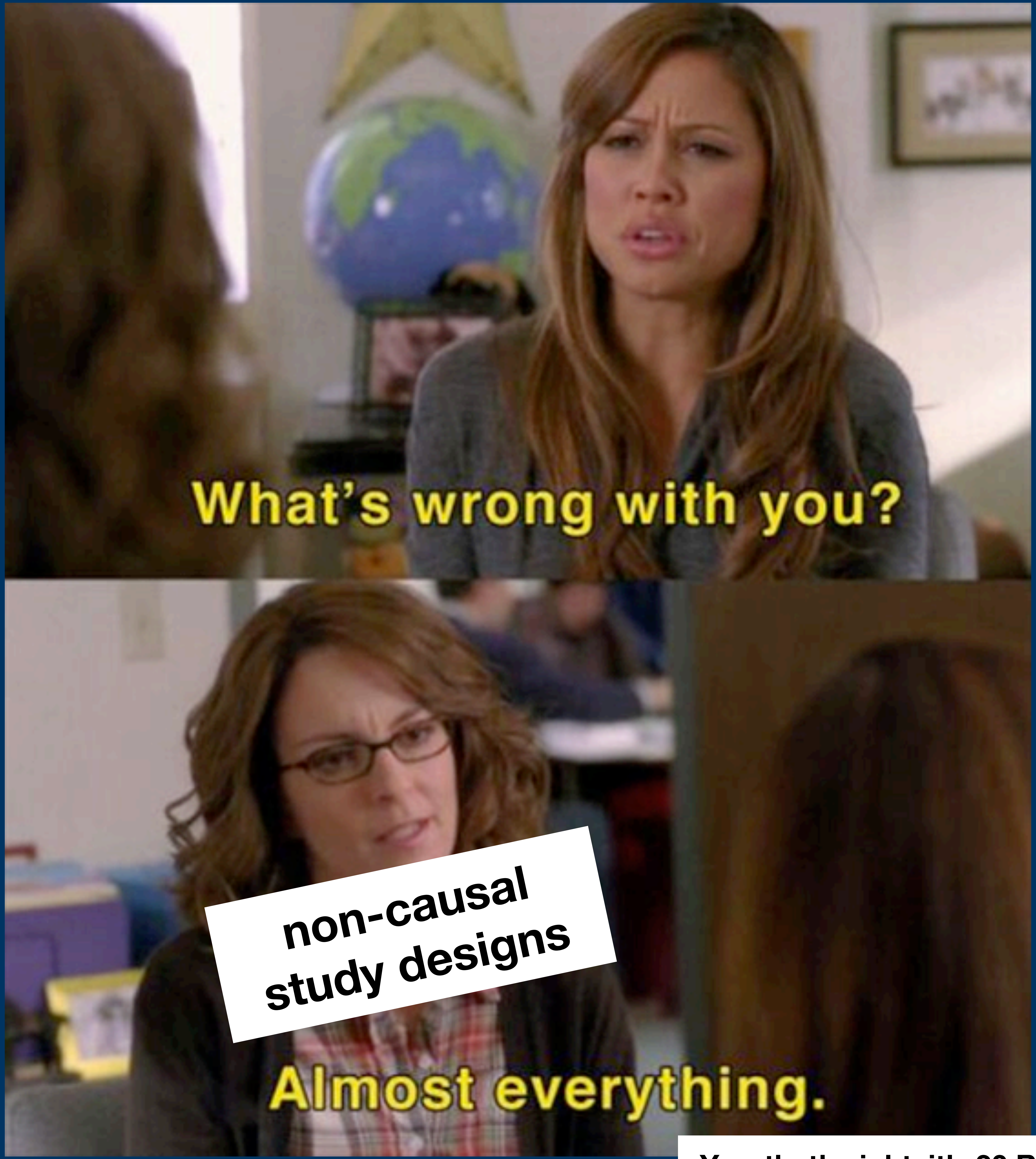
Let's talk about ex-periments, baby

API 202: TF Session 4

ALL

Nolan M. Kavanagh
February 16, 2024



Yes, that's right, it's 30 Rock day.

Goals for today

1. Review the benefits of randomization.
2. Discuss bias in randomized experiments.
3. Experience the magic of difference-in-differences.
4. Practice interpreting difference-in-differences.

Overview of our sample data

Dataset of over 2 million U.S. adults from 2011–2019

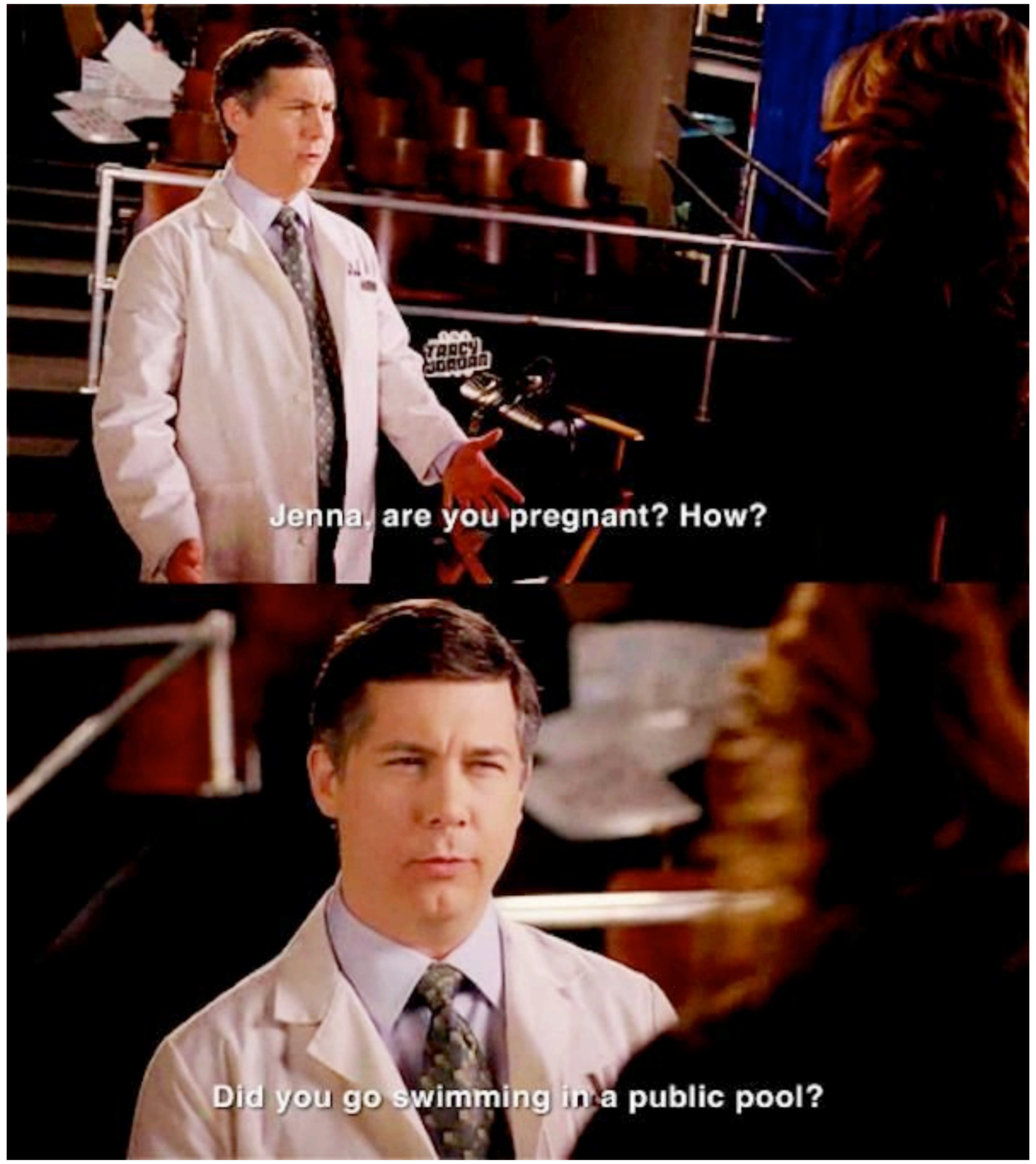
`state`	**State of respondent**	*Behavioral Risk Factor Surveillance System*
`year`	**Year when surveyed**	*Behavioral Risk Factor Surveillance System*
`age`	**Dummy for under 50 (1) or not (0)**	*Behavioral Risk Factor Surveillance System*
`gender`	**Dummy for man (1) or woman (0)**	*Behavioral Risk Factor Surveillance System*
`race_eth`	**Dummy for white/non-Latin (1) or not (0)**	*Behavioral Risk Factor Surveillance System*
`married`	**Dummy for married (1) or not (0)**	*Behavioral Risk Factor Surveillance System*
`education`	**Dummy for college-educated (1) or not (0)**	*Behavioral Risk Factor Surveillance System*
`income`	**Dummy for income <$35,000 (1) or not (0)**	*Behavioral Risk Factor Surveillance System*
`insurance`	**Dummy for health insurance (1) or not (0)**	*Behavioral Risk Factor Surveillance System*
`expansion`	**Dummy for Medicaid expansion state (1) or not (0)**	*Administrative*
`post_2014`	**Dummy for 6/1/2014 onward (1) or pre-2014 (0)**	*Administrative*

I have an idea!

About 9% of Americans don’t have health insurance.

If they had insurance, they might get healthier.

$$\text{health insurance} \xrightarrow{?} \text{health status}$$



(Disclaimer: No, it doesn’t work that way.)

Why can’t we just compare insured and uninsured folks?

What are we worried about?

Say it with me!

Why can’t we just compare insured and uninsured folks?

There are, uh, a few omitted variables.

	Uninsured (n=213,714)	Insured (n=2,039,416)	Difference	t-test (P-value)
Percent under 50 years	60.9%	33.2%	+27.7 pp	P<0.001
Percent men	48.5%	43.0%	+5.5 pp	P<0.001
Percent non-Latin white	58.8%	80.3%	–21.5 pp	P<0.001
Percent married	38.5%	56.0%	–17.5 pp	P<0.001
Percent college-educated	16.4%	39.0%	–22.6 pp	P<0.001
Percent with income <$35,000	71.3%	35.7%	+35.6 pp	P<0.001

Omitted variable bias will haunt your dreams.

Let's say that in our short regression, having insurance = better health.

$\hat{\alpha}_1 > 0$

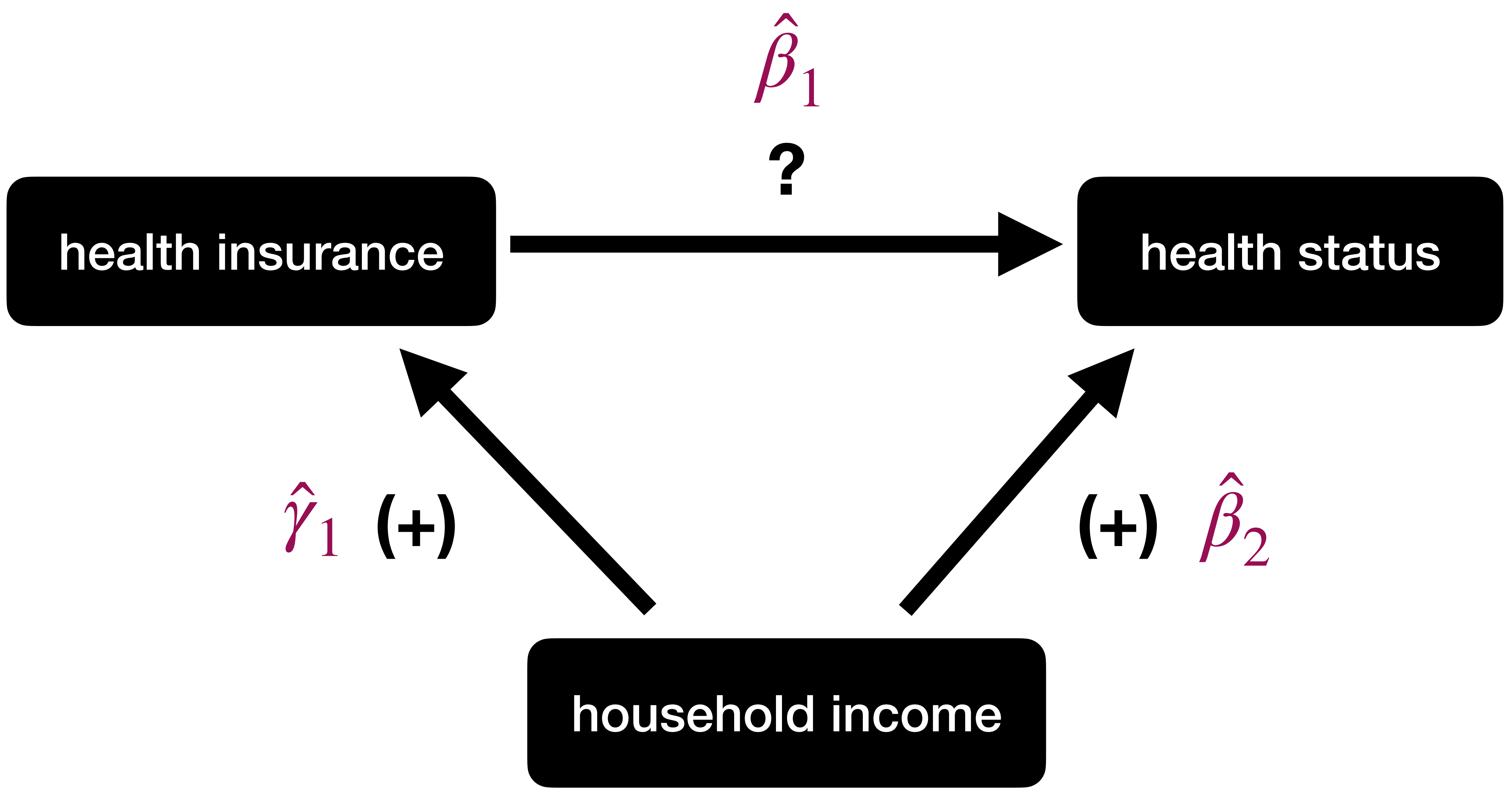
What might be an omitted variable?

Omitted variable bias will haunt your dreams.

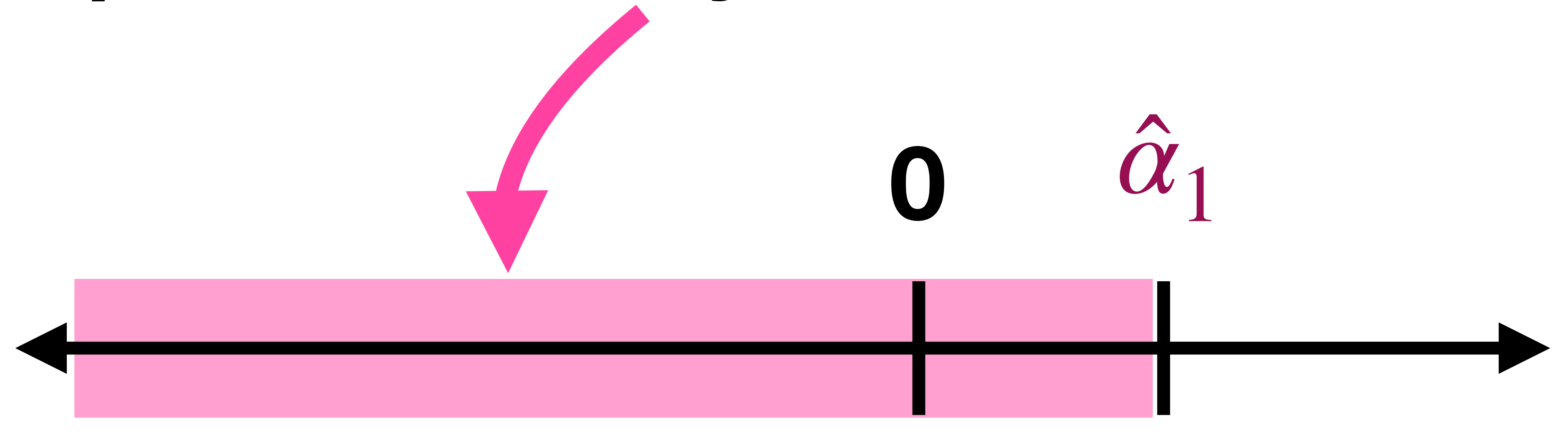
Let's say that in our short regression, having insurance = better health.

$\hat{\alpha}_1 > 0$

What might be an omitted variable?



β_1 could be <u>anywhere</u> in here.



Our bias is positive, so α_1 must be to the right of β_1.

Bias formula

$$\alpha_1 - \beta_1 = \beta_2 * \gamma_1 = (+)(+) = (+)$$

Let’s just randomize insurance!

Randomly pick 10,000 uninsured folks & split them into two groups.

	Uninsured (n=213,714)	Group A (n=5,000)	Group B (n=5,000)	Difference (A – B)	t-test (P-value)
Percent under 50 years	60.9%	62.0%	60.1%	+1.9 pp	P=0.05
Percent men	48.5%	48.6%	49.0%	–0.4 pp	P=0.70
Percent non-Latin white	58.8%	57.9%	58.0%	–0.1 pp	P=0.98
Percent married	38.5%	37.1%	39.4%	–2.3 pp	P=0.02
Percent college-educated	16.4%	15.9%	16.3%	–0.4 pp	P=0.57
Percent with income <$35,000	71.3%	71.3%	70.6%	+0.7 pp	P=0.44

Sometimes we’ll get “significant” values just due to chance. That’s OK!


Ha ha!
High-fiving a million angels.

Someone actually did that!

The Oregon Health Insurance Experiment made an insurance lottery.

Lottery winners were *eligible to enroll* in Medicaid. But not all did.

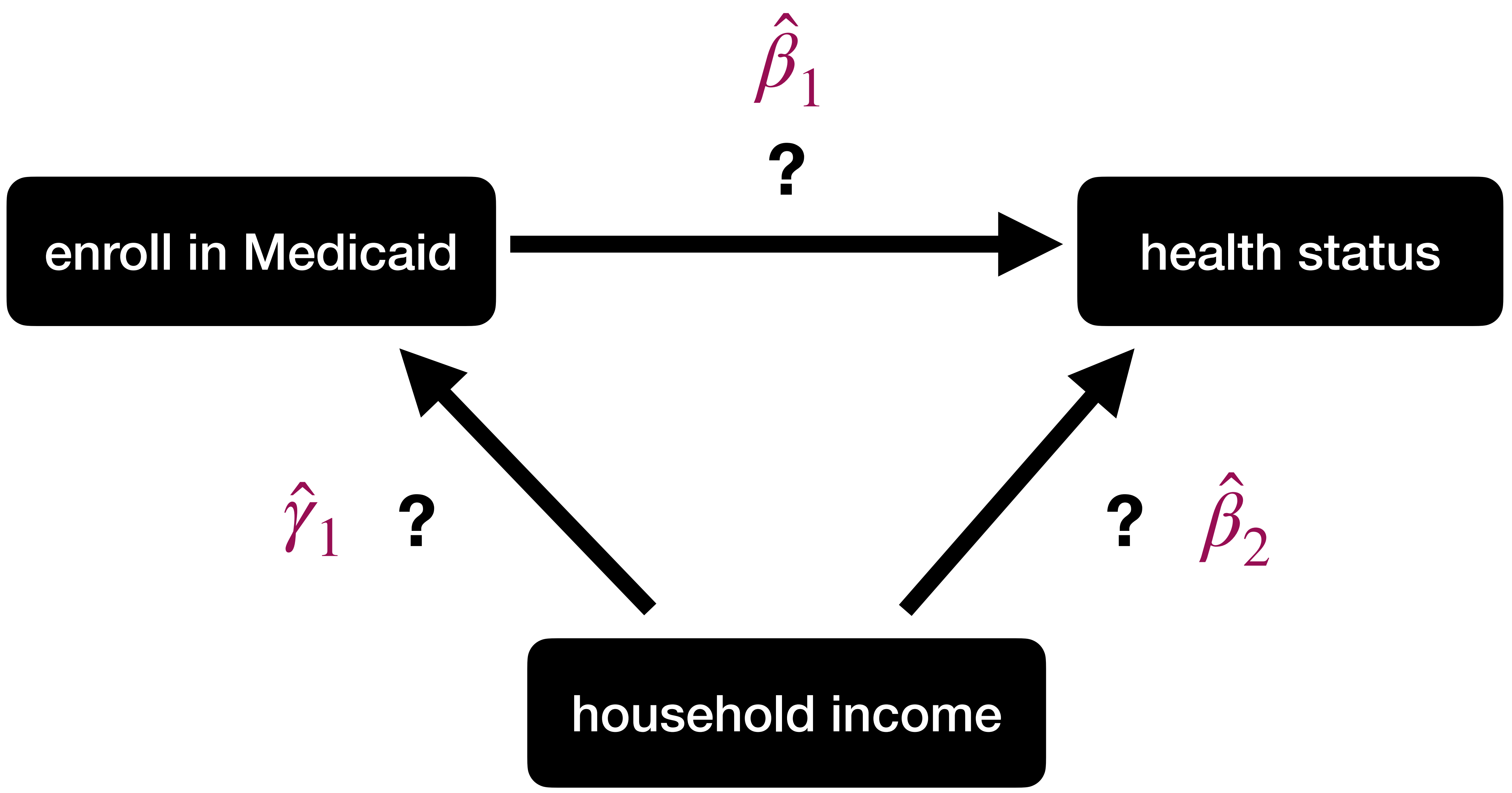
When we analyze our experiment, whom should we compare?

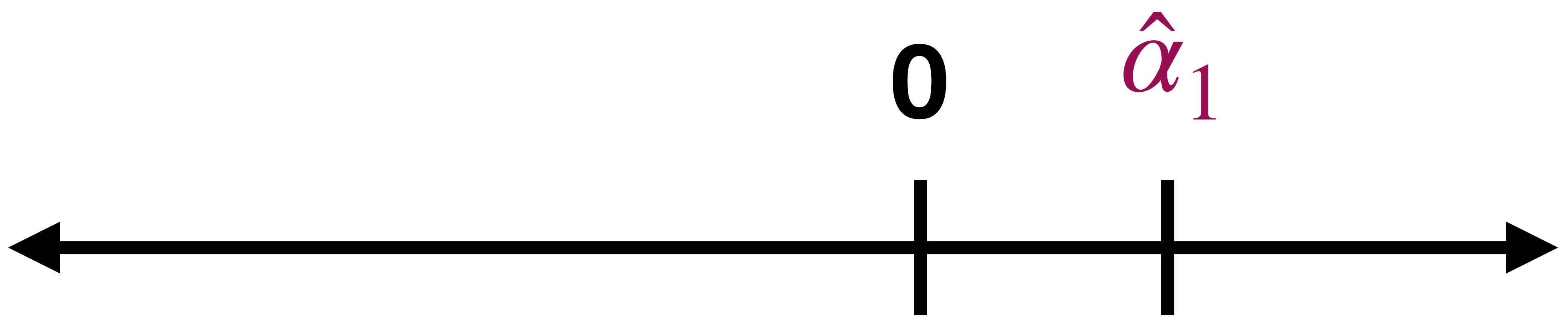
Let’s say we just analyze enrollees.

In our short regression, we find that Medicaid = better health.

$\hat{\alpha}_1 > 0$

What if only “richer” people enrolled?





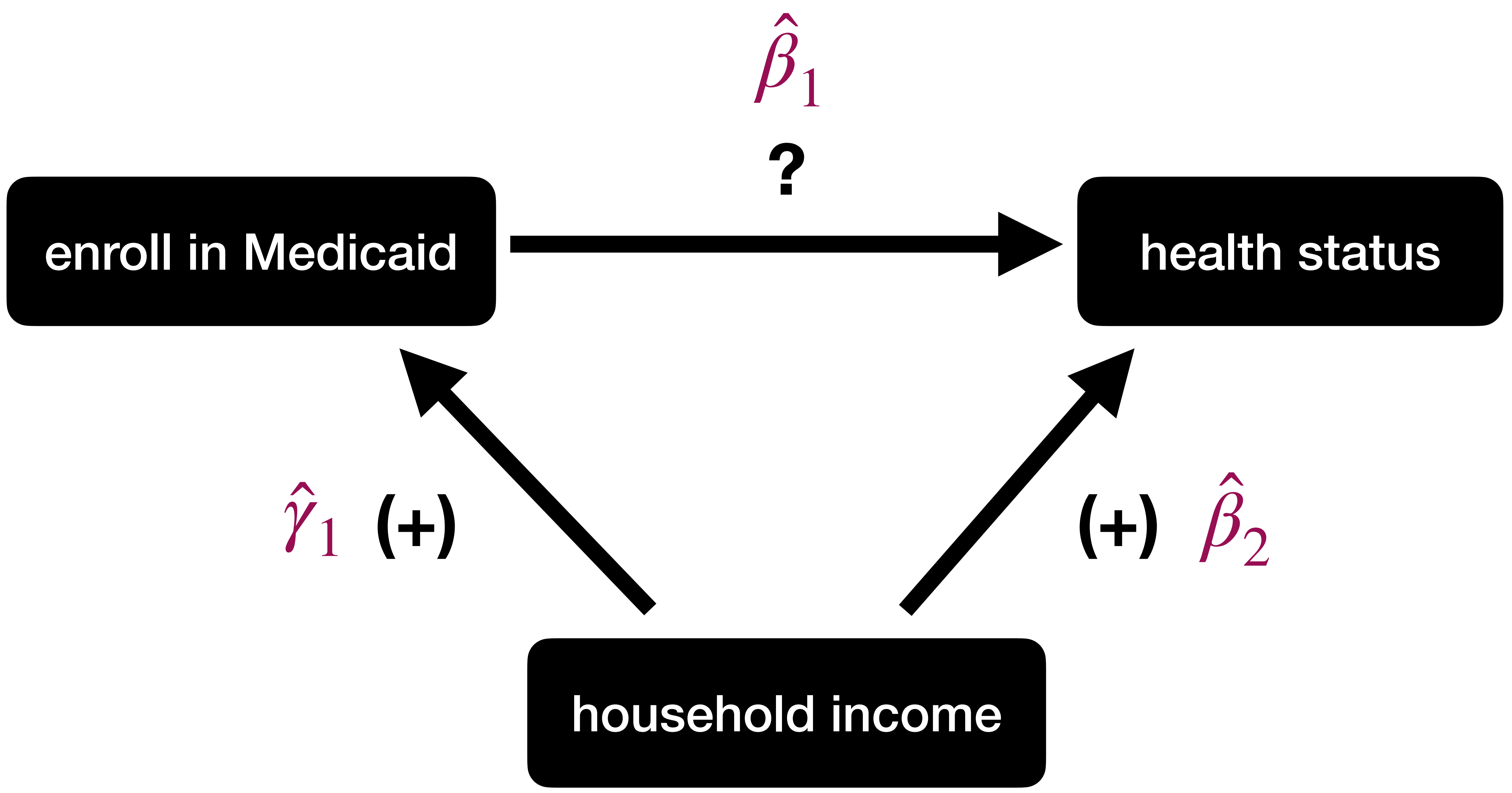
Bias formula $\alpha_1 - \beta_1 = \beta_2 * \gamma_1 =$

Let's say we just analyze enrollees.

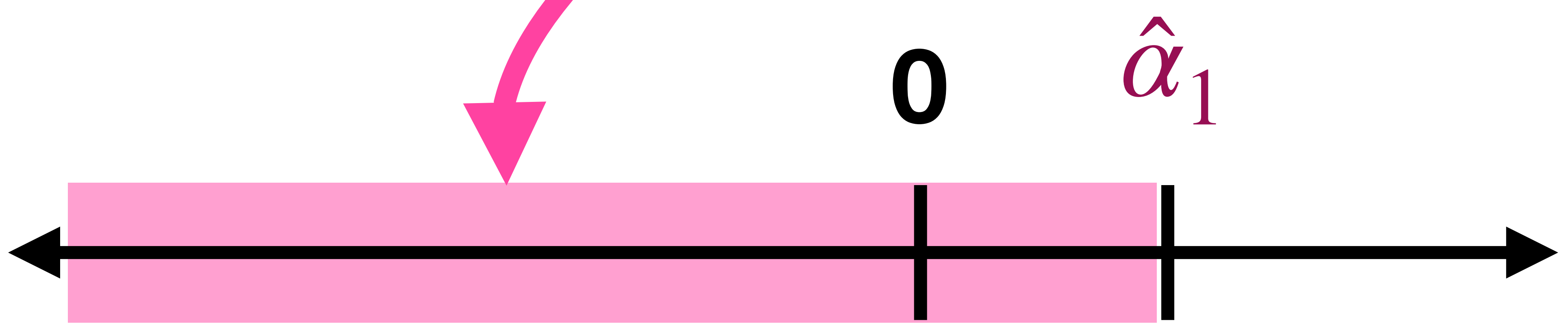
In our short regression, we find that Medicaid = better health.

$\hat{\alpha}_1 > 0$

What if only "richer" people enrolled?



β_1 could be <u>anywhere</u> in here.



Our bias is positive, so α_1 must be to the right of β_1.

Bias formula

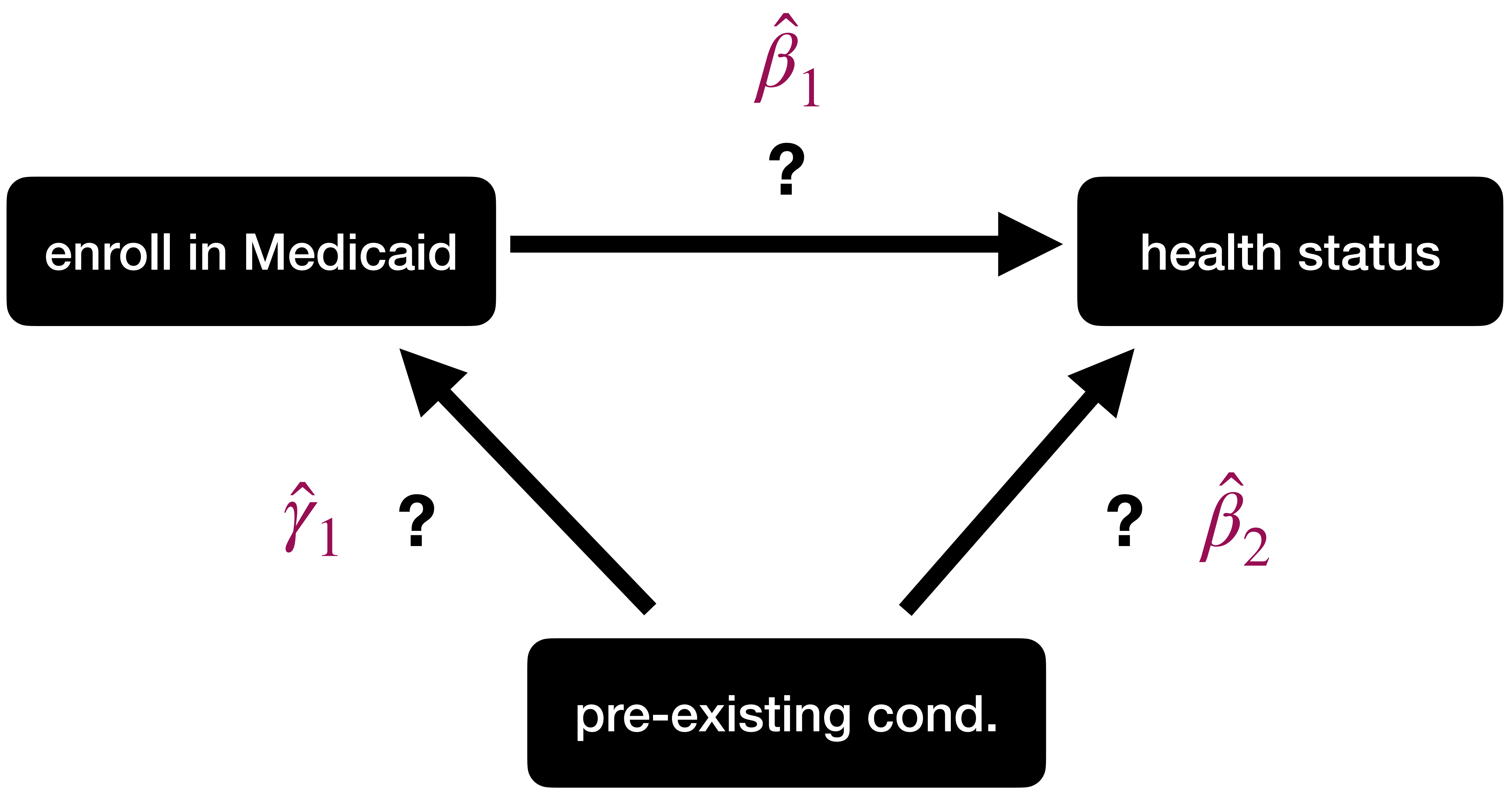
$$\alpha_1 - \beta_1 = \beta_2 * \gamma_1 = (+)(+) = (+)$$

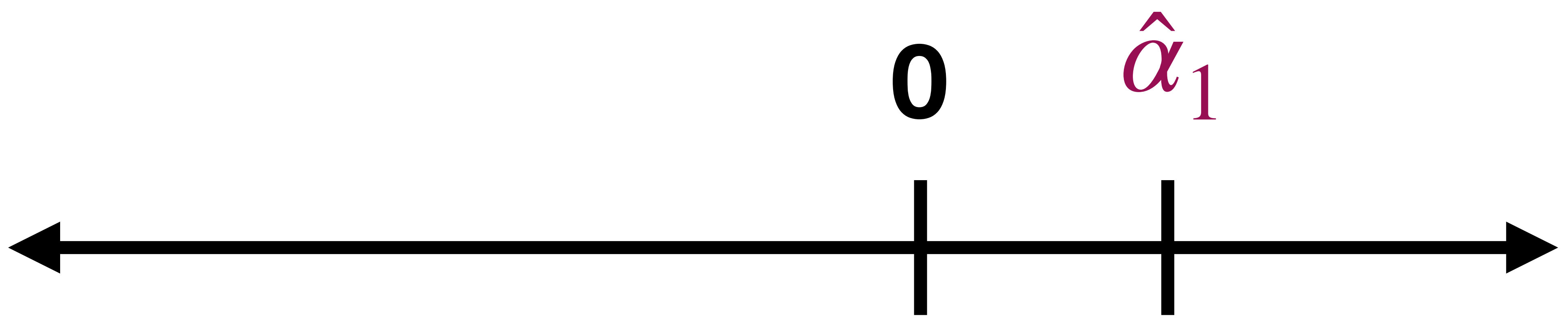
Let's say we just analyze enrollees.

In our short regression, we find that Medicaid = better health.

$\hat{\alpha}_1 > 0$

What if only "sicker" people enrolled?





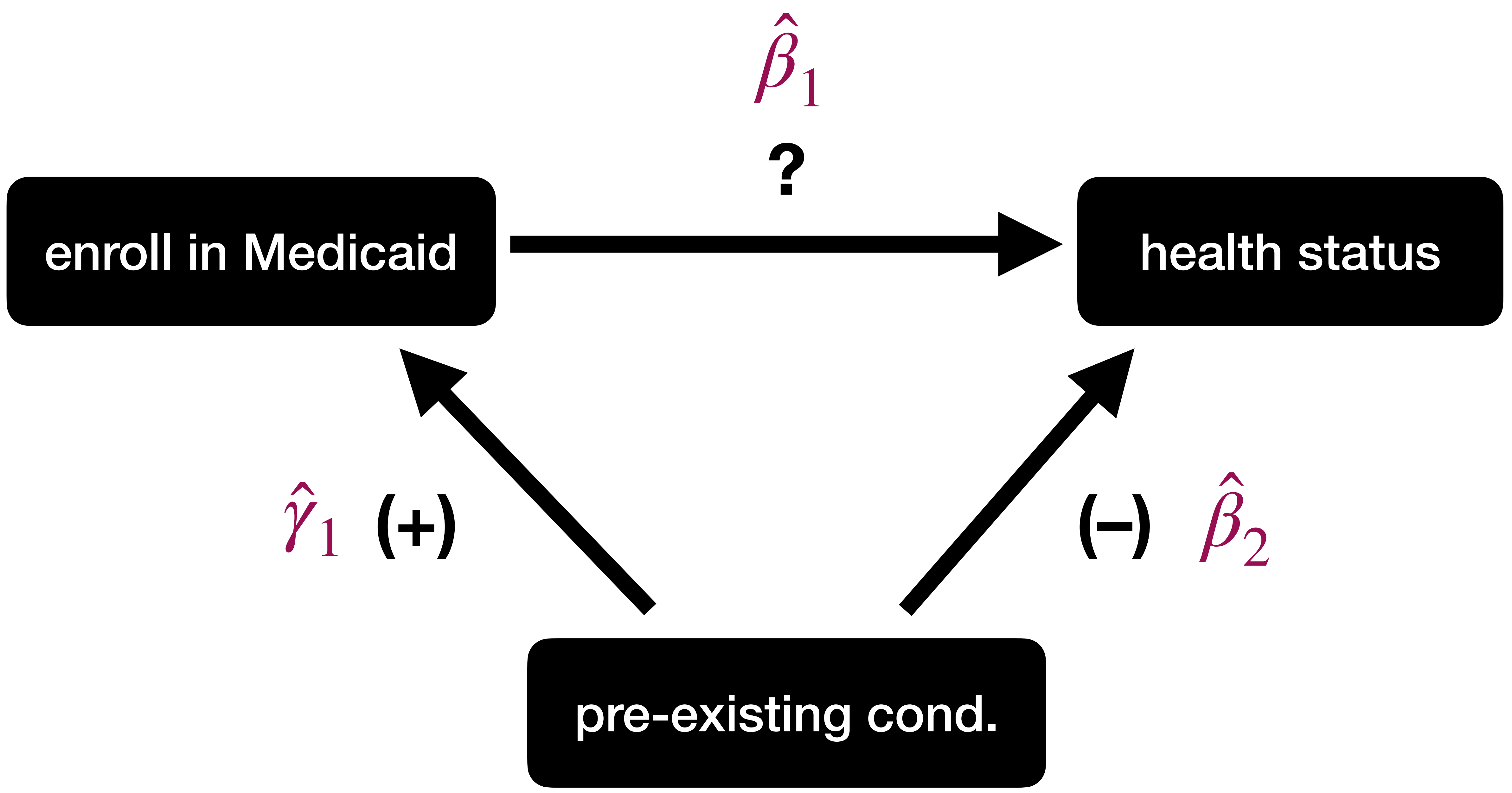
Bias formula $\alpha_1 - \beta_1 = \beta_2 * \gamma_1 =$

Let’s say we just analyze enrollees.

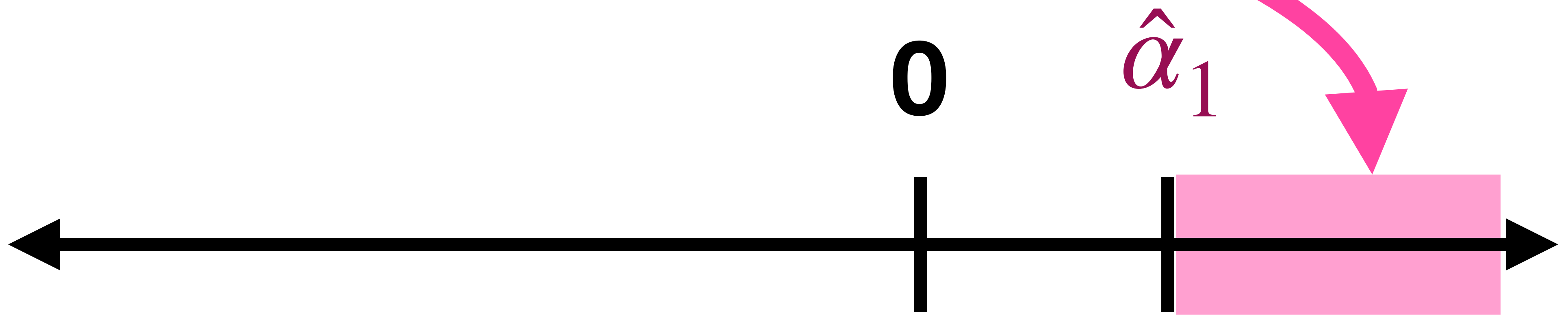
In our short regression, we find that Medicaid = better health.

$\hat{\alpha}_1 > 0$

What if only “richer” people enrolled?



β_1 could be <u>anywhere</u> over here.



Our bias is negative, so α_1 must be to the left of β_1.

Bias formula

$$\alpha_1 - \beta_1 = \beta_2 * \gamma_1 = (+)(-) = (-)$$

This is why intention to treat matters.

We only have “balance” in our original, randomized sample.

Analyzing just enrollees breaks this balance since we now have a different sample that people selected into based on omitted variable(s).

Intention-to-treat analyses preserve the original randomization.

Can experiments have other biases?

Yes!

We are especially concerned about:

1. **Attrition** (hence, intention to treat)

2. **Failures in randomization** (this is like an omitted variable!)

3. **Spillover** (also like an omitted variable!)

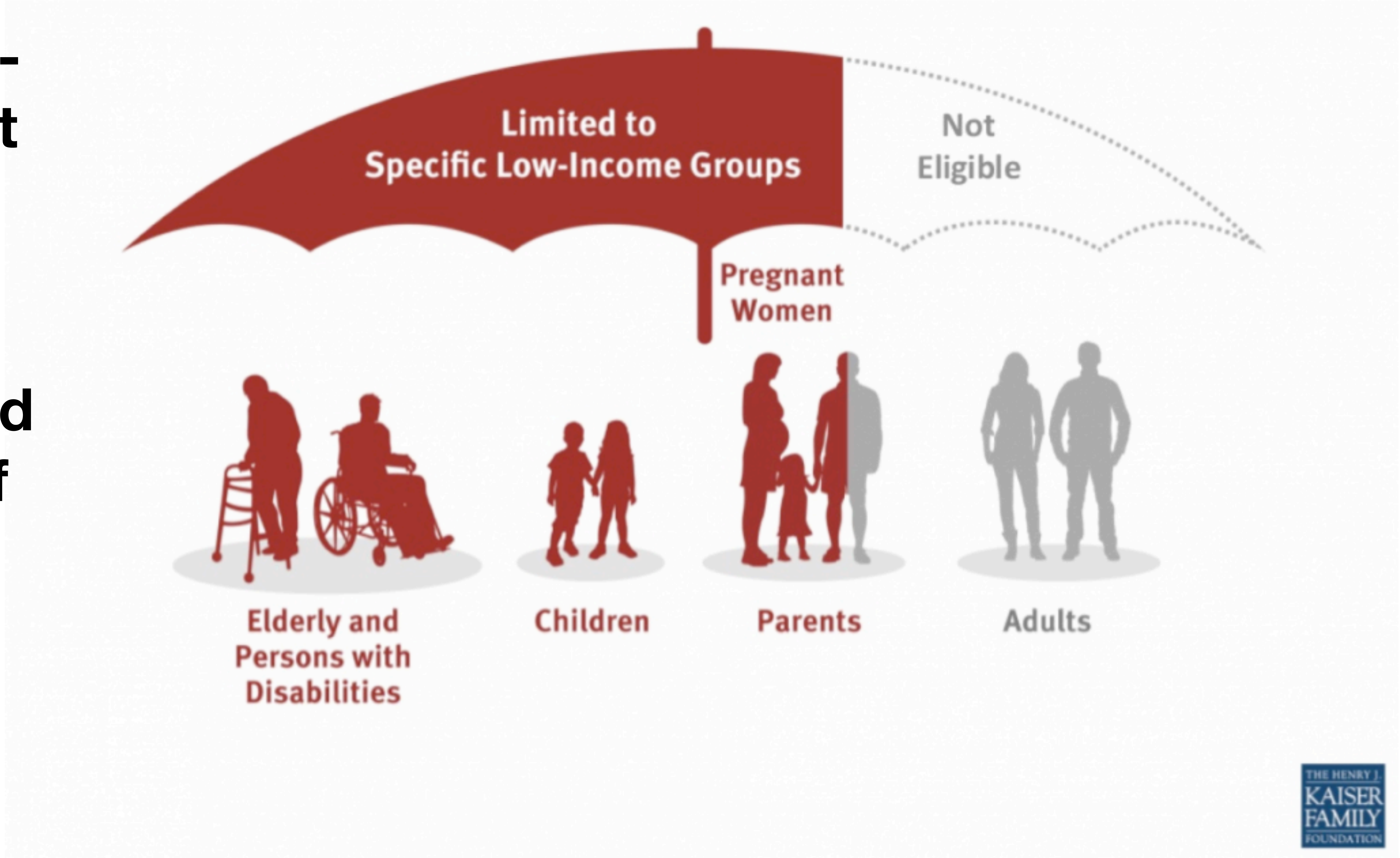
Also, experiments can be expensive, impractical, or unethical.

What about a “natural experiment”?

Medicaid used to be a state-level insurance program just for specific populations.

In 2014, the Affordable Care Act allowed states to expand it to everyone up to 138% of the federal poverty level.

In 2024, 138% FPL for a family of 4 is $43,056.

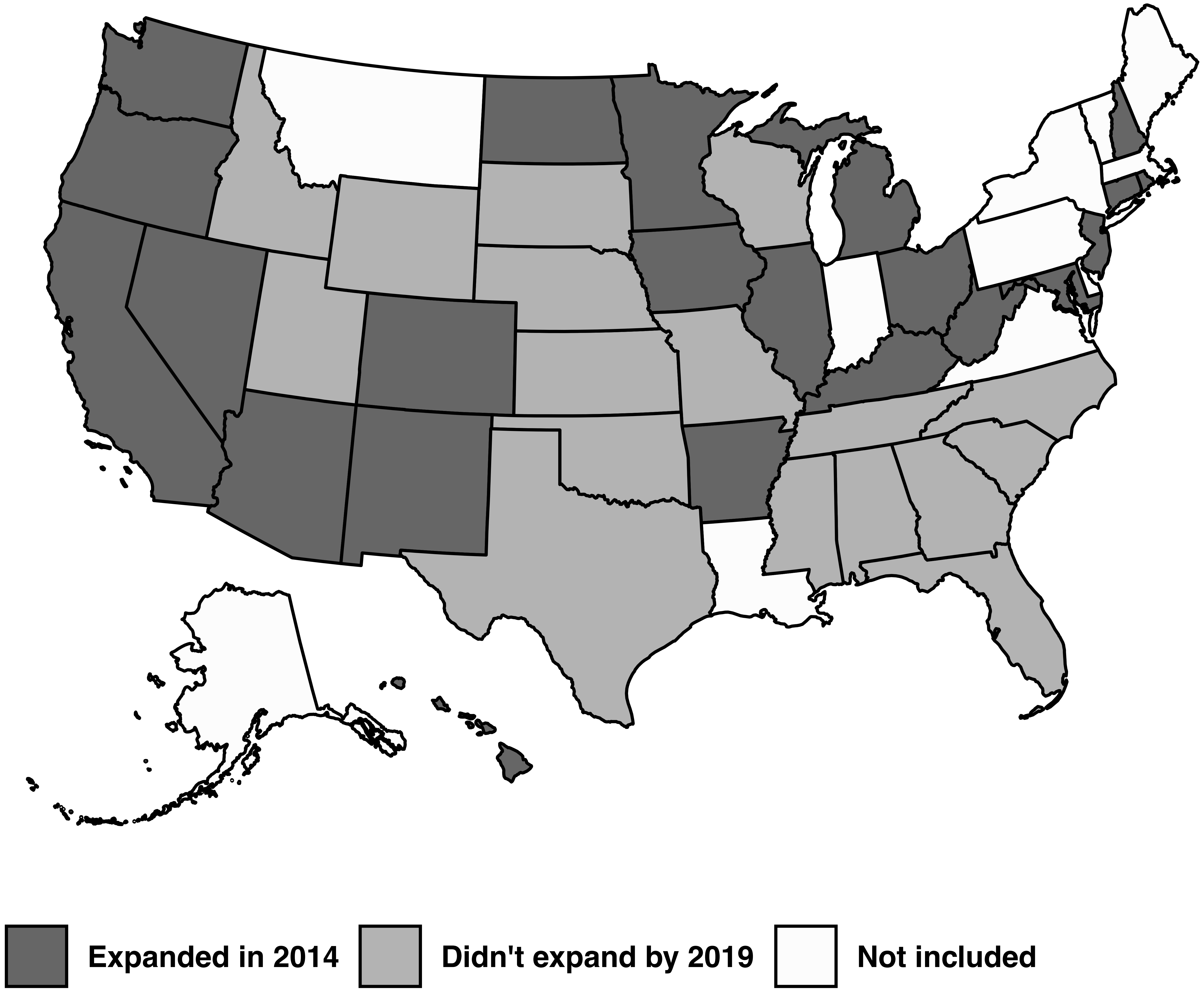


What about a “natural experiment”?

However, not every state chose to expand Medicaid.

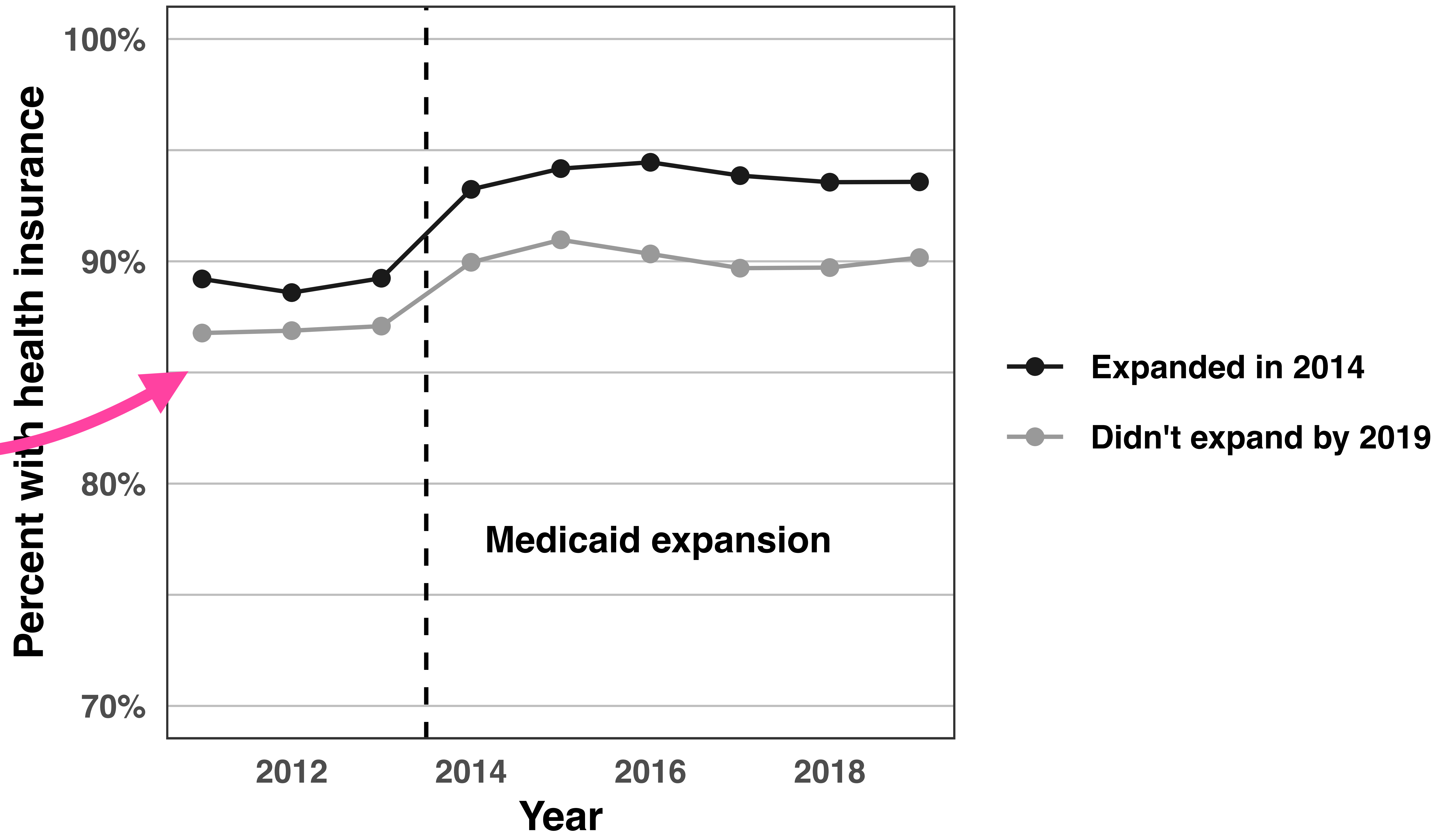
We might call this a “natural experiment” and evaluate it as if it were randomized*!

*Each state’s choice to expand isn’t really random, but if our difference-in-difference assumptions hold, we can pretend like it is!



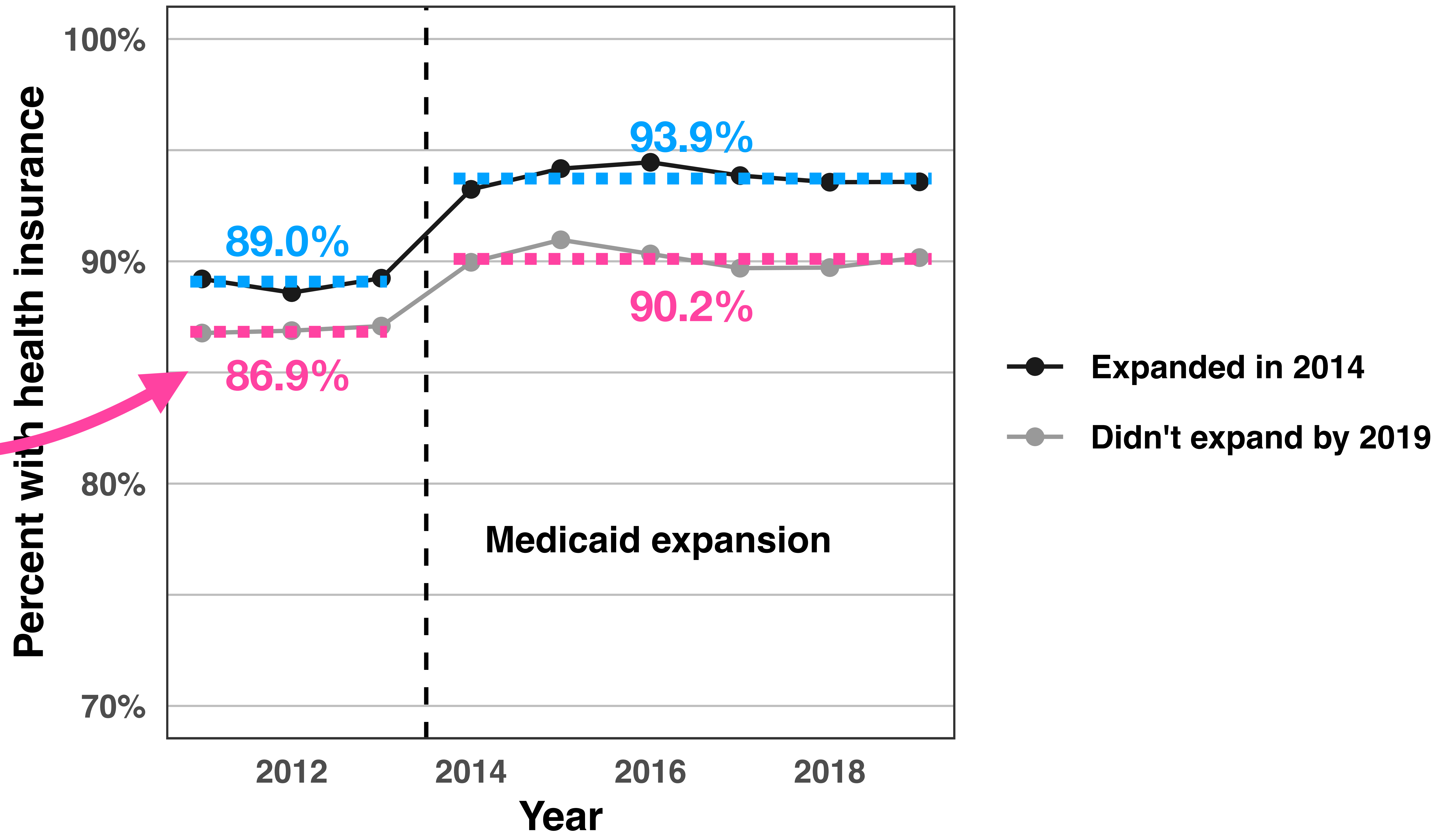
Let’s do a difference-in-differences!

The identifying assumption of a diff-in-diff is parallel trends, i.e. that each group of states has a similar pre-expansion trajectory.



Let’s do a difference-in-differences!

The identifying assumption of a diff-in-diff is parallel trends, i.e. that each group of states has a similar pre-expansion trajectory.



Let's do a difference-in-differences!

$$(insurance)_i = \beta_0 + \beta_1(expansion)_i + \beta_2(post\_2014)_i + \beta_3(expansion * post\_2014)_i + u_i$$

dummy for expansion
1 = Expanded in 2014
0 = Didn't expand by 2019

dummy for time
1 = June 1, 2014 or later
0 = Before 2014

interaction
i.e. the diff-in-diff

	Before 2014 `post_2014 = 0`	**After 6/1/2014** `post_2014 = 1`	**Difference**
Didn't expand by 2019 `expansion = 0`			
Expanded in 2014 `expansion = 1`			
Difference			

Let's do a difference-in-differences!

$$(insurance)_i = \beta_0 + \beta_1(expansion)_i + \beta_2(post\_2014)_i + \beta_3(expansion * post\_2014)_i + u_i$$

dummy for expansion
1 = Expanded in 2014
0 = Didn't expand by 2019

dummy for time
1 = June 1, 2014 or later
0 = Before 2014

interaction
i.e. the diff-in-diff

	Before 2014 `post_2014 = 0`	**After 6/1/2014** `post_2014 = 1`	**Difference**
Didn't expand by 2019 `expansion = 0`	β_0	$\beta_0 + \beta_2$	β_2
Expanded in 2014 `expansion = 1`	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
Difference	β_1	$\beta_1 + \beta_3$	β_3

What omitted variables are we eliminating?

	Before 2014 `post_2014 = 0`	After 6/1/2014 `post_2014 = 1`	Difference
Didn't expand by 2019 `expansion = 0`	β_0	$\beta_0 + \beta_2$	β_2
Expanded in 2014 `expansion = 1`	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
Difference	β_1	$\beta_1 + \beta_3$	β_3

What omitted variables are we eliminating?

What's left? State-variant omitted variables.
e.g. changes in state budgets or politics

state-invariant omitted variables
e.g. state demographics

time-varying omitted variables
e.g. national economic trends

	Before 2014 `post_2014 = 0`	**After 6/1/2014** `post_2014 = 1`	**Difference**
Didn't expand by 2019 `expansion = 0`	β_0	$\beta_0 + \beta_2$	β_2
Expanded in 2014 `expansion = 1`	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
Difference	β_1	$\beta_1 + \beta_3$	β_3

Show me the regression already!

	insurance
Intercept	**0.869*****
	(0.007)
expansion	**0.021***
	(0.009)
post_2014	**0.033*****
	(0.003)
expansion * post_2014	**0.016****
	(0.005)
Num.Obs.	2,253,130
R2	0.008
R2 Adj.	0.008

*p<0.05, **p<0.01, ***p<0.001

Show me the regression already!

	insurance
Intercept	**0.869*** **
	(0.007)
`expansion`	**0.021***
	(0.009)
`post_2014`	**0.033*** **
	(0.003)
`expansion * post_2014`	**0.016** **
	(0.005)
Num.Obs.	2,253,130
R2	0.008
R2 Adj.	0.008

*p<0.05, **p<0.01, ***p<0.001

The insurance rate in the non-expansion states was 86.9% before 2014.

(Note: Since we're comparing groups within the interaction, we don't have to say "holding time constant" since that's necessarily implied by the interaction terms.)

The difference is statistically significant.

Show me the regression already!

	insurance
Intercept	**0.869*** **
	(0.007)
expansion	**0.021***
	(0.009)
post_2014	**0.033*** **
	(0.003)
expansion * post_2014	**0.016** **
	(0.005)
Num.Obs.	2,253,130
R2	0.008
R2 Adj.	0.008

*p<0.05, **p<0.01, ***p<0.001

The difference in insurance rates between expansion and non-expansion states was 2.1 percentage points <u>before</u> 2014.

The difference is statistically significant.

Show me the regression already!

	insurance
Intercept	**0.869*** **
	(0.007)
expansion	**0.021***
	(0.009)
post_2014	**0.033*** **
	(0.003)
expansion * post_2014	**0.016** **
	(0.005)
Num.Obs.	2,253,130
R2	0.008
R2 Adj.	0.008

*p<0.05, **p<0.01, ***p<0.001

The difference in insurance rates for non-expansion states was 3.3 pp when comparing after 2014 to before 2014.

The difference is statistically significant.

Show me the regression already!

	insurance
Intercept	**0.869*****
	(0.007)
expansion	**0.021***
	(0.009)
post_2014	**0.033*****
	(0.003)
expansion * post_2014	**0.016****
	(0.005)
Num.Obs.	2,253,130
R2	0.008
R2 Adj.	0.008

*p<0.05, **p<0.01, ***p<0.001

The difference-in-differences for insurance rates was 1.6 percentage points.

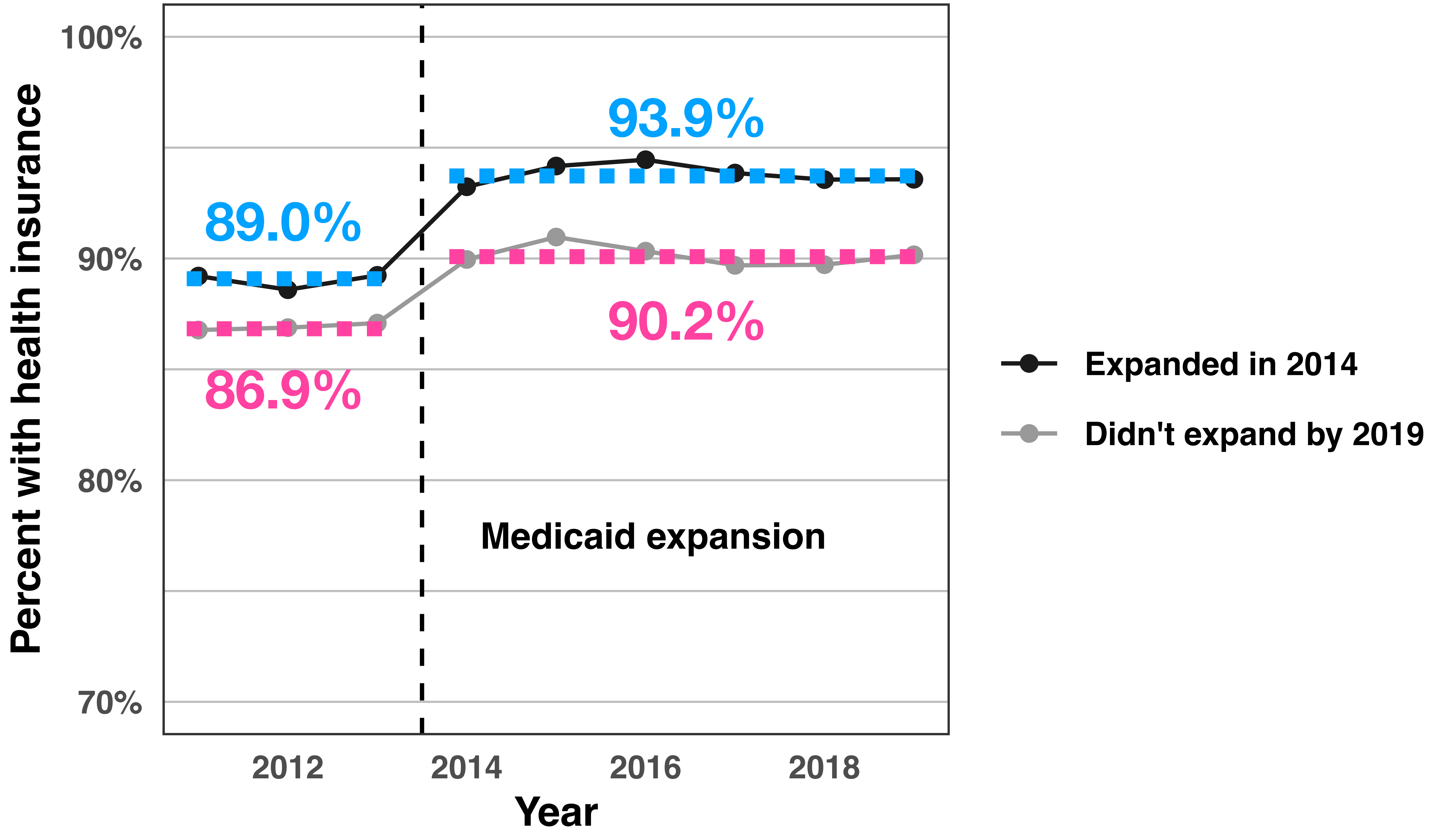
(That is, the difference in the change in insurance rates for expansion states, compared to the difference in change for non-expansion states, was 1.6 pp.)

The difference is statistically significant.

Show me the regression already!

	insurance
Intercept	**0.869*** **
	(0.007)
expansion	**0.021***
	(0.009)
post_2014	**0.033*** **
	(0.003)
expansion * post_2014	$\boxed{\textbf{0.016**}}$
	$\boxed{(0.005)}$
Num.Obs.	2,253,130
R2	0.008
R2 Adj.	0.008

*p<0.05, **p<0.01, ***p<0.001



$$(insurance)_i = \beta_0 + \beta_1(expansion)_i + \beta_2(post\_2014)_i + \beta_3(expansion * post\_2014)_i + u_i$$

But is it causal?

Again, causality is a spectrum.

We have to evaluate the assumptions.

But here, I would say yes!

…unless we can think of “killer” state-variant omitted variables.