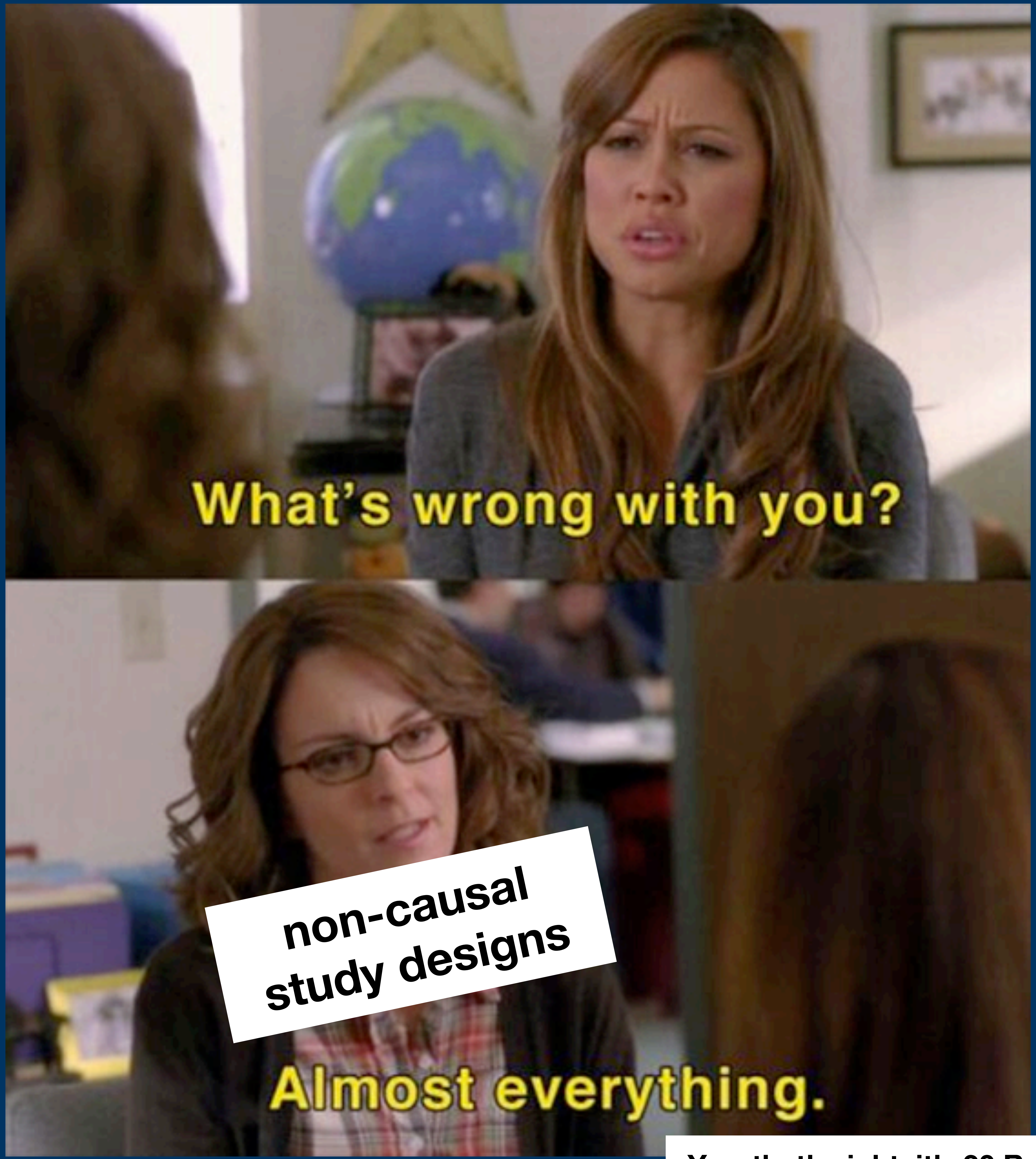
Experiments. Blergh!

API 202: TF Session 4

ALL

Nolan M. Kavanagh
February 20, 2026



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

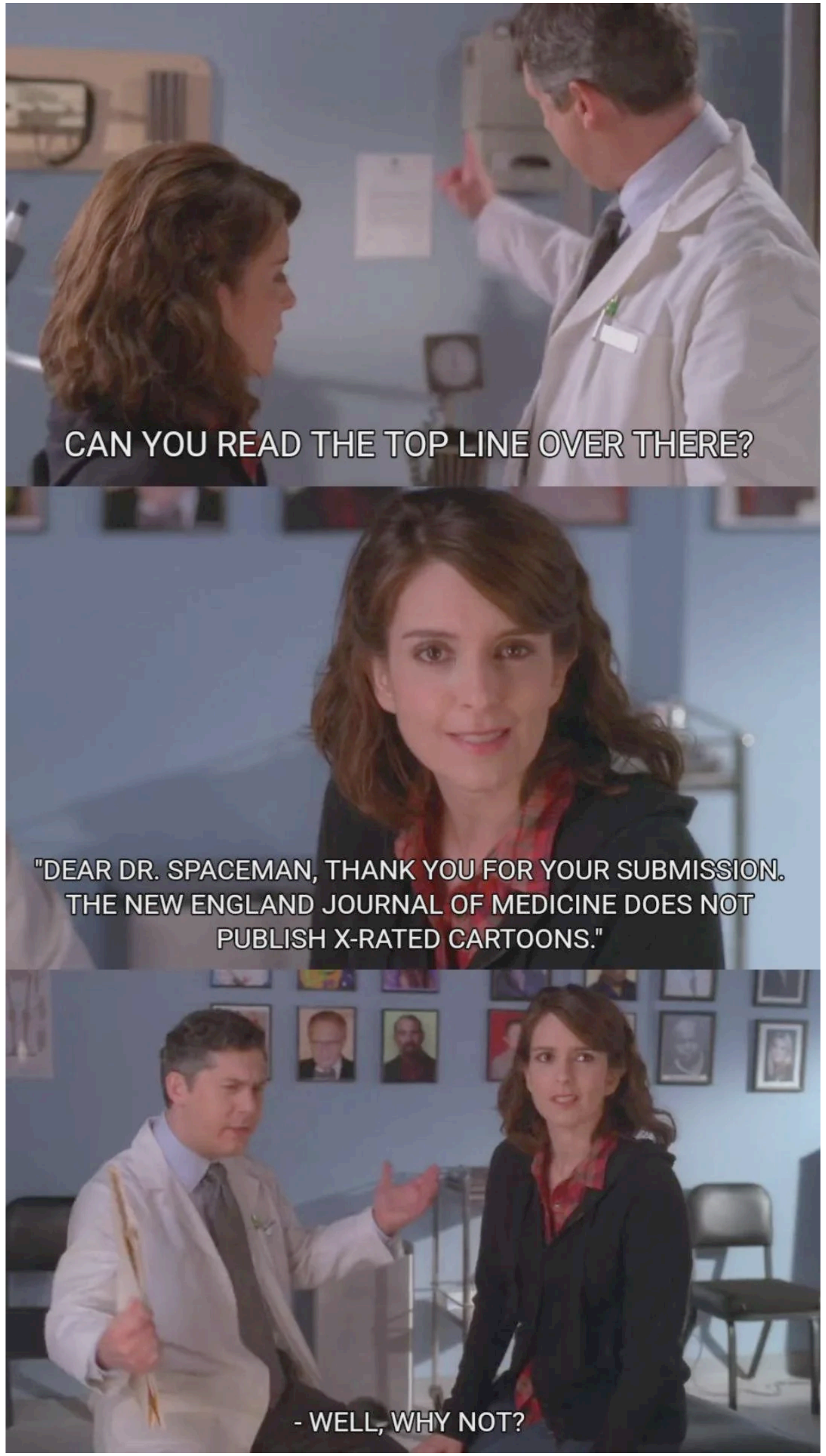
`state`	**State of respondent**	*Behavioral Risk Factor Surveillance System*
`year`	**Year when surveyed**	*Behavioral Risk Factor Surveillance System*
`age`	**Dummy for under 50 (1) or not (0)**	*Behavioral Risk Factor Surveillance System*
`gender`	**Dummy for man (1) or woman (0)**	*Behavioral Risk 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.

health insurance $\xrightarrow{?}$ health status



I, too, have been rejected by the New England Journal of Medicine.

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.

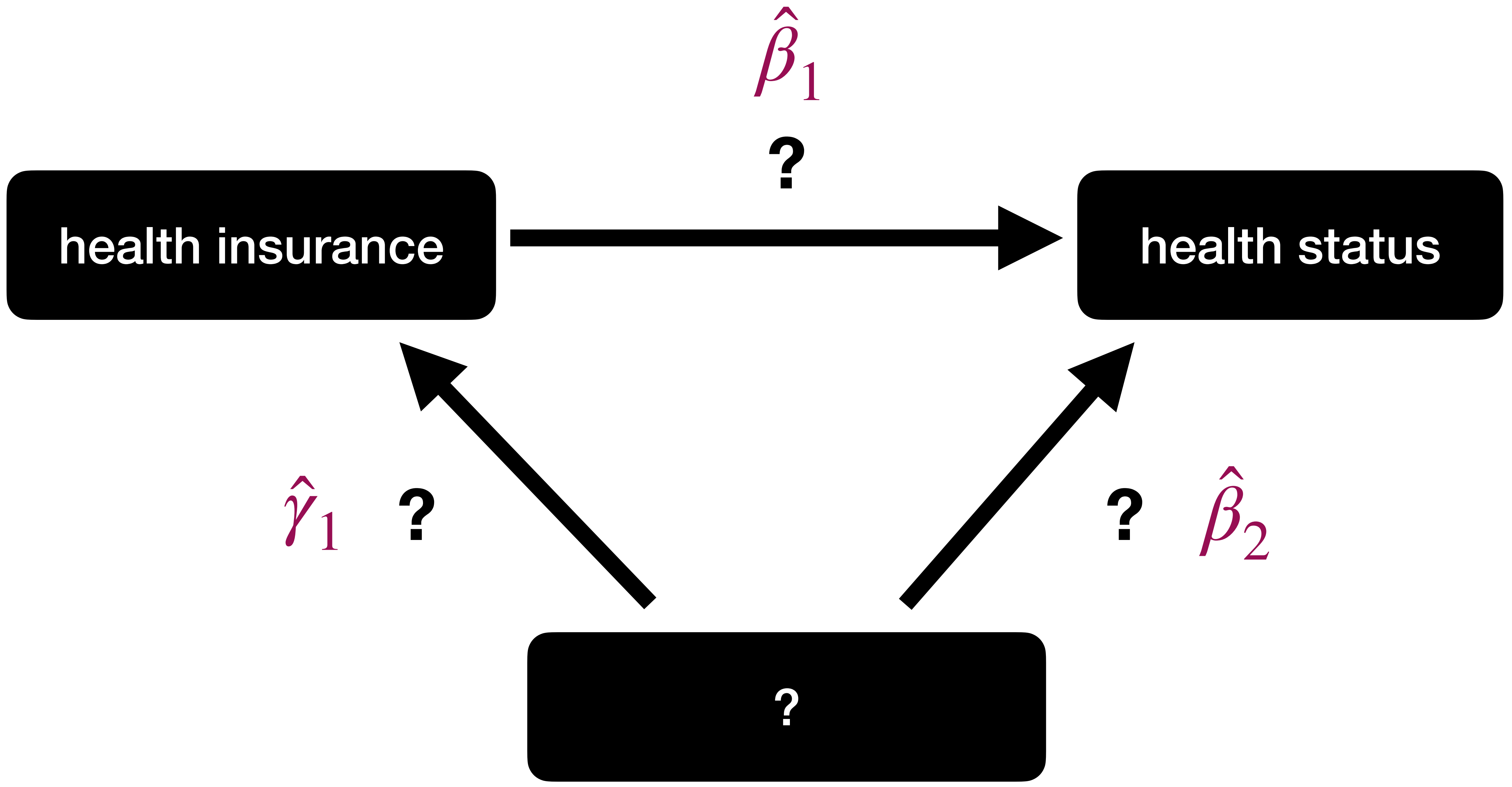
	Uninsured (n=213,714)	Insured (n=2,039,416)	Difference	t-test (P-value)
Percent under 50 years	60.9%	33.2%	+27.7 pp	P<0.001
Percent men	48.5%	43.0%	+5.5 pp	P<0.001
Percent non-Latin white	58.8%	80.3%	–21.5 pp	P<0.001
Percent married	38.5%	56.0%	–17.5 pp	P<0.001
Percent college-educated	16.4%	39.0%	–22.6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

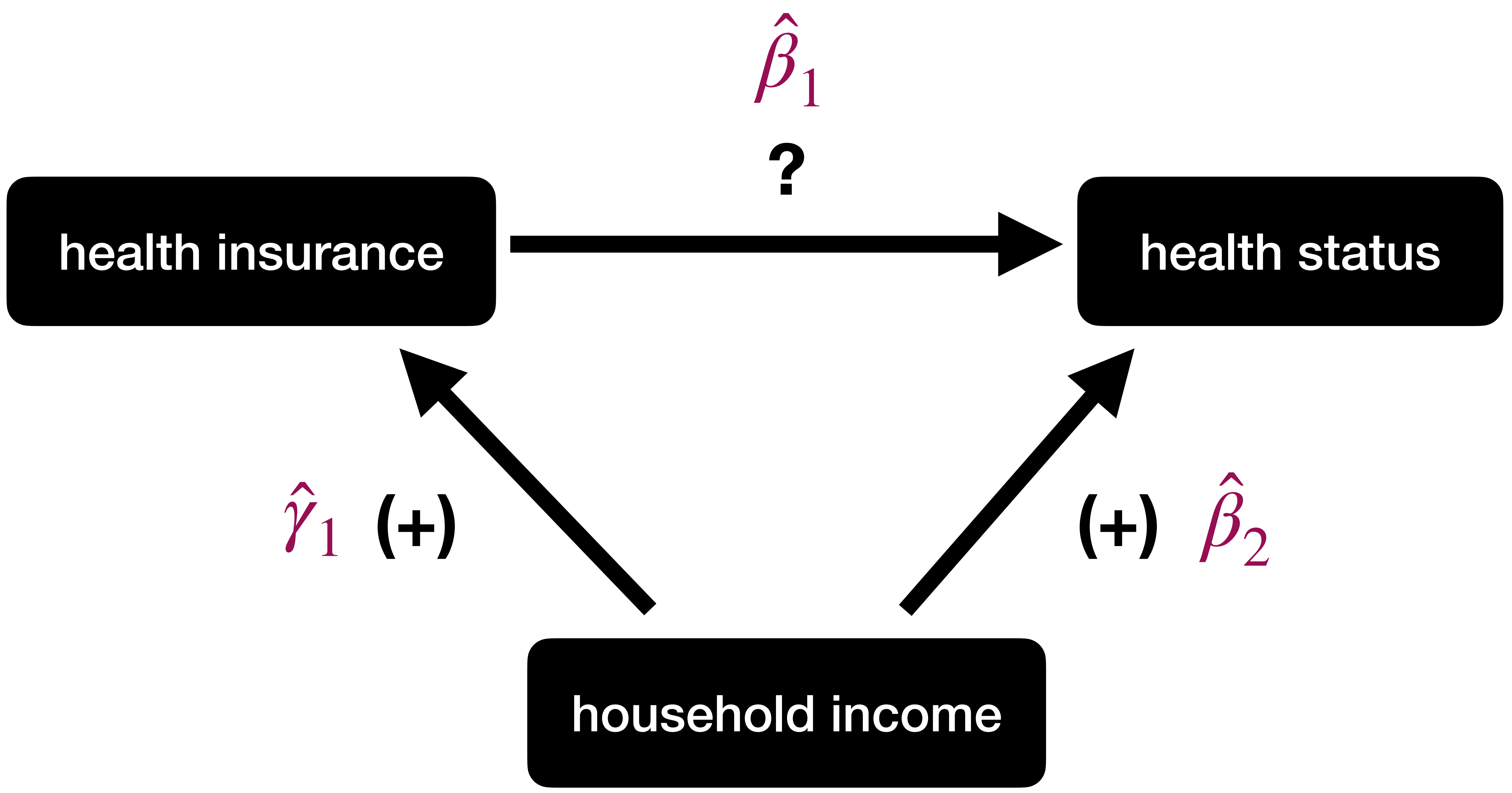


Omitted variable bias will haunt your dreams.

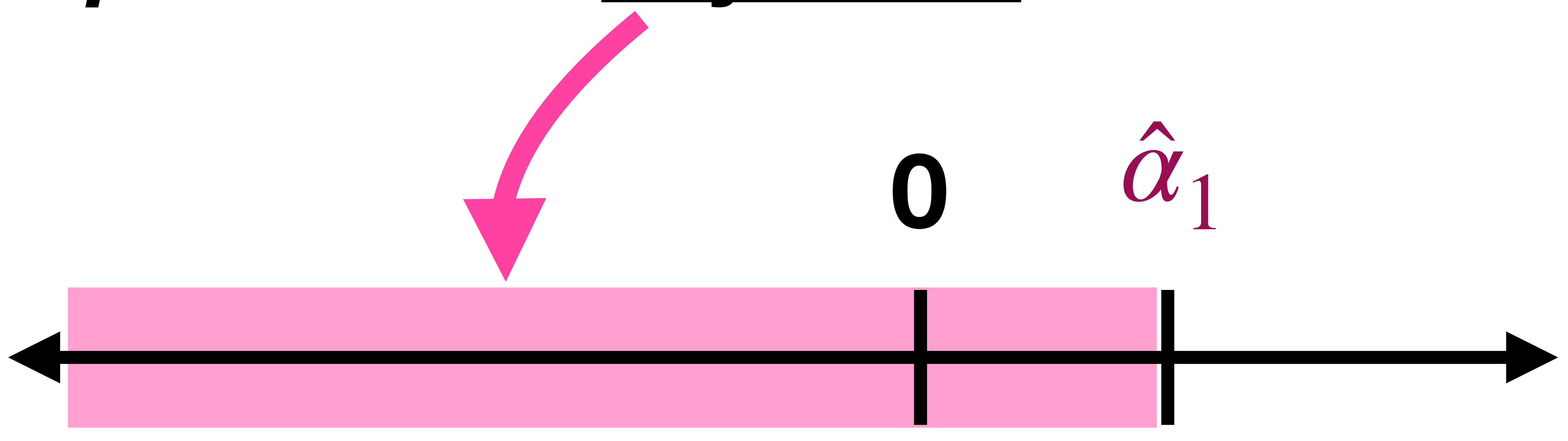
Let's say that in our short regression, having insurance = better health.

$\hat{\alpha}_1 > 0$

What might be an omitted variable?



$\hat{\beta}_1$ could be anywhere in here.



Our bias is positive, so $\hat{\alpha}_1$ must be to the right of $\hat{\beta}_1$.

Bias formula

$$\hat{\alpha}_1 - \hat{\beta}_1 = \hat{\beta}_2 * \hat{\gamma}_1 = (+)(+) = (+)$$


Nolan
If anything should happen to me, I want you to read
Mostly Harmless Econometrics
at my funeral.
Nolan's PhD adviser
I'm not doing that.

Let’s just randomize people!

Pick 10,000 uninsured folks & randomly 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%				
Percent men	48.5%				
Percent non-Latin white	58.8%				
Percent married	38.5%				
Percent college-educated	16.4%				
Percent with income <$35,000	71.3%				

What do we expect the two groups to look like?

Let’s just randomize people!

Pick 10,000 uninsured folks & randomly 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

Notice that the groups are much more “balanced” on all variables than before.

Let's just randomize people!

Pick 10,000 uninsured folks & randomly 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

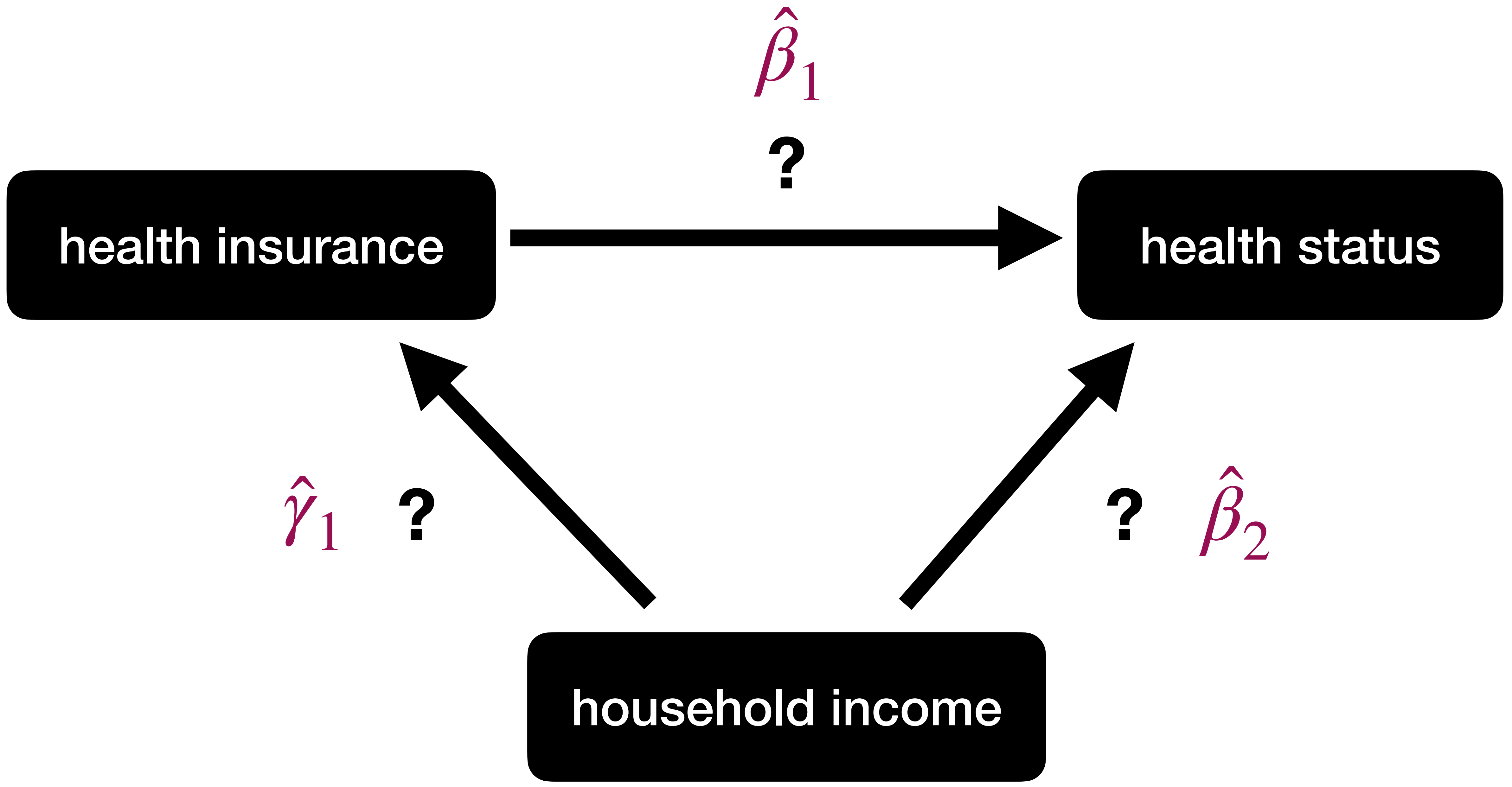
Sometimes we'll get "significant" values just due to chance. We should expect that about 5% of the time!

What happens to the omitted variable bias?

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

$\hat{\alpha}_1 > 0$

Now, sign the bias again.

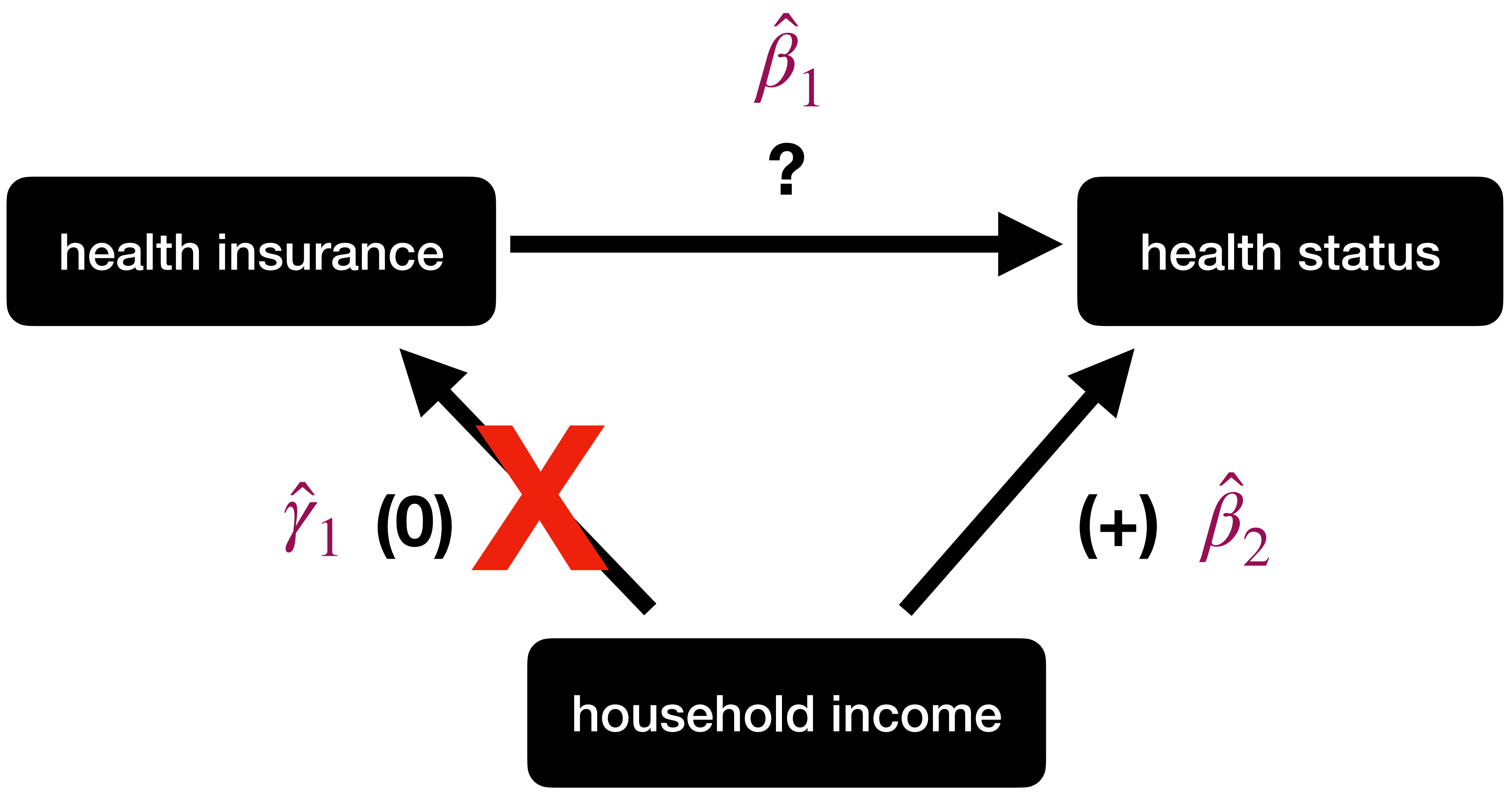


What happens to the omitted variable bias?

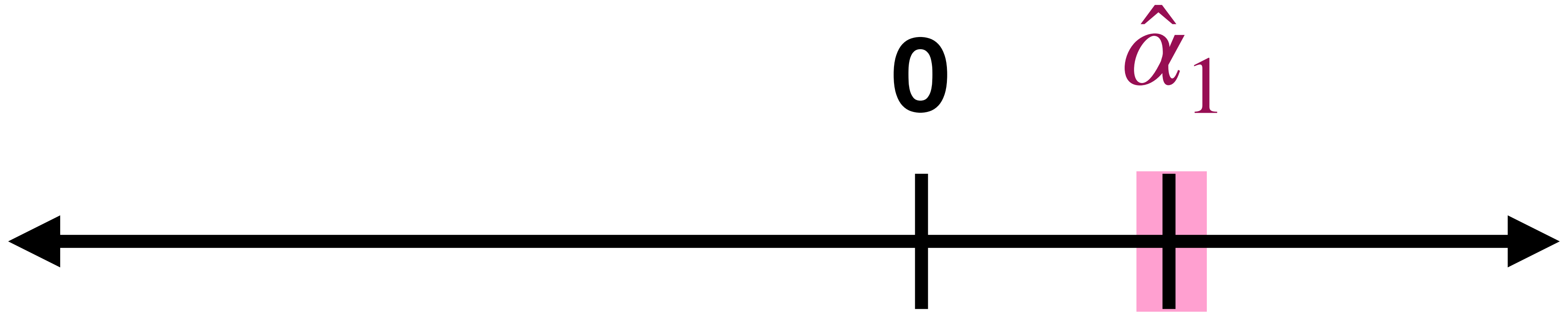
Let's say that in our short regression, having insurance = better health.

$\hat{\alpha}_1 > 0$

Now, sign the bias again.



In a well-executed experiment, $\hat{\beta}_1$ is still right here.



We have no bias, so $\hat{\alpha}_1 = \hat{\beta}_1$.

Bias formula $\widehat{\alpha}_1 - \widehat{\beta}_1 = \widehat{\beta}_2 * \widehat{\gamma}_1 = (+) * 0 = 0$


Ha ha!
Eliminating a million omitted variables

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.

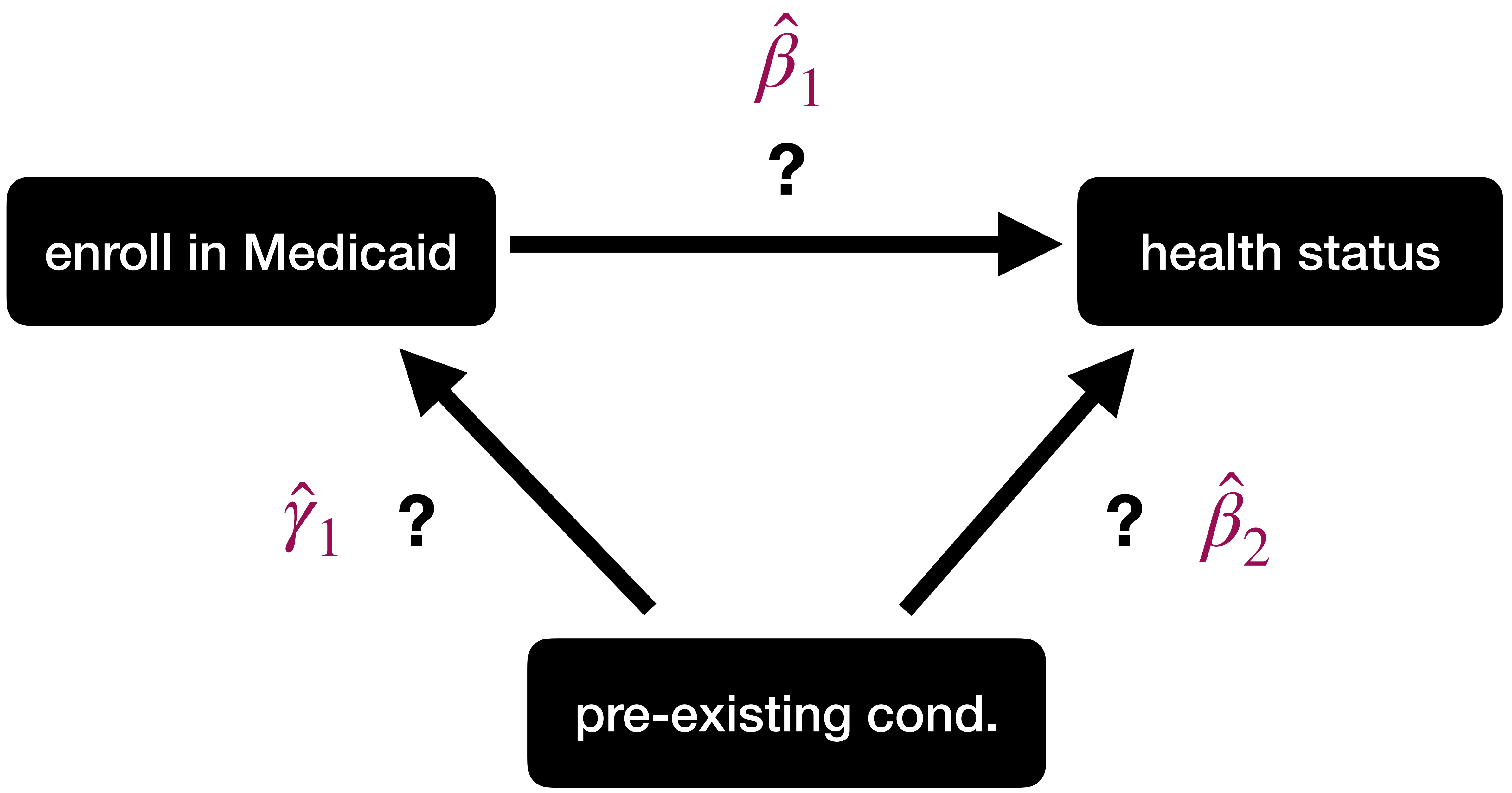
When we analyze our experiment, whom should we compare?

Let's say we just analyze enrollees.

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

$\hat{\alpha}_1 > 0$

What if only "sicker" people enrolled?

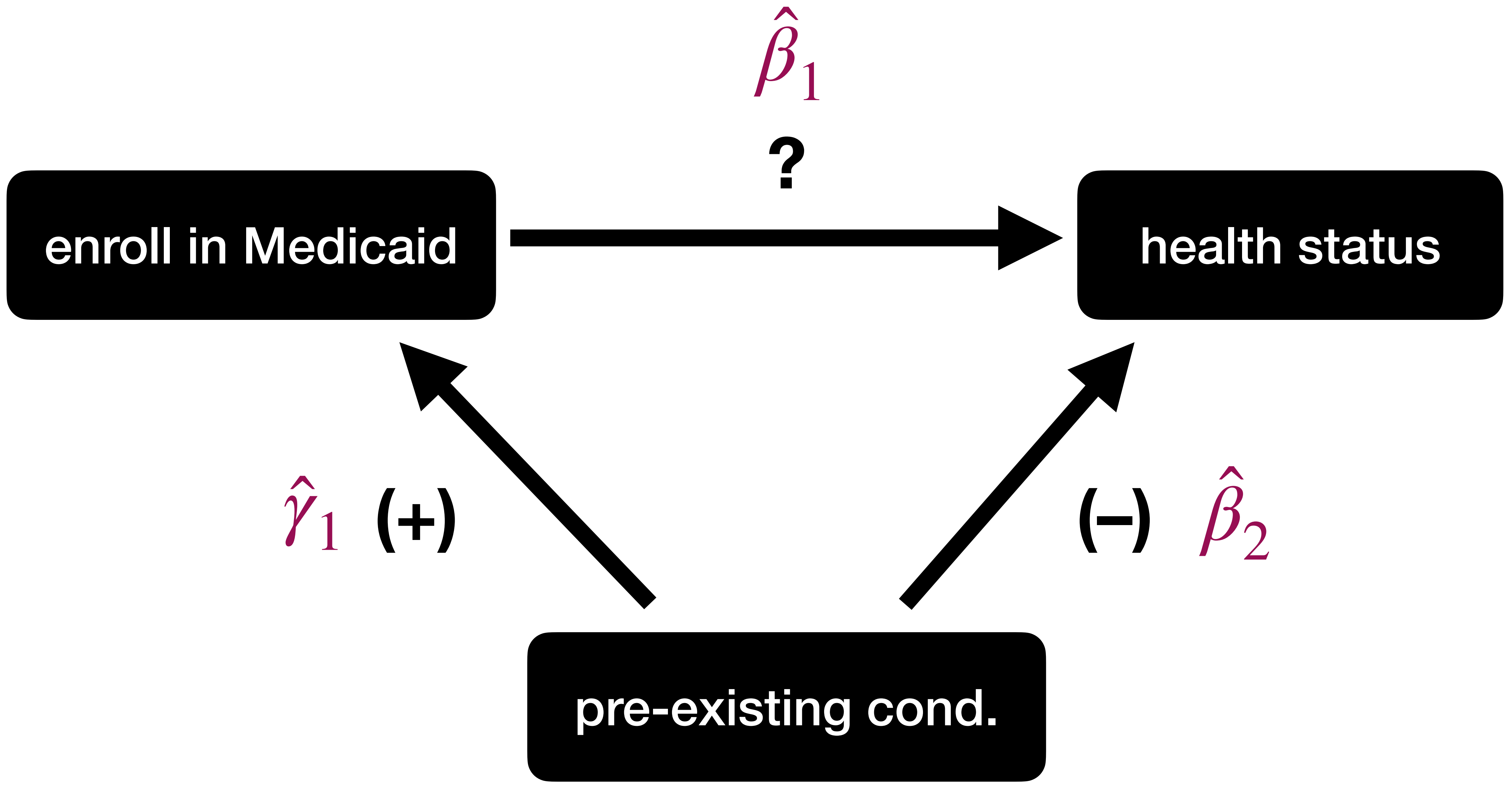


Let's say we just analyze enrollees.

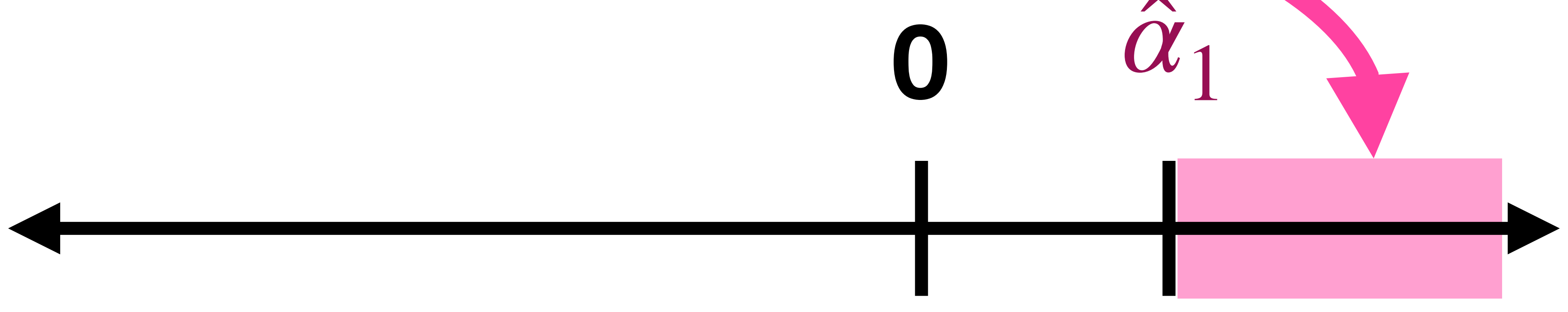
In our short regression, we find that Medicaid = better health.

$\hat{\alpha}_1 > 0$

What if only "sicker" people enrolled?



$\hat{\beta}_1$ could be <u>anywhere</u> over here.



Our bias is negative, so $\hat{\alpha}_1$ must be to the left of $\hat{\beta}_1$.

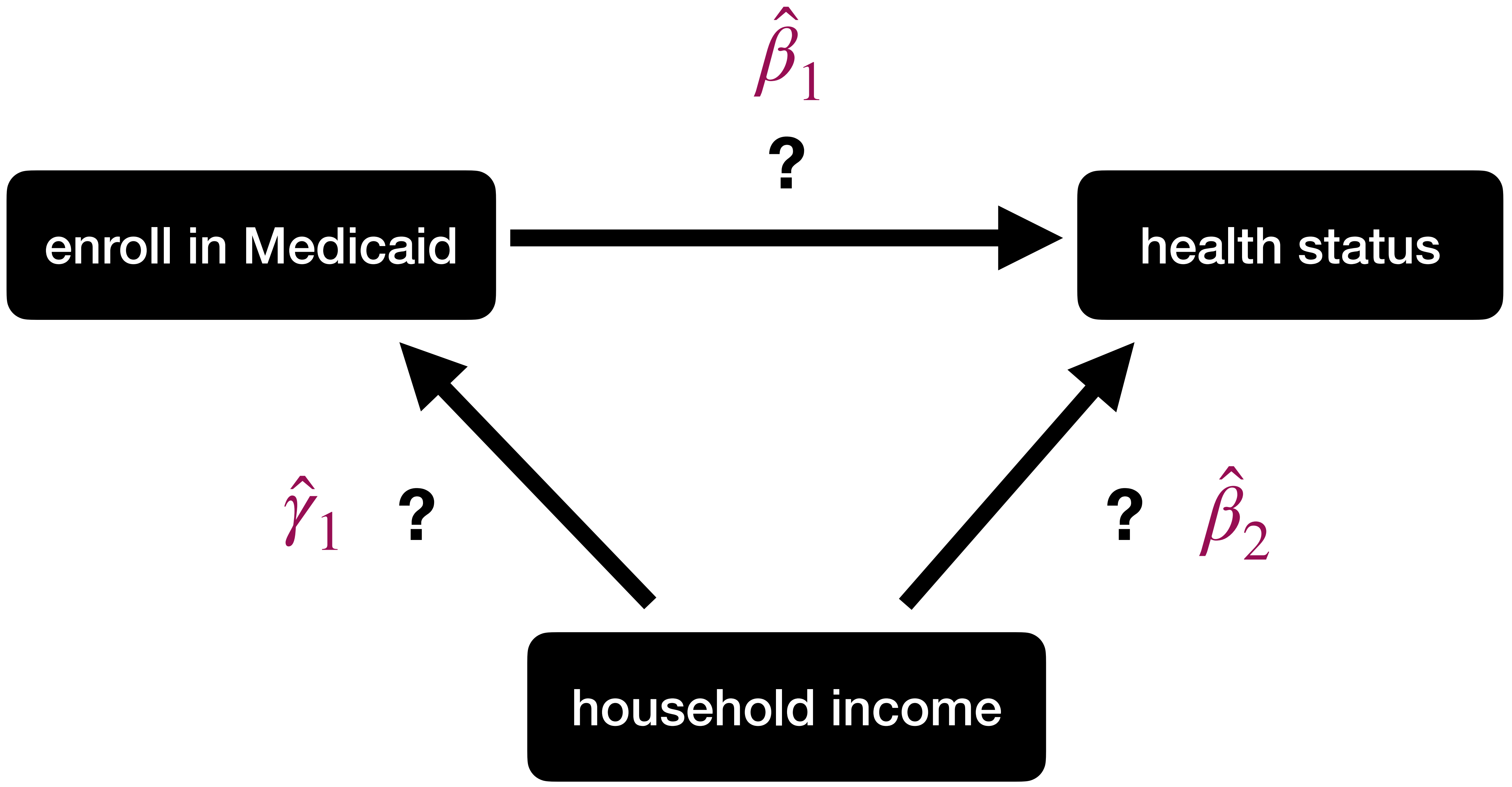
Bias formula $\hat{\alpha}_1 - \hat{\beta}_1 = \hat{\beta}_2 * \hat{\gamma}_1 = (+)(-) = (-)$

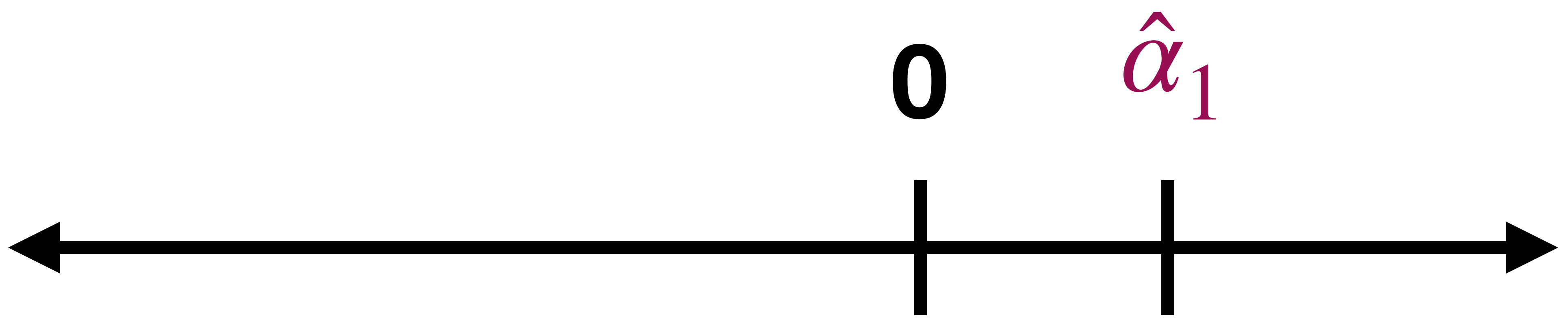
Let’s say we just analyze enrollees.

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

$\hat{\alpha}_1 > 0$

What if only “richer” people enrolled?





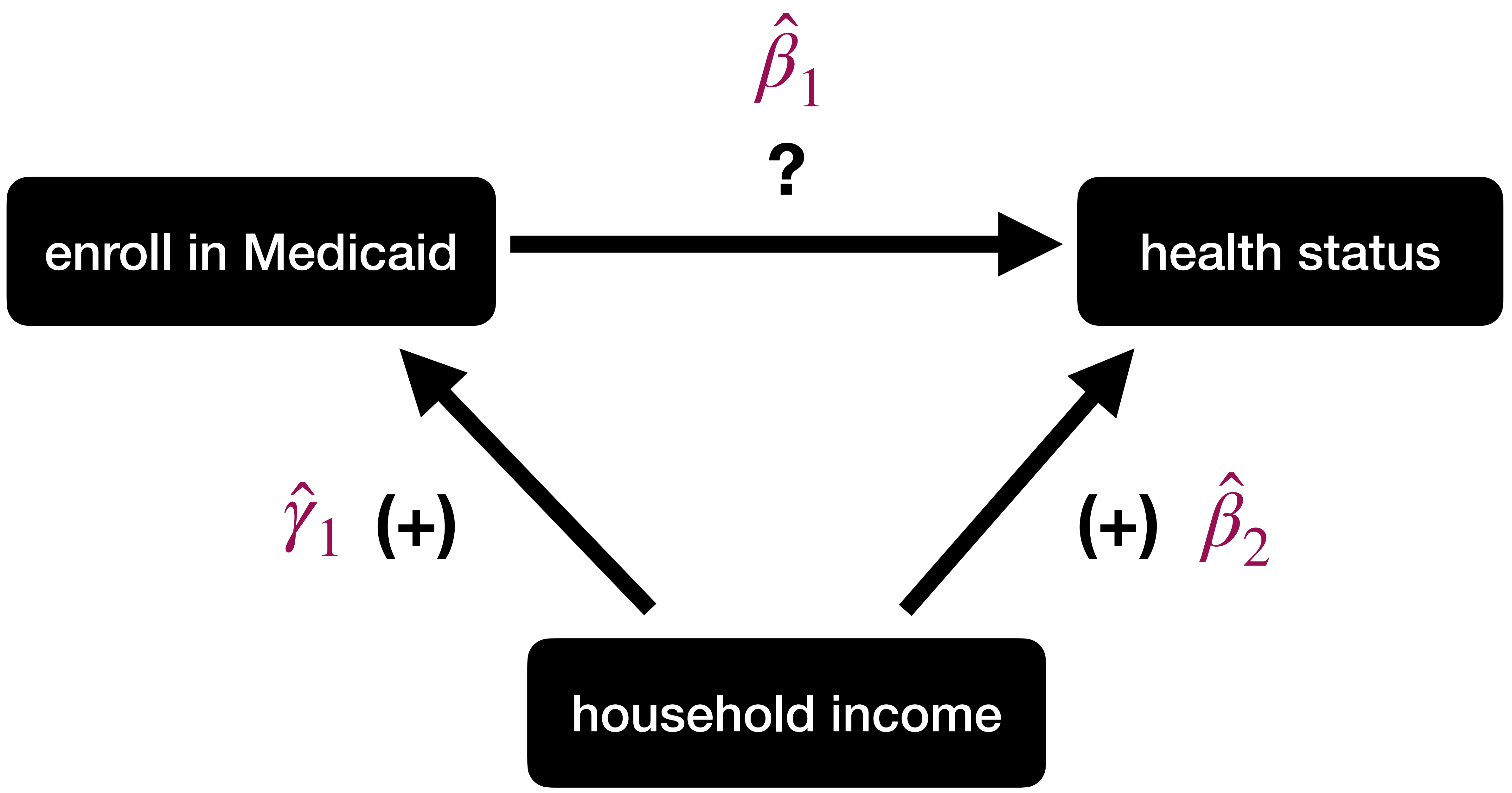
Bias formula $\hat{\alpha}_1 - \hat{\beta}_1 = \hat{\beta}_2 * \hat{\gamma}_1 =$

Let’s say we just analyze enrollees.

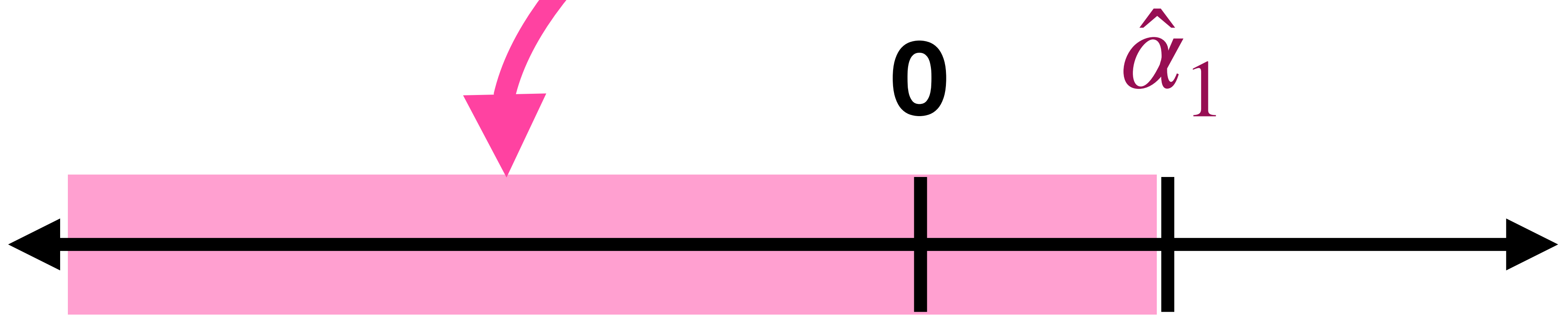
In our short regression, we find that Medicaid = better health.

$\hat{\alpha}_1 > 0$

What if only “richer” people enrolled?



$\hat{\beta}_1$ could be <u>anywhere</u> in here.



Our bias is positive, so $\hat{\alpha}_1$ must be to the right of $\hat{\beta}_1$.

Bias formula $\hat{\alpha}_1 - \hat{\beta}_1 = \hat{\beta}_2 * \hat{\gamma}_1 = (+)(+) = (+)$

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).

<u>Intention-to-treat</u> analyses preserve the original randomization.

We just have a different interpretation: the effect of winning the lottery.

Can experiments have other biases?

Yes!

We are especially concerned about:

1. **Attrition** (which is like an omitted variable — hence, intention to treat)

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

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

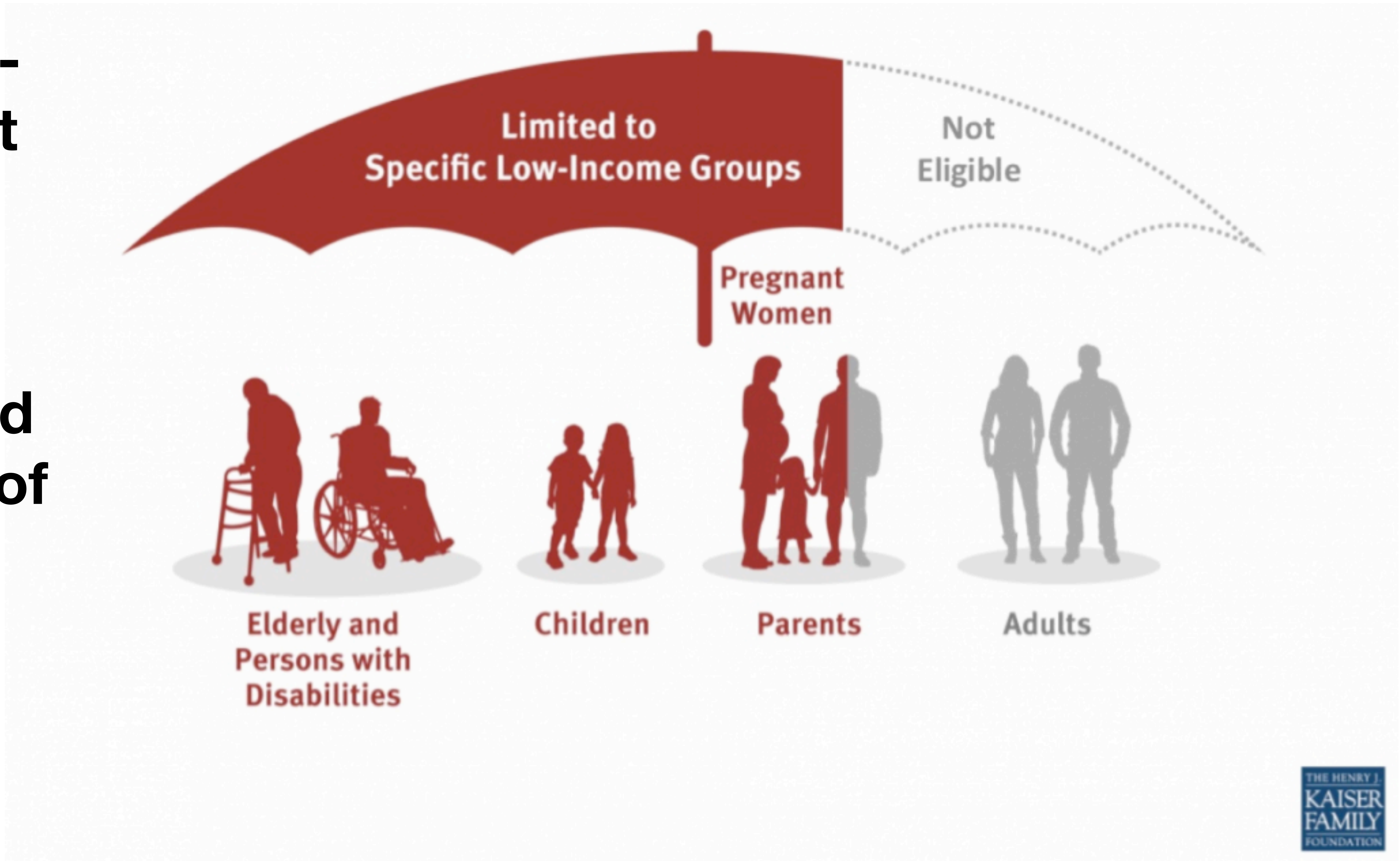
Also, experiments can be expensive, impractical, or unethical.

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 all citizens up to 138% of the federal poverty level.

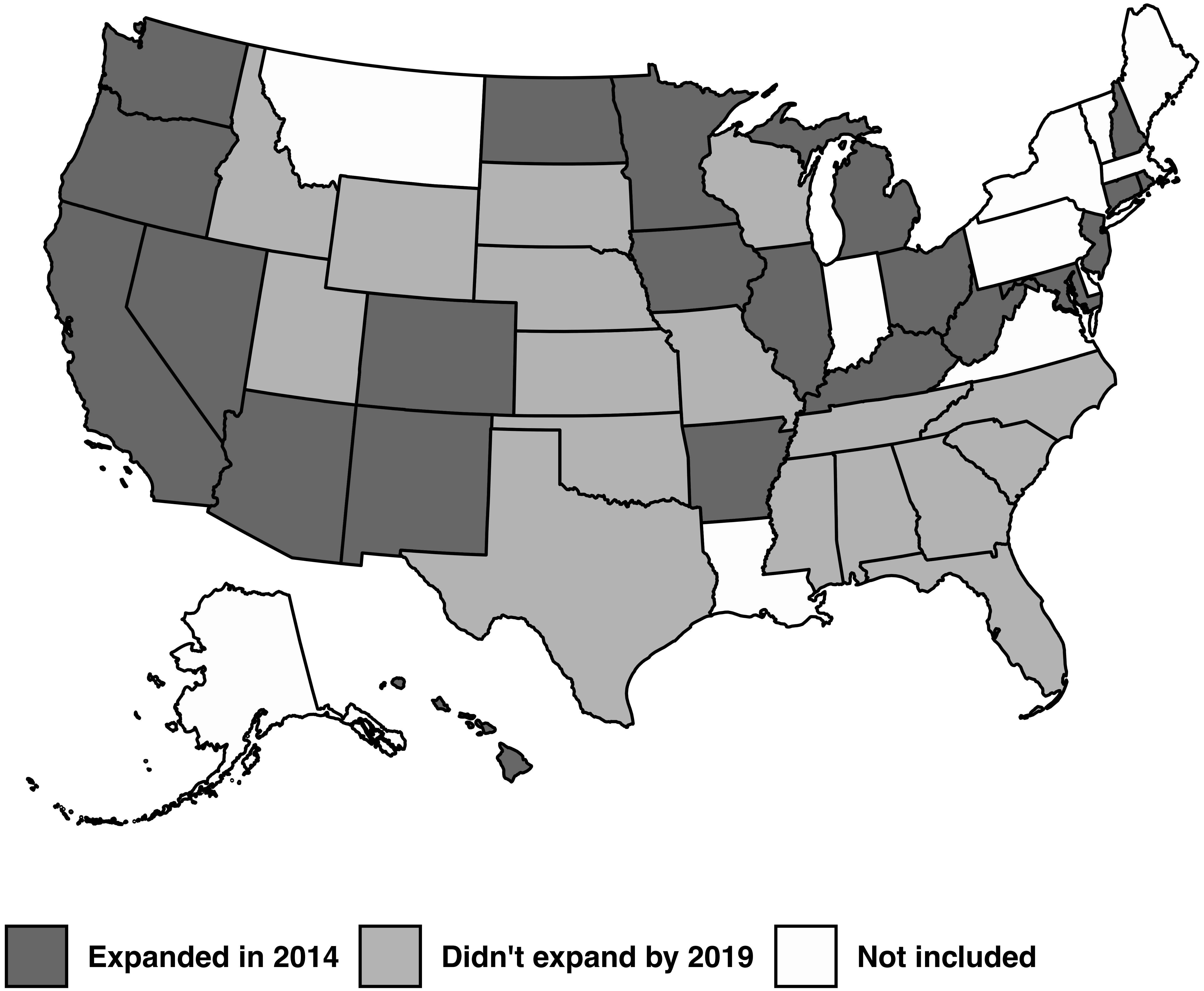
In 2026, 138% FPL for a family of 4 is $45,540.



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*!

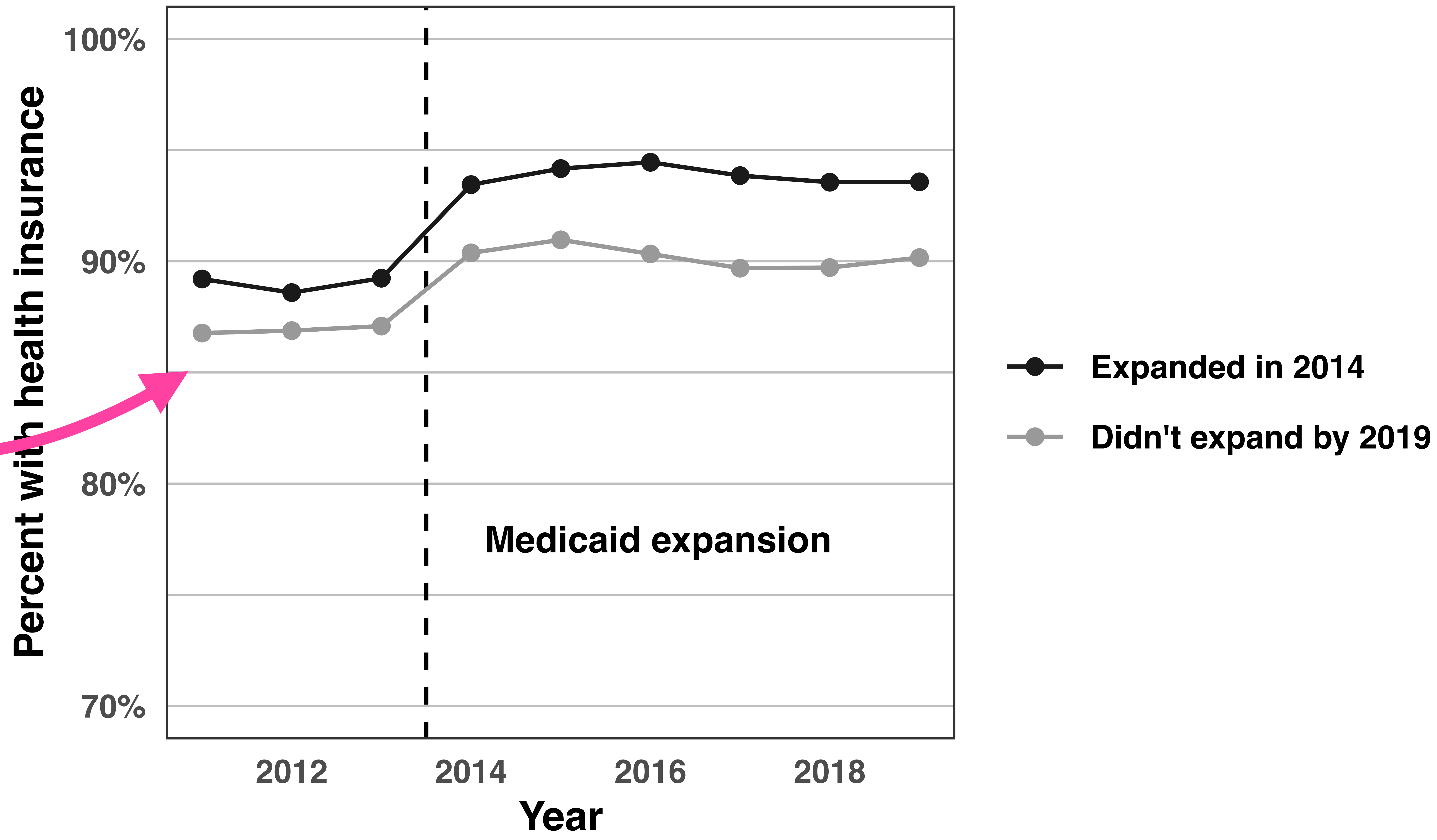


***Each state’s choice to expand isn’t really random, but if our difference-in-difference assumptions hold, we can pretend like it is!**


AMERICAN GLADIATORS MONDAY MAY 12
You're choosing a natural experiment over an RCT
Nolan, writing his thesis papers
I can have it all!

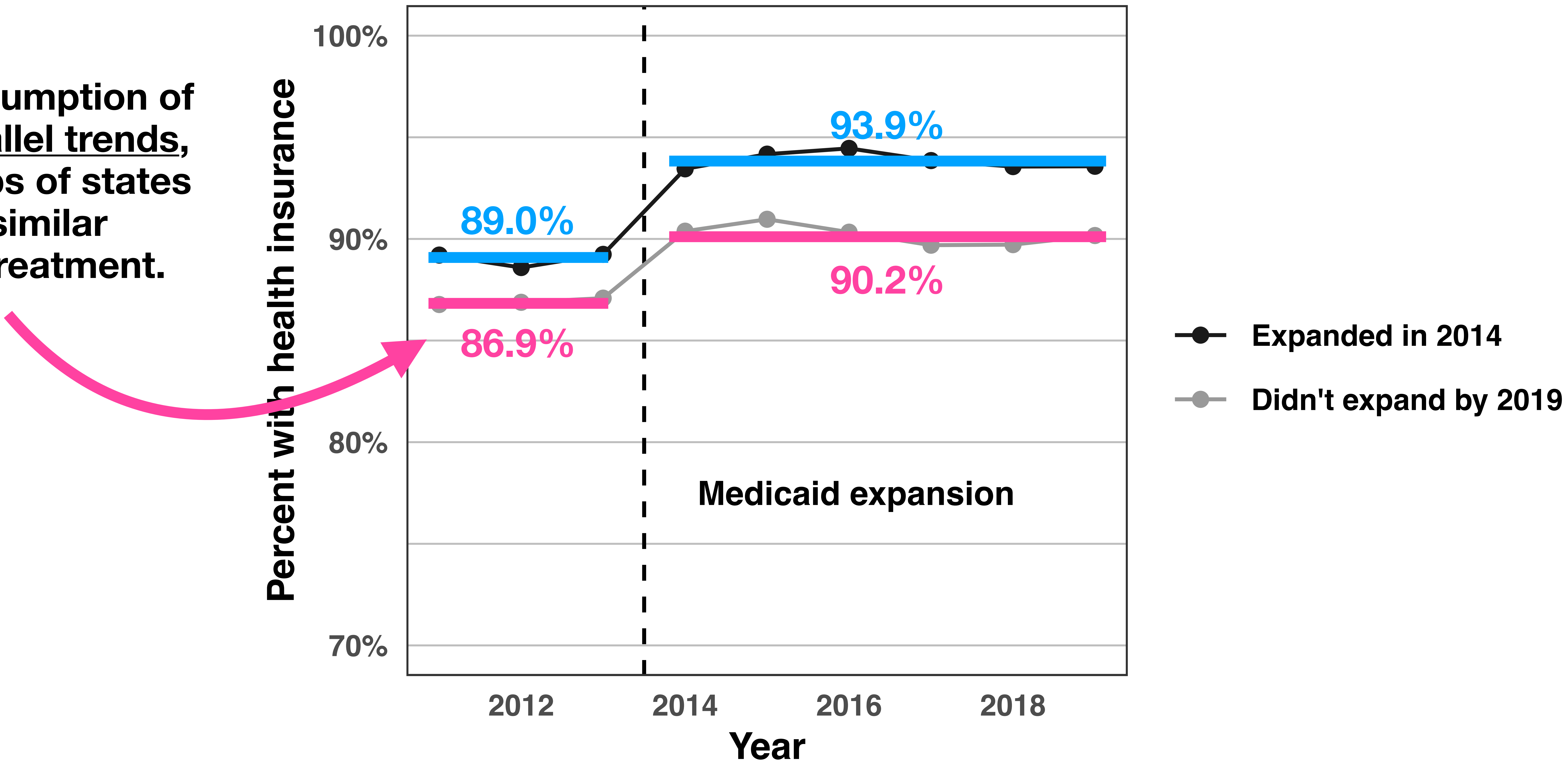
Let's do a difference-in-differences!

The identifying assumption of a diff-in-diff is <u>parallel trends</u>, i.e. that both groups of states would have had a similar trajectory absent treatment.



Let’s do a difference-in-differences!

The identifying assumption of a diff-in-diff is <u>parallel trends</u>, i.e. that both groups of states would have had a similar trajectory absent treatment.



Let's do a difference-in-differences!

$$(insurance)_i = \widehat{\beta}_0 + \widehat{\beta}_1(expansion)_i + \widehat{\beta}_2(post\_2014)_i + \widehat{\beta}_3(expansion * post\_2014)_i + \widehat{u}_i$$

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!

$$(insurance)_i = \hat{\beta}_0 + \hat{\beta}_1(expansion)_i + \hat{\beta}_2(post\_2014)_i + \hat{\beta}_3(expansion * post\_2014)_i + \hat{u}_i$$

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`	$\hat{\beta}_0$	$\hat{\beta}_0 + \hat{\beta}_2$	$\hat{\beta}_2$
Expanded in 2014 `expansion = 1`	$\hat{\beta}_0 + \hat{\beta}_1$	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$	$\hat{\beta}_2 + \hat{\beta}_3$
Difference	$\hat{\beta}_1$	$\hat{\beta}_1 + \hat{\beta}_3$	$\hat{\beta}_3$

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`	$\hat{\beta}_0$	$\hat{\beta}_0 + \hat{\beta}_2$	$\hat{\beta}_2$
Expanded in 2014 `expansion = 1`	$\hat{\beta}_0 + \hat{\beta}_1$	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$	$\hat{\beta}_2 + \hat{\beta}_3$
Difference	$\hat{\beta}_1$	$\hat{\beta}_1 + \hat{\beta}_3$	$\hat{\beta}_3$

What omitted variables are we eliminating?

What's left? Omitted variables that vary by state <u>and</u> time.
e.g. changes in state policies that differently affect the groups

omitted variables that vary between states but not over time
e.g. demographics that don't change within a state over time

omitted variables that vary over time but are the same between states
e.g. national economic trends that affect both groups of states

	Before 2014 `post_2014 = 0`	**After 6/1/2014** `post_2014 = 1`	**Difference**
Didn't expand by 2019 `expansion = 0`	$\hat{\beta}_0$	$\hat{\beta}_0 + \hat{\beta}_2$	$\hat{\beta}_2$
Expanded in 2014 `expansion = 1`	$\hat{\beta}_0 + \hat{\beta}_1$	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$	$\hat{\beta}_2 + \hat{\beta}_3$
Difference	$\hat{\beta}_1$	$\hat{\beta}_1 + \hat{\beta}_3$	$\hat{\beta}_3$


What's your religion, policymaker?
I pretty much just do whatever diff-in-diffs tell me

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

Note: `insurance` was measured as 0–1.

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

Note: **insurance** was measured as 0–1.

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

(Note: We're comparing groups within the interaction, so 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

Note: `insurance` was measured as 0–1.

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

Note: `insurance` was measured as 0–1.

The insurance rate in <u>non-expansion</u> states rose by 3.3 pp after 2014.*

Alternatively, the insurance rate in non-expansion states was 3.3 pp higher after 2014 than it was before 2014.

The difference is statistically significant.

*This is not a causal statement; it just describes what happened.

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

Note: `insurance` was measured as 0–1.

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

That is, the difference in insurance rates over time for expansion states, compared to the difference over time 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	$\boxed{\textbf{0.016**}}$
	(0.005)
Num.Obs.	2,253,130
R2	0.008
R2 Adj.	0.008

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

Note: `insurance` was measured as 0–1.

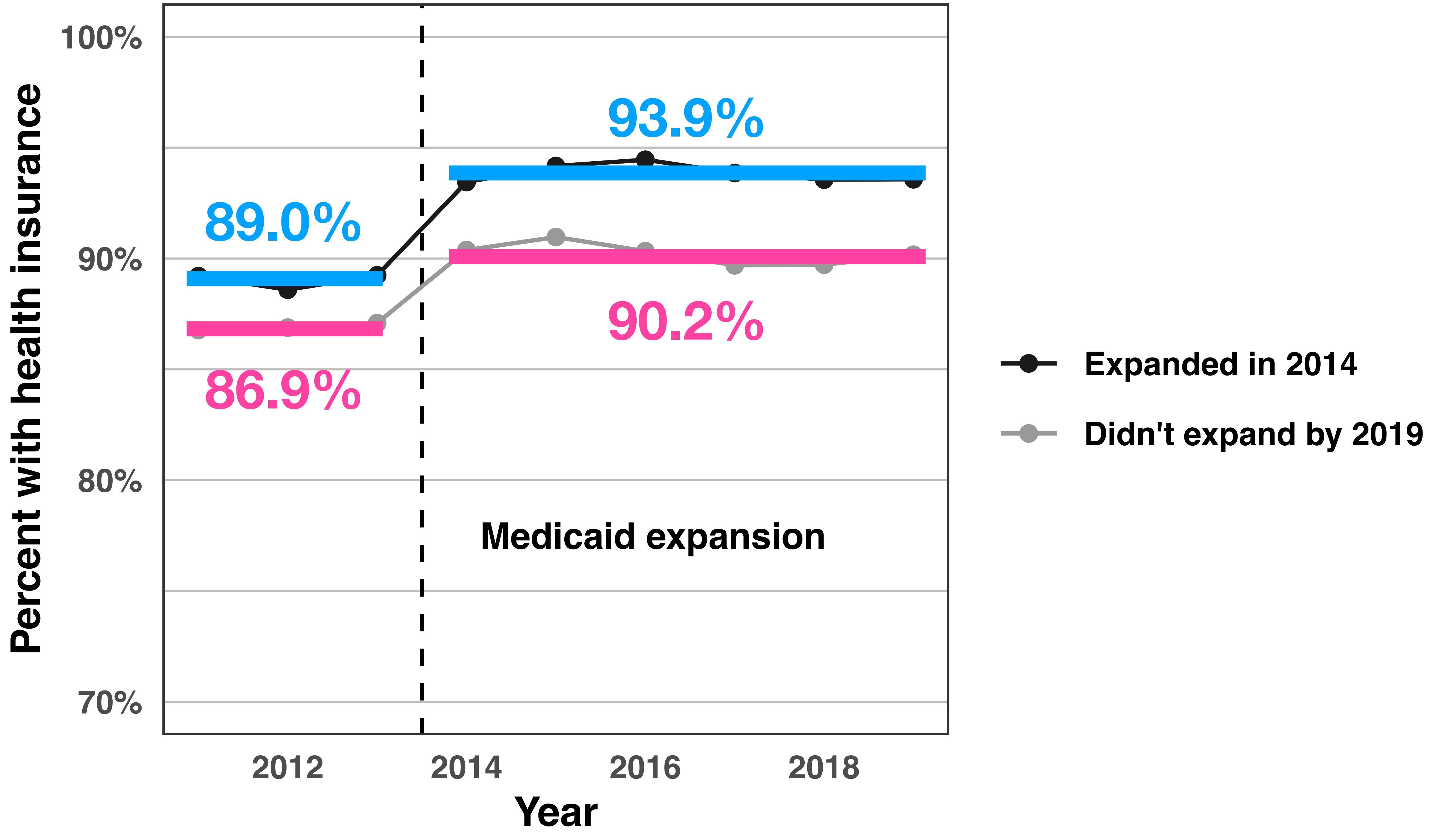
To a policymaker, we could say, "The causal effect of Medicaid expansion was to increase health insurance coverage by 1.6 percentage points in expansion states, relative to non-expansion states."

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



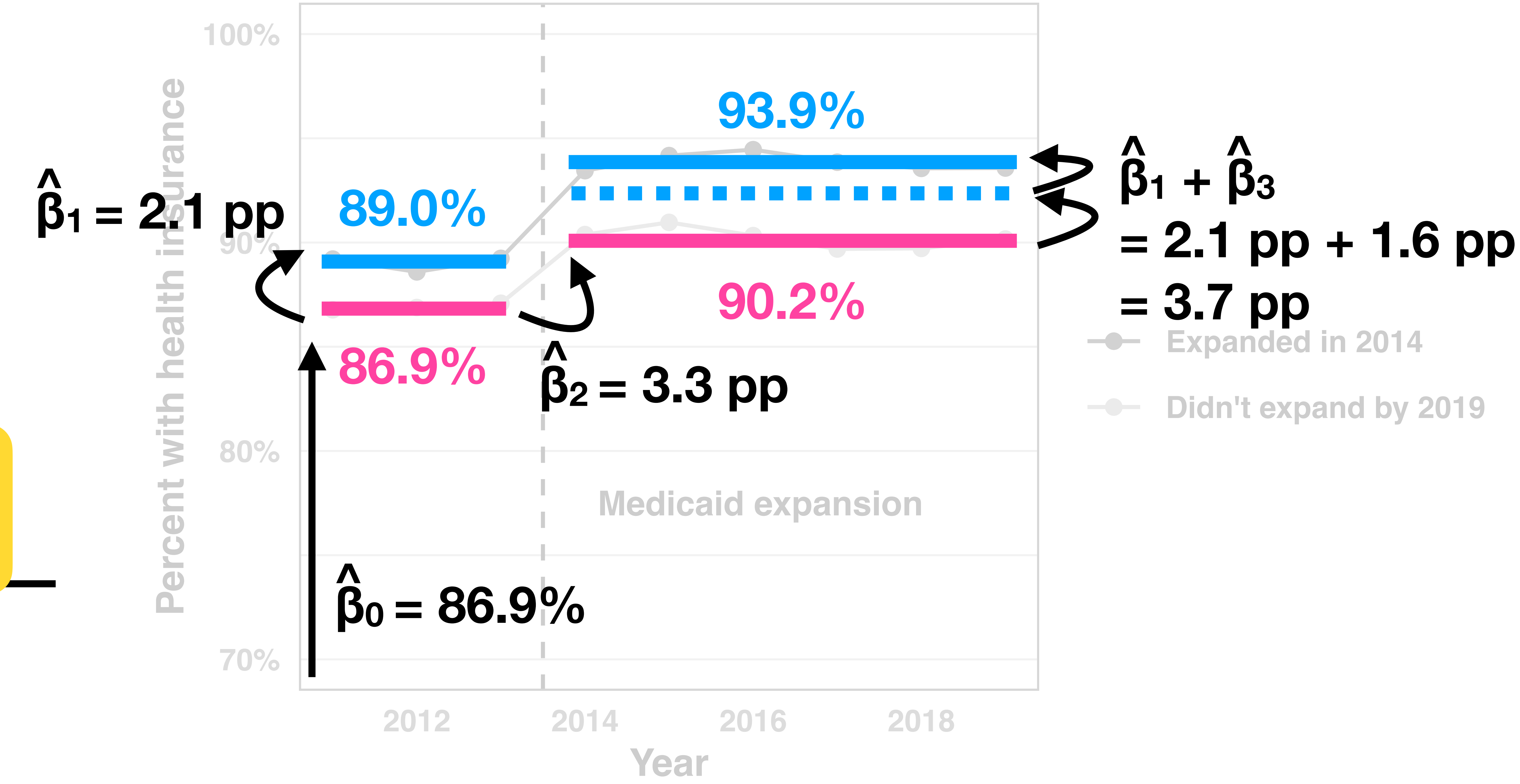
Note: `insurance` was measured as 0–1.

$$(insurance)_i = \widehat{\beta}_0 + \widehat{\beta}_1(expansion)_i + \widehat{\beta}_2(post\_2014)_i + \widehat{\beta}_3(expansion * post\_2014)_i + \widehat{u}_i$$

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



$$(insurance)_i = \hat{\beta}_0 + \hat{\beta}_1(expansion)_i + \hat{\beta}_2(post\_2014)_i + \hat{\beta}_3(expansion * post\_2014)_i + \hat{u}_i$$

Note: `insurance` was measured as 0–1.

But is it causal?

Again, causality is a spectrum.

We have to evaluate the assumptions.

But here, I would say yes, causal!

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