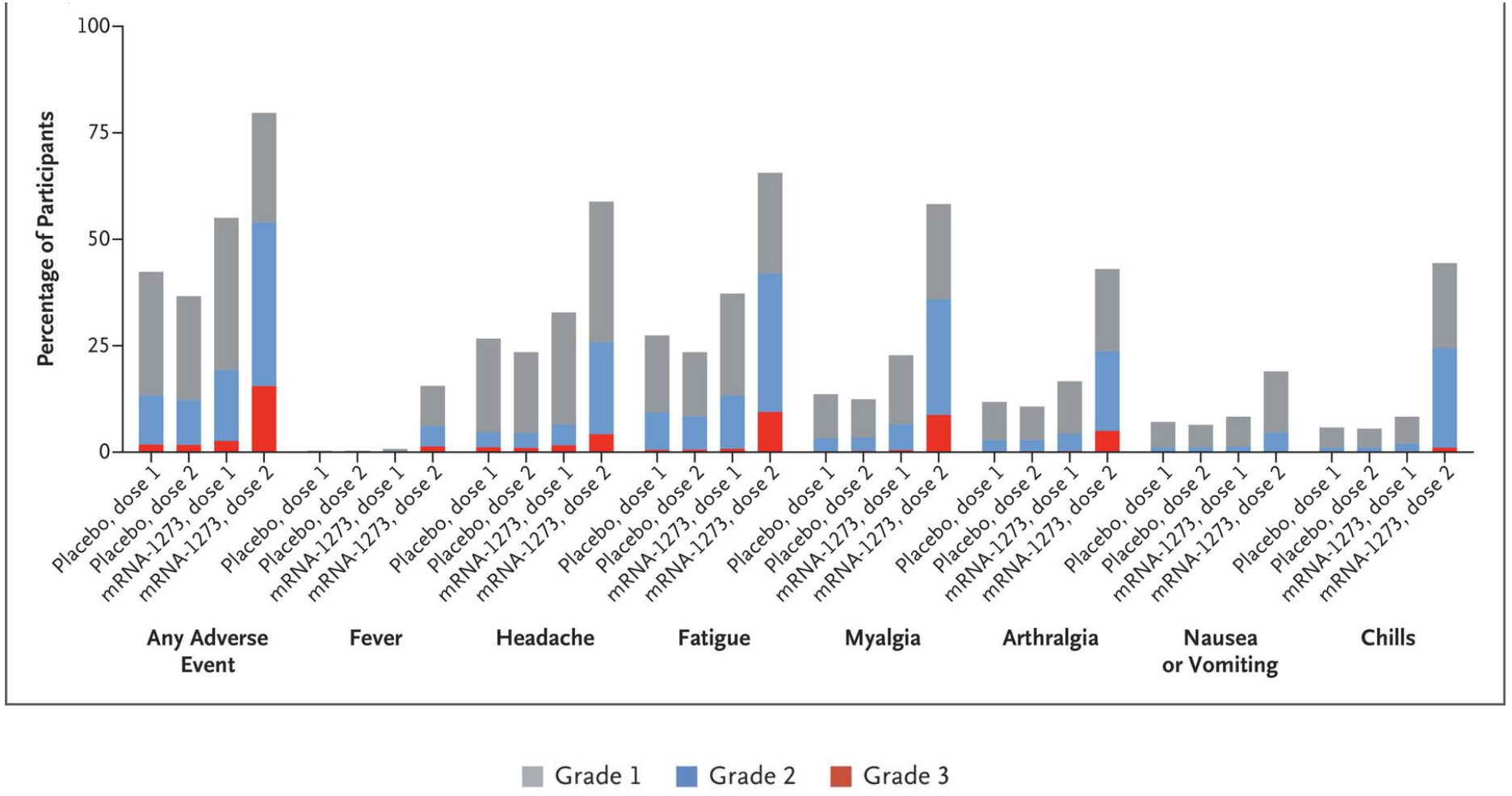
- Because of randomization, characteristics of individuals are well balanced between the two groups
- Trial is not trustable if covariates are not balanced even in a randomized experiment

Table 1. Demographic and Clinical Characteristics at Baseline.*

Characteristics	Placebo (N=15,170)	mRNA-1273 (N=15,181)	Total (N=30,351)
Sex — no. of participants (%)			
Male	8,062 (53.1)	7,923 (52.2)	15,985 (52.7)
Female	7,108 (46.9)	7,258 (47.8)	14,366 (47.3)
Mean age (range) — yr	51.3 (18–95)	51.4 (18–95)	51.4 (18–95)
Age category and risk for severe Covid-19 — no. of participants (%)†			
18 to <65 yr, not at risk	8,886 (58.6)	8,888 (58.5)	17,774 (58.6)
18 to <65 yr, at risk	2,535 (16.7)	2,530 (16.7)	5,065 (16.7)
≥65 yr	3,749 (24.7)	3,763 (24.8)	7,512 (24.8)
Hispanic or Latino ethnicity — no. of participants (%)‡			
Hispanic or Latino	3,114 (20.5)	3,121 (20.6)	6,235 (20.5)
Not Hispanic or Latino	11,917 (78.6)	11,918 (78.5)	23,835 (78.5)
Not reported and unknown	139 (0.9)	142 (0.9)	281 (0.9)
Race or ethnic group — no. of participants (%)‡			
White	11,995 (79.1)	12,029 (79.2)	24,024 (79.2)
Black or African American	1,527 (10.1)	1,563 (10.3)	3,090 (10.2)
Asian	731 (4.8)	651 (4.3)	1,382 (4.6)
American Indian or Alaska Native	121 (0.8)	112 (0.7)	233 (0.8)
Native Hawaiian or Other Pacific Islander	32 (0.2)	35 (0.2)	67 (0.2)

Phase 3 randomized trial for the Moderna vaccine

- Vaccine Safety (even the placebo group can observe some adverse events)

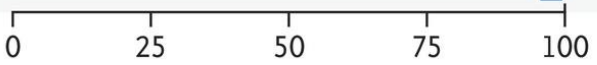


Phase 3 randomized trial for the Moderna vaccine

- Vaccine Efficacy in subgroups

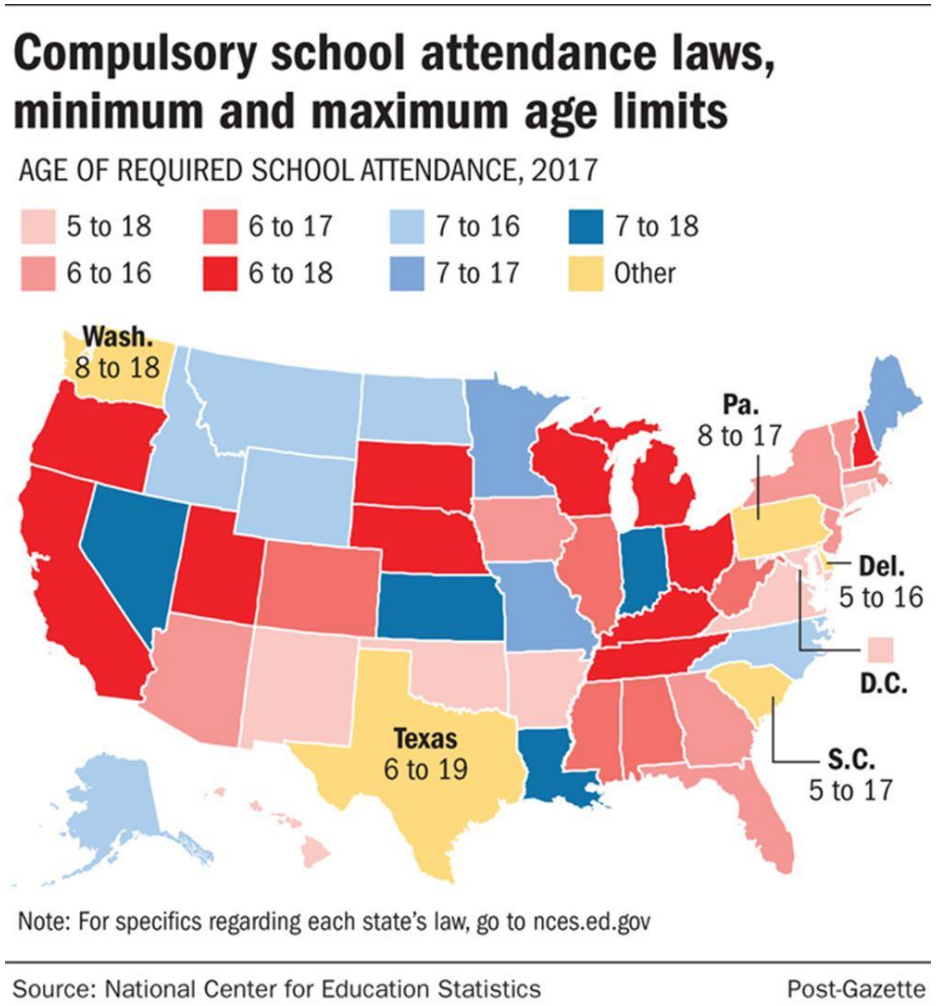
$$\text{Vaccine efficacy} = 1 - \frac{P(\text{disease cases} \mid \text{vaccinated})}{P(\text{disease cases} \mid \text{not vaccinated})}$$

Subgroup	Placebo (N=14,073)	mRNA-1273 (N=14,134)	Vaccine Efficacy (95% CI)
	no. of events/total no.		
All patients	185/14,073	11/14,134	94.1 (89.3–96.8)
Age			
≥18 to <65 yr	156/10,521	7/10,551	95.6 (90.6–97.9)
≥65 yr	29/3552	4/3583	86.4 (61.4–95.2)
Age, risk for severe Covid-19			
18 to <65 yr, not at risk	121/8403	5/8396	95.9 (90.0–98.3)
18 to <65 yr, at risk	35/2118	2/2155	94.4 (76.9–98.7)
≥65 yr	29/3552	4/3583	86.4 (61.4–95.2)
Sex			
Male	87/7462	4/7366	95.4 (87.4–98.3)
Female	98/6611	7/6768	93.1 (85.2–96.8)
At risk for severe Covid-19			
Yes	43/3167	4/3206	90.9 (74.7–96.7)
No	142/10,906	7/10,928	95.1 (89.6–97.7)
Race and ethnic group			
White	144/8916	10/9023	93.2 (87.1–96.4)
Communities of color	41/5132	1/5088	97.5 (82.2–99.7)



Example 4

Effect of compulsory school attendance on schooling and earnings



https://www.post-gazette.com/news/education/2019/03/14/Pennsylvania-Tom-Wolf-students-school-age-requirement-Allegheny-County-Pittsburgh/stories/201903070125

The impact of compulsory schooling on earnings

- Scientists aim to assess whether students who attend school longer receive higher earnings as a result of their increased schooling

- Say, we think that the weekly wages can depend on both the education levels and a person's own ability, family background et. al…

$$\log(\text{Weekly wage}_i) = \beta_0 + \beta_1 \text{Schooling}_i + \beta_2 \text{Ability}_i + \text{noise}_i$$

- If we can not measure one's ability, and one's ability is correlated with one's education level, then if we only regress weekly wages on schooling, then we will have a biased estimate of the effect of schooling

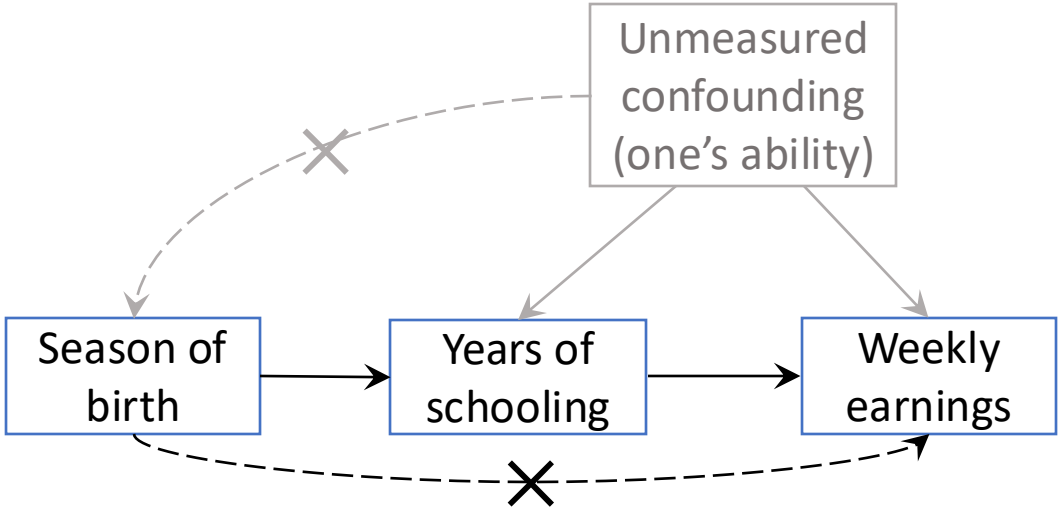
- How can we adjust for the bias? (ability here is called an unmeasured confounding factor)

- We want to find an instrument that is associated with schooling and is guaranteed to be independent from the unmeasured confounding of ability

The impact of compulsory schooling on earnings

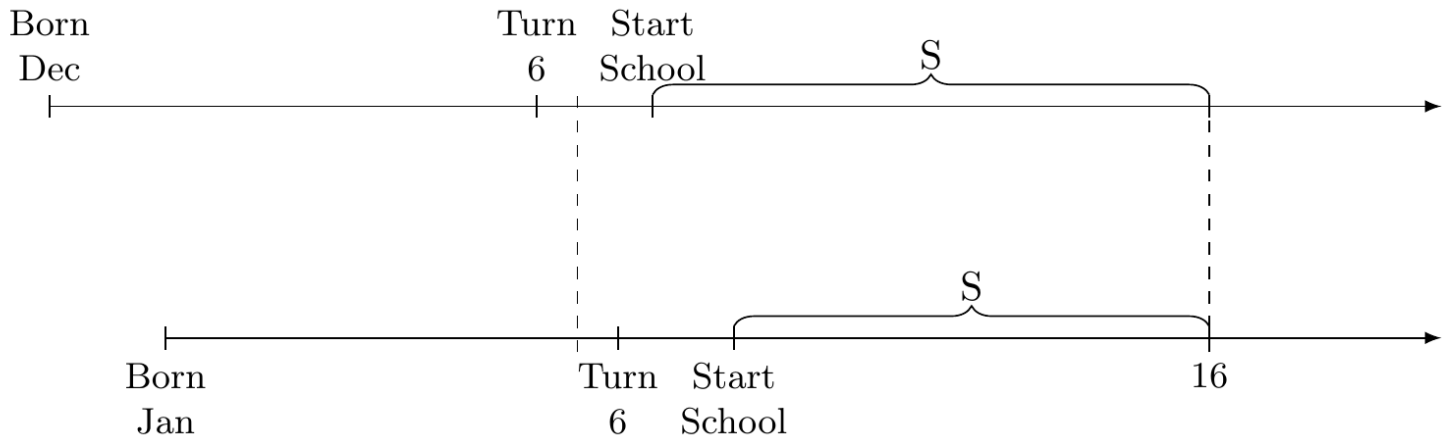
A study by two economists J.D. Angrist and A.B. Krueger
[Does compulsory school attendance affect schooling and earnings?. *The Quarterly Journal of Economics*, 1991.]

- Because of the compulsory schooling laws, children born in different months of the year can start school at different ages and have different years of education
- So, they used season of birth as an instrument to understand the causal effect of schooling:



Why is season of birth an instrument?

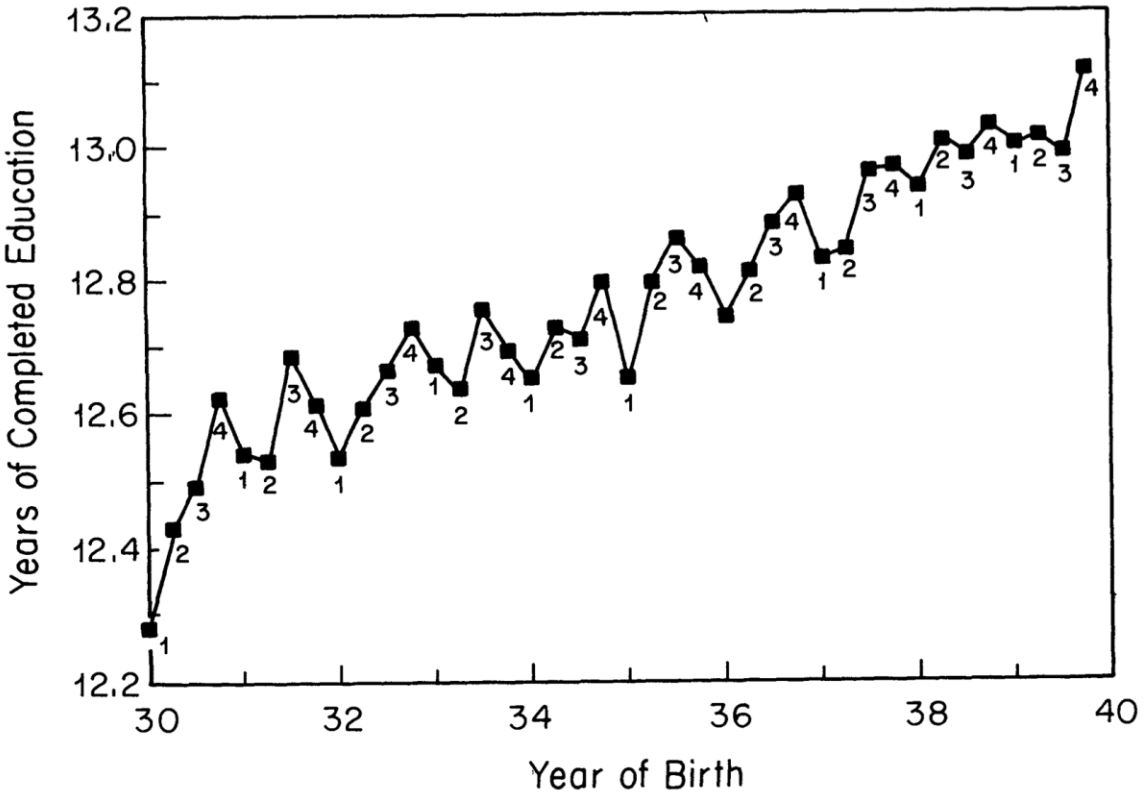
- School districts typically require a student to have turned age six by January 1 of the year in which he or she enters school
- students born earlier in the year enter school at an older age and attain the legal dropout age at an earlier point in their educational careers than students born later in the year
- Season of birth should be independent from other unmeasured confounding factors like ability, family of background …



https://mixtape.scunning.com/instrumental-variables.html

Season of birth affects school years

- Empirical evidence showing that kids born in earlier seasons indeed have a shorter length of education



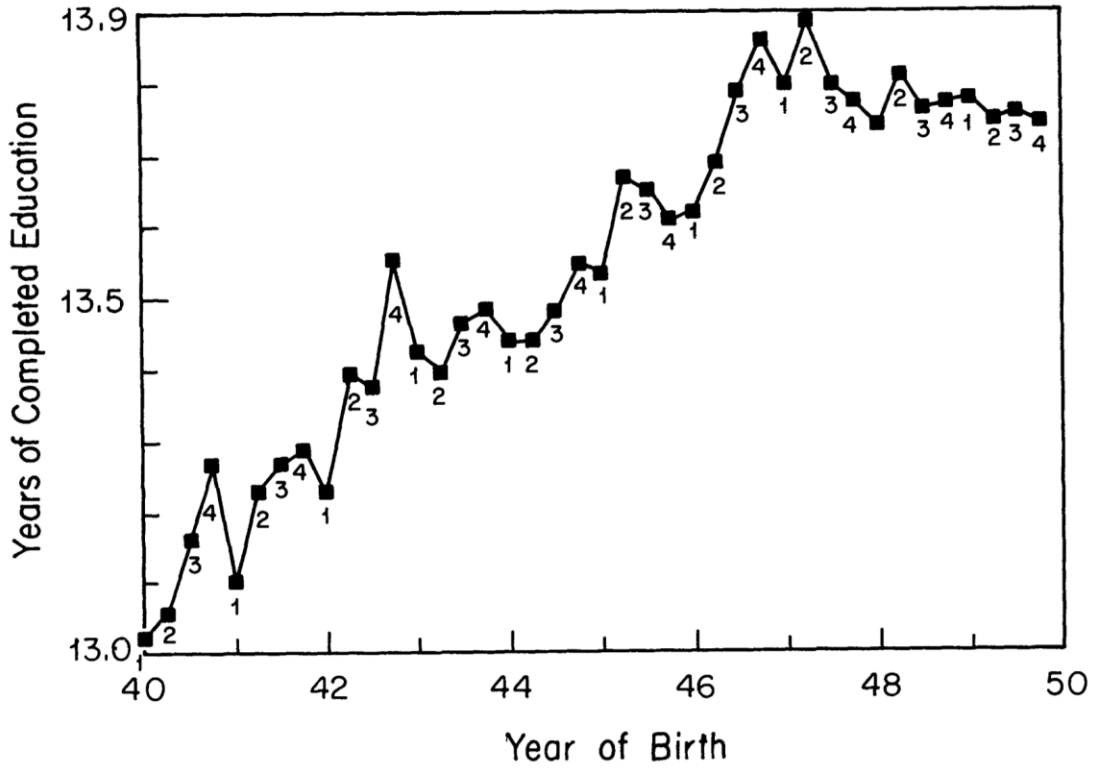
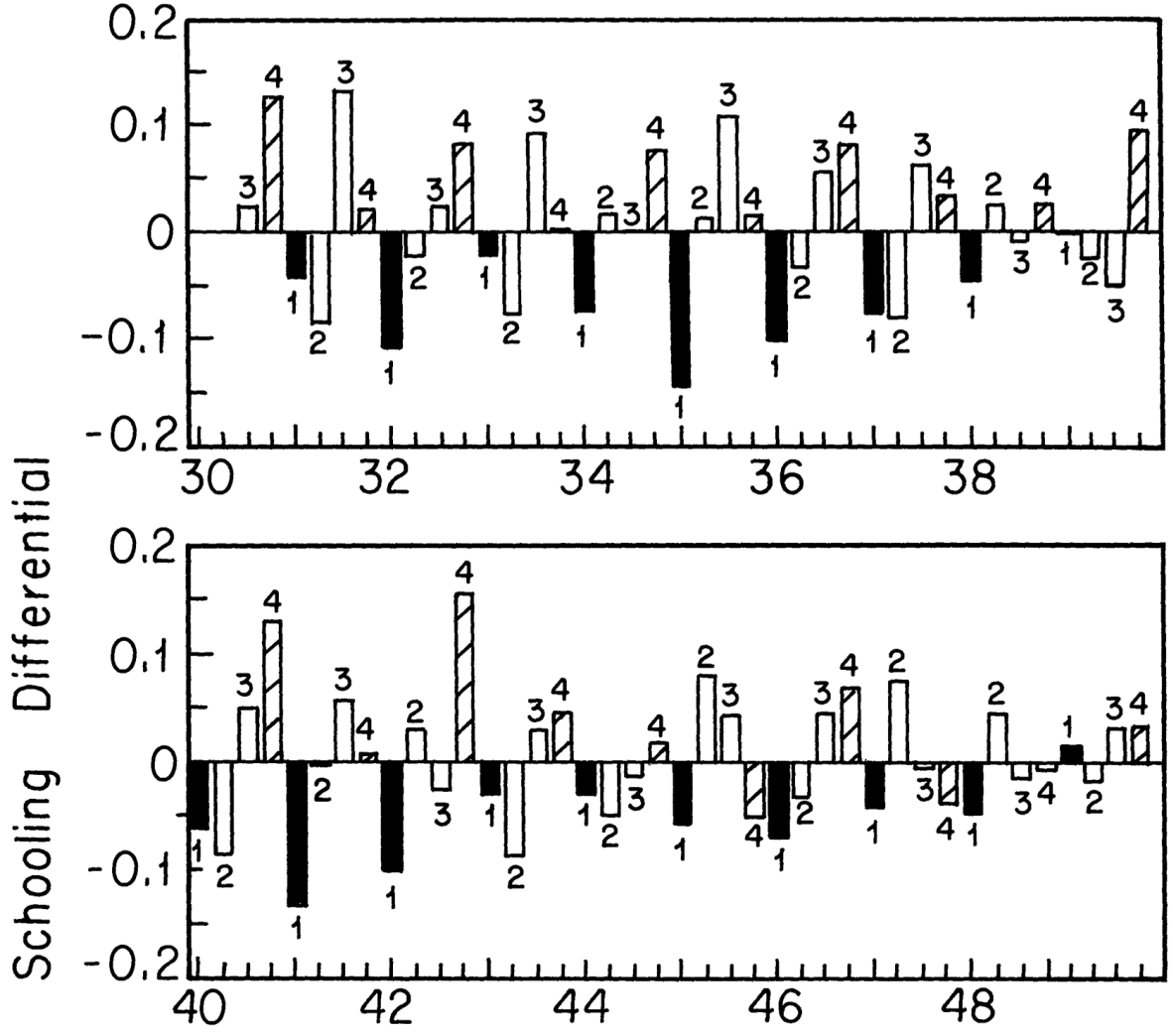


FIGURE I
Years of Education and Season of Birth
1980 Census
Note. Quarter of birth is listed below each observation.

Season of birth affects school years

- After removing the year trend



Season of birth affects weekly earnings

- Kids born at earlier seasons are associated with a lower weekly earnings

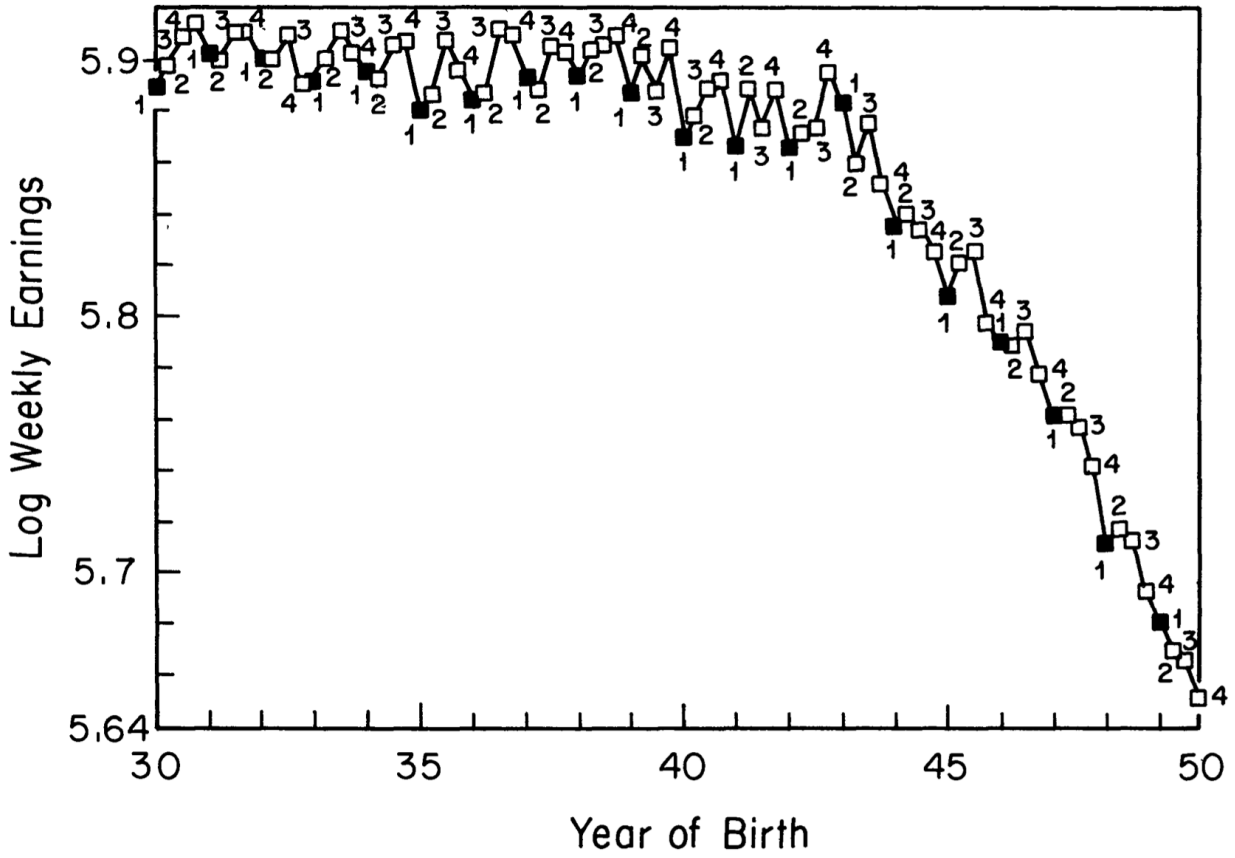


FIGURE V
Mean Log Weekly Wage, by Quarter of Birth
All Men Born 1930–1949; 1980 Census

The logic of using the instrument:

- Born later in the year -> More years at school -> higher weekly earnings

Causal inference

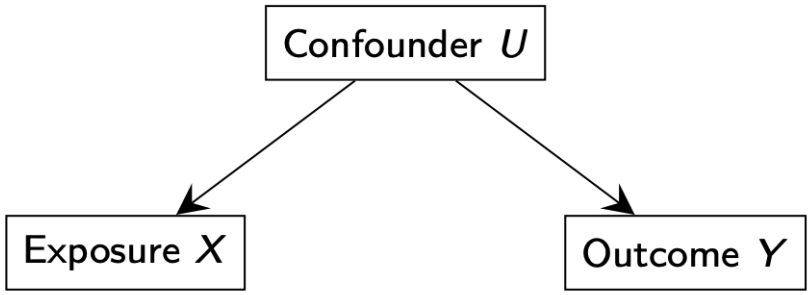
- To summarize, most scientific questions are causal questions
- We know what causal effects mean as a human being

> *I would rather discover one causal law than be King of Persia.*
> *— Democritus*
>
> *We have knowledge of a thing only when we have grasped its cause.*
> *— Aristotle, Posterior Analytics*

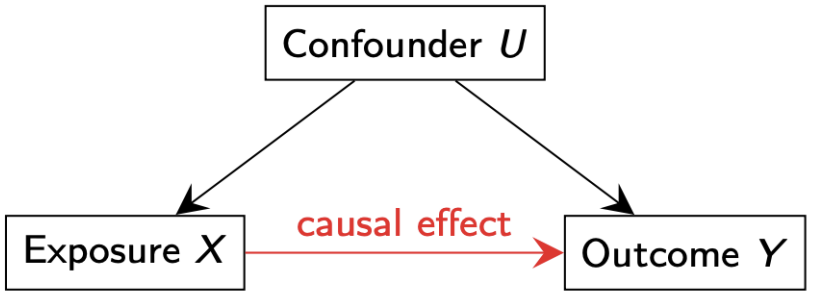
- How to quantitatively define “causal effects” with mathematical notations?
  - When are regression adjustments correct? What variables should I include?
  - Are there other approaches that can help us identify and estimate a causal effect more reliably (require less assumptions and robust to violations of assumptions)

Association ≠ Causation

- Confounding



(a) Correlated but not causal



(b) Causal

- In randomized experiments, treatment is independent from confounders and is the gold standard for causal inference
  - How to perform statistical estimation for randomized experiments with minimum assumptions?
  - What if our randomized experiments are not perfect?

- In observational studies, we are always worried about Confounding
  - How do we adjust for known confounders?
  - How do we deal with unmeasured hidden confounders?

Reference papers to read

Example 1:

- Assessment, R. E. L. I. A. B. L. E. (2009). Major lipids, apolipoproteins, and risk of vascular disease. *Jama*, *302*(18), 1993-2000. https://jamanetwork.com/journals/jama/article-abstract/184863
- Barter, P. J., Caulfield, M., Eriksson, M., Grundy, S. M., Kastelein, J. J., Komajda, M., ... & Brewer, B. (2007). Effects of torcetrapib in patients at high risk for coronary events. New England journal of medicine, 357(21), 2109-2122. https://www.nejm.org/doi/10.1056/NEJMoa0706628?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200www.ncbi.nlm.nih.gov

Example 2:

- Wilcox, A. J. (1993). Birth weight and perinatal mortality: the effect of maternal smoking. American journal of epidemiology, 137(10), 1098-1104. https://academic.oup.com/aje/article-abstract/137/10/1098/128195
- Hernández-Díaz, S., Schisterman, E. F., & Hernán, M. A. (2006). The birth weight "paradox" uncovered?. American journal of epidemiology, 164(11), 1115-1120. https://academic.oup.com/aje/article/164/11/1115/61454

Example 3:

- Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., ... & Zaks, T. (2020). Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. New England journal of medicine. https://www.nejm.org/doi/full/10.1056/nejmoa2035389

Example 4:

- Angrist, J. D., & Keueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings?. The Quarterly Journal of Economics, 106(4), 979-1014. https://www.jstor.org/stable/2937954

Please use Uchicago proxy! to open the links if you can not get the full pdf off campus