MIT 14.32/14.320 **Spring 2023**
J. Angrist

Lecture Note 4

Causality, Experiments, and Potential Outcomes

1 Casual vs Causal Effects

In an argument that's far from casual, Americans debate the causal effects of health insurance. How does health insurance, or the lack thereof, affect health? In Breaking Bad, bad health insurance is responsible for Walter White's transformation from high school chemistry teacher to Heisenberg the Meth Dealer. The view that health insurance is the key to good health motivated the 2010 Affordable Care Act, also known as Obamacare.

The Affordable Care Act imposed tax penalties on those who fail to sign up for health insurance, but it remains true that some Americans are covered and some aren't (The 2017 Tax Cuts and Jobs Act ended tax penalties for the uninsured). This brings us to the question at the heart of MM Chapter 1:

> Are the insured healthier than *they* would have been had they not been insured?

Implicit in this question is a *what if* comparison. The answer is not obvious: after all, anyone can go to a hospital emergency department in an hour of need (federal law requires EDs to treat all comers). This might be coverage enough.

The insured are indeed substantially healthier than the uninsured. But perhaps this reflects something about the people who are lucky enough to have access to low-cost insurance (like public sector workers) or are rich enough to pay for it (like MIT faculty). The insured differ from the uninsured for many reasons besides their insurance.

Formal notation for *potential outcomes* makes causal questions precise. For each person, indexed by i, define two possible outcomes:

- Health of person i when i is insured: Y_{1i}
- Health of person i when i is uninsured: Y_{0i}

In practice, health can be measured by self-reports on a survey, or objective measures like blood pressure.

Potential outcomes indexed to alternative treatments are the keys to causal inference. The causal effect of insurance on person i is:

$$Y_{1i} - Y_{0i}.$$

We never see this individual-level causal effect because, at any point in time, i is either insured or not.

- Still, we can hope to measure the *average treatment effect* (ATE) of insurance, an *average causal effect* defined as:
$$E[Y_{1i} - Y_{0i}].$$
- We might also want to gauge the average causal effect of health insurance on the insured:
$$E[Y_{1i} - Y_{0i}|D_i = 1],$$
where D_i is a dummy variable equal to 1 for the insured. This parameter is called the *effect of treatment on the treated* (TOT).
  - ATE characterizes the impact of insurance averaged over all those in the population of interest, while TOT tells us whether the insured population benefits (on average) from coverage.

1.1 Selection Bias

Research on causal effects is often motivated by an interest in TOT. This parameter can be written:

$$E[Y_{1i} - Y_{0i}|D_i = 1] = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \tag{1}$$

TOT compares the health of the insured, $E[Y_{1i}|D_i = 1]$, with *their* health when uninsured, $E[Y_{0i}|D_i = 1]$. We can reliably estimate $E[Y_{1i}|D_i = 1]$ in a random sample, but $E[Y_{0i}|D_i = 1]$ is *never* seen.

- $E[Y_{0i}|D_i = 1]$ is therefore said to be *counterfactual*
- The importance of counterfactuals emerges in a comparison of health between insured and uninsured, which can be written:

  $$E[Y_i|D_i = 1] - E[Y_i|D_i = 0]. \tag{2}$$

  Note that the Y_i's in (2) have no 0 and 1 subscripts because they signify *observed outcomes* as opposed to *potential outcomes.* We're also ignoring the fact that, in practice, we make comparisons using sample means and not expectations; for the moment, this is a detail.
- Observed and potential outcomes are related. Specifically, we have:

  $$Y_i = Y_{0i}(1 - D_i) + Y_{1i}D_i \tag{3}$$

- In other words, we see Y_{0i} for the uninsured and Y_{1i} for the insured:

  $$\begin{aligned} &E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\ &\quad = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]. \end{aligned} \tag{4}$$

  It stands to reason that the difference in average outcomes between insured and uninsured in equation (4) tells us *something* about the average causal effect we're after in equation (1). But not necessarily what we most want to know.
- Adding and subtracting counterfactuals in equation (4), we get:

  $$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] + \{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]\} \\ &= \underbrace{E[Y_{1i} - Y_{0i}|D_i = 1]}_{TOT} + \underbrace{\{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{selection\ bias} \end{aligned}$$

  The difference in average health between the insured and uninsured is the causal effect of insurance on the insured (TOT) plus the term in curly brackets.
  - This important term is called *selection bias.* Selection bias arises when the Y_{0i}'s of the insured differ, on average, from those of the uninsured.

1.2 Insured and Otherwise in the NHIS

- With health measured on a five point scale, the insured feel a lot better! Check out MM Table 1.1, constructed from the 2009 National Health Interview Survey
- The statistical significance of gaps in health by insurance status is not in doubt
- Statistical inference is easy: with estimates and standard errors in hand, you're good to go!

TABLE 1.1
Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
A. Health						
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
B. Characteristics						
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47 (.05)	3.49	3.93	−.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Notes: This table reports average characteristics for insured and uninsured married couples in the 2009 National Health Interview Survey (NHIS). Columns (1), (2), (4), and (5) show average characteristics of the group of individuals specified by the column heading. Columns (3) and (6) report the difference between the average characteristic for individuals with and without health insurance (HI). Standard deviations are in brackets; standard errors are reported in parentheses.

- Still, we must ask: what do these differences in means mean?
- Panel B should worry those invested in causal claims: when it comes to comparisons by insurance status, *ceteris* is surely not *paribus*
  - Why is covariate imbalance important?
  - What does this imbalance suggest for the sign of the HI/no-HI contrast in expected Y_{0i}, that is, for the selection-bias term, $E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$?

1.3 Random Assignment Eliminates Selection Bias

When insurance coverage is randomly assigned, as in a clinical trial or insurance lottery, selection bias disappears. Suppose that D_i is determined by a coin toss: heads you're covered; tails you're not.

- By virtue of random assignment, the insured and uninsured in this experiment are similar in every way except their insurance status. Most importantly, they have the same *potential* outcomes:

$$E[Y_{1i}|D_i = 1] = E[Y_{1i}|D_i = 0]$$
$$E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$$

- Consequently,

$$\begin{aligned} &E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\ &\quad = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \quad (5) \\ &\quad = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \quad (6) \\ &\quad = E[Y_{1i} - Y_{0i}|D_i = 1] \end{aligned}$$

  The step from line (5) to line (6) justifies the central role of randomized trials in social science and clinical research: random assignment eliminates selection bias.

- In a simple RCT, we also have:

$$E[Y_{1i} - Y_{0i}|D_i = 1] = E[Y_{1i} - Y_{0i}].$$

  In a randomized experiment where everyone does what they're assigned to do, the average causal effect on the treated, $E[Y_{1i} - Y_{0i}|D_i = 1]$, is the same as the population average causal effect, $E[Y_{1i} - Y_{0i}]$.

  - This consequence of randomization is important, but not as important as the elimination of selection bias. (And more complicated experiments need not have the feature that ATE=TOT.)

2 A Healthy Debate

2.1 The RAND HIE

- Dateline 1974: Kung Fu enters its 3rd season; The RAND Health Insurance Experiment begins
  - In the 1970s, the RAND Corporation randomly assigned about 6,000 people (who agreed to drop their own insurance) to experimental insurance plans that offered either no cost-sharing, a modest deductible, or imposed 25%, 50% or 95% coinsurance rates on subscribers, capped at a maximum annual payment of $1000.
- The next table shows RAND descriptive statistics and checks for balance
  - We look at four groups defined by different levels and types of cost sharing: (i) the catastrophic coverage plan approximates an uninsured condition; the (ii) deductible and (iii) coinsurance plans provided partial coverage; (iv) free care is what it sounds like.

TABLE 1.3
Demographic characteristics and baseline health in the RAND HIE

	Means	Differences between plan groups			
	Catastrophic plan (1)	Deductible – catastrophic (2)	Coinsurance – catastrophic (3)	Free – catastrophic (4)	Any insurance – catastrophic (5)
A. Demographic characteristics					
Female	.560	−.023	−.025	−.038	−.030
		(.016)	(.015)	(.015)	(.013)
Nonwhite	.172	−.019	−.027	−.028	−.025
		(.027)	(.025)	(.025)	(.022)
Age	32.4	.56	.97	.43	.64
	[12.9]	(.68)	(.65)	(.61)	(.54)
Education	12.1	−.16	−.06	−.26	−.17
	[2.9]	(.19)	(.19)	(.18)	(.16)
Family income	31,603	−2,104	970	−976	−654
	[18,148]	(1,384)	(1,389)	(1,345)	(1,181)
Hospitalized last year	.115	.004	−.002	.001	.001
		(.016)	(.015)	(.015)	(.013)
B. Baseline health variables					
General health index	70.9	−1.44	.21	−1.31	−.93
	[14.9]	(.95)	(.92)	(.87)	(.77)
Cholesterol (mg/dl)	207	−1.42	−1.93	−5.25	−3.19
	[40]	(2.99)	(2.76)	(2.70)	(2.29)
Systolic blood pressure (mm Hg)	122	2.32	.91	1.12	1.39
	[17]	(1.15)	(1.08)	(1.01)	(.90)
Mental health index	73.8	−.12	1.19	.89	.71
	[14.3]	(.82)	(.81)	(.77)	(.68)
Number enrolled	759	881	1,022	1,295	3,198

Notes: This table describes the demographic characteristics and baseline health of subjects in the RAND Health Insurance Experiment (HIE). Column (1) shows the average for the group assigned catastrophic coverage. Columns (2)–(5) compare averages in the deductible, cost-sharing, free care, and any insurance groups with the average in column (1). Standard errors are reported in parentheses in columns (2)–(5); standard deviations are reported in brackets in column (1).

- What's this table suggest regarding the selection-bias term, $E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$ in the RAND experiment?

- Impact

TABLE 1.4
Health expenditure and health outcomes in the RAND HIE

	Means	Differences between plan groups			
	Catastrophic plan (1)	Deductible – catastrophic (2)	Coinsurance – catastrophic (3)	Free – catastrophic (4)	Any insurance – catastrophic (5)
A. Health-care use					
Face-to-face visits	2.78 [5.50]	.19 (.25)	.48 (.24)	1.66 (.25)	.90 (.20)
Outpatient expenses	248 [488]	42 (21)	60 (21)	169 (20)	101 (17)
Hospital admissions	.099 [.379]	.016 (.011)	.002 (.011)	.029 (.010)	.017 (.009)
Inpatient expenses	388 [2,308]	72 (69)	93 (73)	116 (60)	97 (53)
Total expenses	636 [2,535]	114 (79)	152 (85)	285 (72)	198 (63)
B. Health outcomes					
General health index	68.5 [15.9]	−.87 (.96)	.61 (.90)	−.78 (.87)	−.36 (.77)
Cholesterol (mg/dl)	203 [42]	.69 (2.57)	−2.31 (2.47)	−1.83 (2.39)	−1.32 (2.08)
Systolic blood pressure (mm Hg)	122 [19]	1.17 (1.06)	−1.39 (.99)	−.52 (.93)	−.36 (.85)
Mental health index	75.5 [14.8]	.45 (.91)	1.07 (.87)	.43 (.83)	.64 (.75)
Number enrolled	759	881	1,022	1,295	3,198

Notes: This table reports means and treatment effects for health expenditure and health outcomes in the RAND Health Insurance Experiment (HIE). Column (1) shows the average for the group assigned catastrophic coverage. Columns (2)–(5) compare averages in the deductible, cost-sharing, free care, and any insurance groups with the average in column (1). Standard errors are reported in parentheses in columns (2)–(5); standard deviations are reported in brackets in column (1).

- Unlike NHIS Table 1.1, the comparisons in Table 1.4 carry causal weight.
- We see little here to suggest insurance *causes* the insured to be healthier (though the RAND HIE was motivated more by Panel A than Panel B)

2.2 HI on the Oregon Trail

- Elderly Americans get publicly provide health insurance through Medicare, while many of the poor are covered by Medicaid
- In 2008, Oregon's Medicaid agency offered coverage to about 30,000 otherwise uninsured low-income adults who didn't qualify for Medicaid by the usual rules. These lucky 30,000 were chosen by lottery from about 75,000 applicants.
  - This just in from Portlandia ...

TABLE 1.5
OHP effects on insurance coverage and health-care use

	Oregon		Portland area	
Outcome	Control mean (1)	Treatment effect (2)	Control mean (3)	Treatment effect (4)
A. Administrative data				
Ever on Medicaid	.141	.256 (.004)	.151	.247 (.006)
Any hospital admissions	.067	.005 (.002)		
Any emergency department visit			.345	.017 (.006)
Number of emergency department visits			1.02	.101 (.029)
Sample size	74,922		24,646	
B. Survey data				
Outpatient visits (in the past 6 months)	1.91	.314 (.054)		
Any prescriptions?	.637	.025 (.008)		
Sample size	23,741			

Notes: This table reports estimates of the effect of winning the Oregon Health Plan (OHP) lottery on insurance coverage and use of health care. Odd-numbered columns show control group averages. Even-numbered columns report the regression coefficient on a dummy for lottery winners. Standard errors are reported in parentheses.

- Later, we'll learn that an OHP insurance lottery offer dummy is an *instrumental variable* (IV) for insurance coverage, with *first-stage* effect given by $\hat{\pi} = .256$ (in column 2 above) and an IV estimate of TOT equal to the *reduced-form* estimates in column 4 divided by this
- Hey, where's my health divided?

TABLE 1.6
OHP effects on health indicators and financial health

	Oregon		Portland area	
Outcome	Control mean (1)	Treatment effect (2)	Control mean (3)	Treatment effect (4)
A. Health indicators				
Health is good	.548	.039 (.008)		
Physical health index			45.5	.29 (.21)
Mental health index			44.4	.47 (.24)
Cholesterol			204	.53 (.69)
Systolic blood pressure (mm Hg)			119	−.13 (.30)
B. Financial health				
Medical expenditures >30% of income			.055	−.011 (.005)
Any medical debt?			.568	−.032 (.010)
Sample size	23,741		12,229	

Notes: This table reports estimates of the effect of winning the Oregon Health Plan (OHP) lottery on health indicators and financial health. Odd-numbered columns show control group averages. Even-numbered columns report the regression coefficient on a dummy for lottery winners. Standard errors are reported in parentheses.

- Health insurance makes household *finances* healthier
- As in Table 1.4, the statistical significance of reduced health expenditures in Panel B supports causal conclusions: that's the miracle of random assignment

- Masters of 'metrics know to distinguish between:
  - random sampling, which facilitates statistical inference about population parameters, causal or otherwise
  - random assignment, which supports causal inference, that is, comparisons of potential outcomes free of selection bias

2.3 Capturing Causal effects Without Random Assignment

Randomized research designs represent an often-unattainable ideal. Masters of 'metrics therefore develop and implement empirical methods that reduce or eliminate selection bias in settings where RCTs are prohibitively expensive, time-consuming, impractical, or unethical.

- Even so, the experimental ideal disciplines our thinking. The first question a researcher considers is always thus:

  What's the experiment you'd like to do?