Evaluating policies – the use of quasi-experiments

Sam Harper

Epidemiology, Biostatistics & Occupational Health, McGill University

Institute for Health and Social Policy, McGill University

Endowed Chair on Health Inequalities, Erasmus University

HS02c: Public health research: intervention development and evaluation, 8 Nov, 2020, Erasmus Medical Center

Outline

1. Problems with non-randomized studies
2. Natural and quasi-experiments
3. Example study designs
   - Difference-in-Differences
   - Instrumental variables
   - Regression discontinuity
4. Concluding thoughts
5. References

Our objective (mostly)

- We are mainly (though not exclusively) interested in causal effects.
- We want to know:
  - Did the program work? If so, for whom? If not, why not?
  - If we implement the program elsewhere, should we expect the same result?
- These questions involve counterfactuals about what would happen **if** we intervened to do something.
- These are causal questions.

Our objective (mostly)

- We are mainly (though not exclusively) interested in causal effects.
- We want to know:
  - Did the program work? If so, for whom? If not, why not?
  - If we implement the program elsewhere, should we expect the same result?
- These questions involve counterfactuals about what would happen **if** we intervened to do something.
- These are causal questions.

Our objective (mostly)

- We are mainly (though not exclusively) interested in causal effects.
- We want to know:
  - Did the program work? If so, for whom? If not, why not?
  - If we implement the program elsewhere, should we expect the same result?
- These questions involve counterfactuals about what would happen **if** we intervened to do something.
- These are causal questions.

Causation, Association, and Confounding

- **Causal effect**: Do individuals randomly assigned (i.e., SET) to the intervention have better outcomes?

$$E\left(Y|SET\left[Treated\right]\right) - E\left(Y|SET\left[Untreated\right]\right)$$

- **Association**: Do individuals who choose to take the intervention have better outcomes?

$$E\left(Y|Treated\right) - E(Y|Untreated)$$

- **Confounding** :

$$E\left(Y|SET\left[Treated\right]\right) - E\left(Y|SET\left[Untreated\right]\right) \neq E\left(Y|Treated\right) - E\left(Y|Untreated\right)$$

Causation, Association, and Confounding

- **Causal effect**: Do individuals randomly assigned (i.e., SET) to the intervention have better outcomes?

$$E\left(Y|SET\left[Treated\right]\right)-E\left(Y|SET\left[Untreated\right]\right)$$

- **Association**: Do individuals who choose to take the intervention have better outcomes?

$$E\left(Y|Treated\right)-E(Y|Untreated)$$

- **Confounding** :

$$E\left(Y|SET\left[Treated\right]\right)-E\left(Y|SET\left[Untreated\right]\right)\neq E\left(Y|Treated\right)-E\left(Y|Untreated\right)$$

Causation, Association, and Confounding

- **Causal effect**: Do individuals randomly assigned (i.e., SET) to the intervention have better outcomes?

$$E\left(Y|SET\left[Treated\right]\right)-E\left(Y|SET\left[Untreated\right]\right)$$

- **Association**: Do individuals who choose to take the intervention have better outcomes?

$$E\left(Y|\,Treated\right)-E(Y|\,Untreated)$$

- **Confounding** :

$$E\left(Y|SET\left[Treated\right]\right)-E\left(Y|SET\left[Untreated\right]\right)\neq E\left(Y|\,Treated\right)-E\left(Y|Untreated\right)$$

Randomized Trials vs. Observational Studies

RCTs, Defined

RCTs involve: (1) comparing treated and control groups; (2) the treatment assignment is random; and (3) investigator does the randomizing.

- In an RCT, treatment/exposure is assigned by the investigator
- In observational studies, exposed/unexposed groups exist in the source population and are selected by the investigator.

- Good natural experiments do (1) and (2), but not (3).
- Because there is no control over assignment, the credibility of natural experiments hinges on how good "as-if random" approximates (2).

Randomized Trials vs. Observational Studies

RCTs, Defined

RCTs involve: (1) comparing treated and control groups; (2) the treatment assignment is random; and (3) investigator does the randomizing.

- In an RCT, treatment/exposure is assigned by the investigator
- In observational studies, exposed/unexposed groups exist in the source population and are selected by the investigator.
- Good natural experiments do (1) and (2), but not (3).
- Because there is no control over assignment, the credibility of natural experiments hinges on how good "as-if random" approximates (2).

Randomized Trials vs. Observational Studies

RCTs, Defined

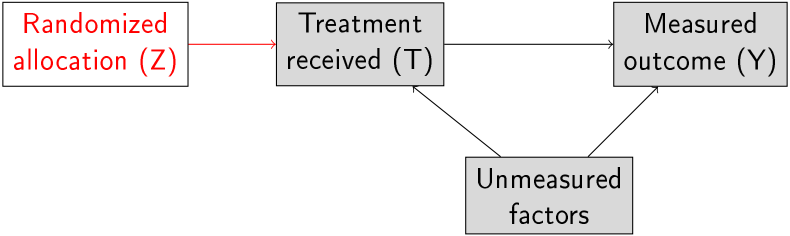
RCTs involve: (1) comparing treated and control groups; (2) the treatment assignment is random; and (3) investigator does the randomizing.

- In an RCT, treatment/exposure is assigned by the investigator
- In observational studies, exposed/unexposed groups exist in the source population and are selected by the investigator.

- Good natural experiments do (1) and (2), but not (3).
- Because there is no control over assignment, the credibility of natural experiments hinges on how good "as-if random" approximates (2).

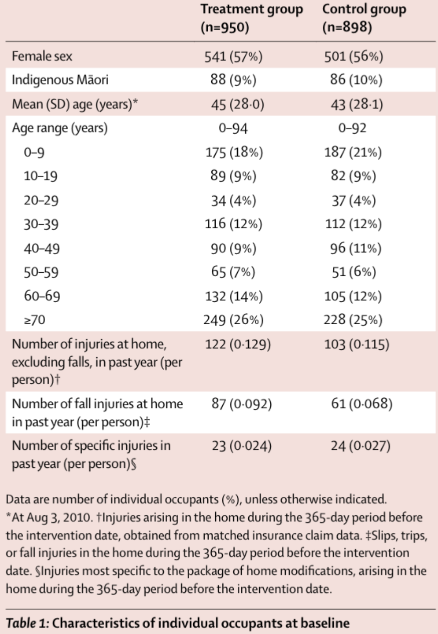
Strength of randomized treatment allocation

- Recall that randomization means that we can generally estimate the causal effect without bias.
- Randomization guarantees exchangeability on measured and unmeasured factors.



Randomize if you can.

- Randomization leads to:
  - balance on measured factors.
  - balance on unmeasured factors.
- Unmeasured factors cannot bias the estimate of the exposure effect.
- Example from Home Injury Prevention Intervention cluster RCT (Keall et al. 2015[1])



	Treatment group (n=950)	Control group (n=898)
Female sex	541 (57%)	501 (56%)
Indigenous Māori	88 (9%)	86 (10%)
Mean (SD) age (years)*	45 (28·0)	43 (28·1)
Age range (years)	0–94	0–92
0–9	175 (18%)	187 (21%)
10–19	89 (9%)	82 (9%)
20–29	34 (4%)	37 (4%)
30–39	116 (12%)	112 (12%)
40–49	90 (9%)	96 (11%)
50–59	65 (7%)	51 (6%)
60–69	132 (14%)	105 (12%)
≥70	249 (26%)	228 (25%)
Number of injuries at home, excluding falls, in past year (per person)†	122 (0·129)	103 (0·115)
Number of fall injuries at home in past year (per person)‡	87 (0·092)	61 (0·068)
Number of specific injuries in past year (per person)§	23 (0·024)	24 (0·027)

Data are number of individual occupants (%), unless otherwise indicated. *At Aug 3, 2010. †Injuries arising in the home during the 365-day period before the intervention date, obtained from matched insurance claim data. ‡Slips, trips, or fall injuries in the home during the 365-day period before the intervention date. §Injuries most specific to the package of home modifications, arising in the home during the 365-day period before the intervention date.

***Table 1:* Characteristics of individual occupants at baseline**

Consequences of non-randomized treatment assignment

- If we are not controlling treatment assignment, then who is?
- Policy programs do not typically select people to treat at random.
  - Programs may target those that they think are most likely to benefit.
  - Programs implemented decisively non-randomly (e.g., states passing drunk driving laws in response to high-profile accidents).
  - Governments deciding to tax (or negatively tax) certain goods.
- People do not choose to participate in programs at random.
  - Welfare programs, health screening programs, etc.
  - People who believe they are likely to benefit from the program.

Consequences of non-randomized treatment assignment

- If we are not controlling treatment assignment, then who is?
- Policy programs do not typically select people to treat at random.
  - Programs may target those that they think are most likely to benefit.
  - Programs implemented decisively non-randomly (e.g., states passing drunk driving laws in response to high-profile accidents).
  - Governments deciding to tax (or negatively tax) certain goods.
- People do not choose to participate in programs at random.
  - Welfare programs, health screening programs, etc.
  - People who believe they are likely to benefit from the program.

Consequences of non-randomized treatment assignment

- If we are not controlling treatment assignment, then who is?
- Policy programs do not typically select people to treat at random.
  - Programs may target those that they think are most likely to benefit.
  - Programs implemented decisively non-randomly (e.g., states passing drunk driving laws in response to high-profile accidents).
  - Governments deciding to tax (or negatively tax) certain goods.
- People do not choose to participate in programs at random.
  - Welfare programs, health screening programs, etc.
  - People who believe they are likely to benefit from the program.

What's the problem?

- We are mainly (though not exclusively) interested in causal effects.
- Randomization is generally great for answering whether treatment assignment Z affects Y.
  - treatment assignment (Z) is independent of potential outcomes and all measured and unmeasured pre-treatment variables.
  - Effect of Z on Y is unconfounded ($Z \rightarrow Y$)
- But RCTs have serious limitations.

What's the problem?

- We are mainly (though not exclusively) interested in causal effects.
- Randomization is generally great for answering whether treatment assignment Z affects Y.
  - treatment assignment (Z) is independent of potential outcomes and all measured and unmeasured pre-treatment variables.
  - Effect of Z on Y is unconfounded ($Z \rightarrow Y$)
- But RCTs have serious limitations.

What's the problem?

- We are mainly (though not exclusively) interested in causal effects.
- Randomization is generally great for answering whether treatment assignment Z affects Y.
    - treatment assignment (Z) is independent of potential outcomes and all measured and unmeasured pre-treatment variables.
    - Effect of Z on Y is unconfounded ($Z \to Y$)
- But RCTs have serious limitations.

Problem of Social Exposures

- Many social exposures/programs cannot be randomized by investigators:
  - Unethical (poverty, parental social class, job loss)
  - Impossible (ethnic background, place of birth)
  - Expensive (neighborhood environments)
- RCT results may not generalize to other population groups.
- Effects may be produced by complex, intermediate pathways.
- Some exposures are hypothesized to have long latency periods (many years before outcomes are observable).
- We need alternatives to RCTs.

Problem of Social Exposures

- Many social exposures/programs cannot be randomized by investigators:
  - Unethical (poverty, parental social class, job loss)
  - Impossible (ethnic background, place of birth)
  - Expensive (neighborhood environments)
- RCT results may not generalize to other population groups.
- Effects may be produced by complex, intermediate pathways.
- Some exposures are hypothesized to have long latency periods (many years before outcomes are observable).
- We need alternatives to RCTs.

Problem of Social Exposures

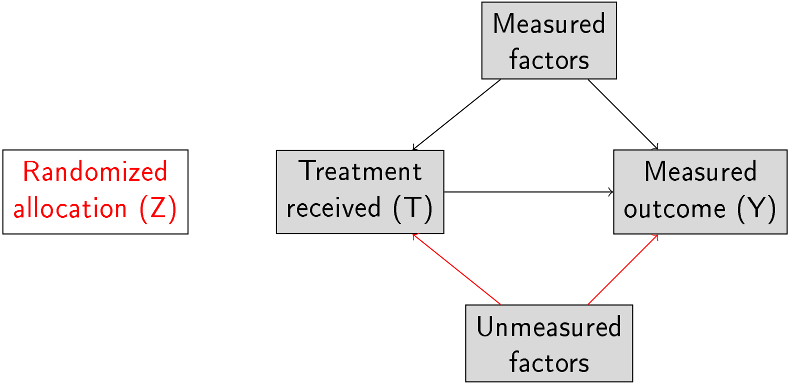
- Many social exposures/programs cannot be randomized by investigators:
    - Unethical (poverty, parental social class, job loss)
    - Impossible (ethnic background, place of birth)
    - Expensive (neighborhood environments)
- RCT results may not generalize to other population groups.
- Effects may be produced by complex, intermediate pathways.
- Some exposures are hypothesized to have long latency periods (many years before outcomes are observable).
- We need alternatives to RCTs.

Problem of Social Exposures

- Many social exposures/programs cannot be randomized by investigators:
  - Unethical (poverty, parental social class, job loss)
  - Impossible (ethnic background, place of birth)
  - Expensive (neighborhood environments)
- RCT results may not generalize to other population groups.
- Effects may be produced by complex, intermediate pathways.
- Some exposures are hypothesized to have long latency periods (many years before outcomes are observable).
- We need alternatives to RCTs.

Illustration of the problem

- Non-randomized designs typically start with observing treated and untreated groups, so more assumptions are necessary.
- In particular we should be worried about unmeasured (or mismeasured) factors that may lead to bias:



Unmeasured confounding is a serious challenge

- We often compare outcomes among socially advantaged and disadvantaged groups.
- Key problem: people choose/end up in treated or untreated group for reasons that are difficult to measure and that may be correlated with their outcomes.
- So what do we do? Typically...**adjust.**
  - Measure and adjust (regression) for C confounding factors.
  - Conditional on C, we are supposed to believe assignment is "as good as random" = causal.

Unmeasured confounding is a serious challenge

- We often compare outcomes among socially advantaged and disadvantaged groups.
- Key problem: people choose/end up in treated or untreated group for reasons that are difficult to measure and that may be correlated with their outcomes.
- So what do we do? Typically...**adjust.**
  - Measure and adjust (regression) for C confounding factors.
  - Conditional on C, we are supposed to believe assignment is "as good as random" = causal.

Unmeasured confounding is a serious challenge

- We often compare outcomes among socially advantaged and disadvantaged groups.
- Key problem: people choose/end up in treated or untreated group for reasons that are difficult to measure and that may be correlated with their outcomes.
- So what do we do? Typically...**adjust.**
  - Measure and adjust (regression) for C confounding factors.
  - Conditional on C, we are supposed to believe assignment is "as good as random" = causal.

Key issue is credibility

- If we have a good design and assume that we have measured all of the confounders, then regression adjustment can give us exactly what we want: an estimate of the causal effect of exposure to T.
- Core issue: How credible is this assumption?

"Now, keep in mind that these numbers are only as accurate as the fictitious data, ludicrous assumptions and wishful thinking they're based upon!"

Ex: SEP and CVD in Netherlands

Many observed differences. Is "no other unmeasured differences" credible?

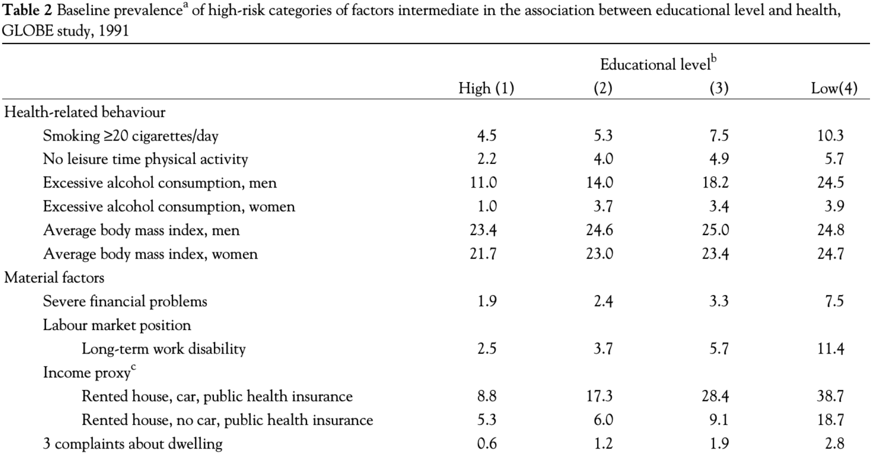


Table 2 Baseline prevalence[a] of high-risk categories of factors intermediate in the association between educational level and health, GLOBE study, 1991

	Educational level[b]			
	High (1)	(2)	(3)	Low(4)
Health-related behaviour				
Smoking ≥20 cigarettes/day	4.5	5.3	7.5	10.3
No leisure time physical activity	2.2	4.0	4.9	5.7
Excessive alcohol consumption, men	11.0	14.0	18.2	24.5
Excessive alcohol consumption, women	1.0	3.7	3.4	3.9
Average body mass index, men	23.4	24.6	25.0	24.8
Average body mass index, women	21.7	23.0	23.4	24.7
Material factors				
Severe financial problems	1.9	2.4	3.3	7.5
Labour market position				
Long-term work disability	2.5	3.7	5.7	11.4
Income proxy[c]				
Rented house, car, public health insurance	8.8	17.3	28.4	38.7
Rented house, no car, public health insurance	5.3	6.0	9.1	18.7
3 complaints about dwelling	0.6	1.2	1.9	2.8

van Lenthe et al. 2004 [2]

Ex: SEP and CVD in Netherlands

Many observed differences. Is “no other unmeasured differences” credible?

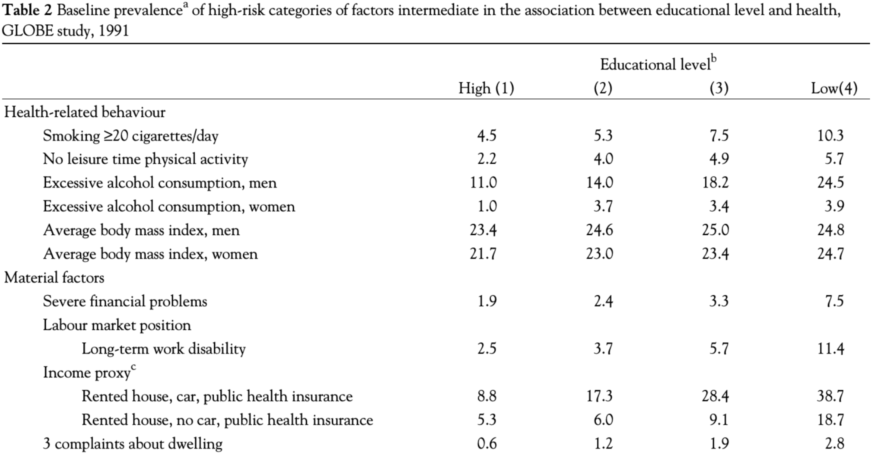


Table 2 Baseline prevalence[a] of high-risk categories of factors intermediate in the association between educational level and health, GLOBE study, 1991

	Educational level[b]			
	High (1)	(2)	(3)	Low(4)
Health-related behaviour				
Smoking ≥20 cigarettes/day	4.5	5.3	7.5	10.3
No leisure time physical activity	2.2	4.0	4.9	5.7
Excessive alcohol consumption, men	11.0	14.0	18.2	24.5
Excessive alcohol consumption, women	1.0	3.7	3.4	3.9
Average body mass index, men	23.4	24.6	25.0	24.8
Average body mass index, women	21.7	23.0	23.4	24.7
Material factors				
Severe financial problems	1.9	2.4	3.3	7.5
Labour market position				
Long-term work disability	2.5	3.7	5.7	11.4
Income proxy[c]				
Rented house, car, public health insurance	8.8	17.3	28.4	38.7
Rented house, no car, public health insurance	5.3	6.0	9.1	18.7
3 complaints about dwelling	0.6	1.2	1.9	2.8

van Lenthe et al. 2004 [2]

Is credibility is getting harder to sell?

- Another example: Does breastfeeding increase child IQ?
- Several observational studies show higher IQs for breastfed children.
  - "The authors of this and other studies claim to find effects of breastfeeding because even once they adjust for the differences they see across women, the effects persist. But this assumes that the adjustments they do are able to remove all of the differences across women. This is extremely unlikely to be the case."
  - "I would argue that in the case of breastfeeding, this issue is impossible to ignore and therefore any study that simply compares breast-fed to formula-fed infants is deeply flawed. That doesn't mean the results from such studies are necessarily wrong, just that we can't learn much from them."

Oster (2015). http://fivethirtyeight.com/features/everybody-calm-down-about-breastfeeding/

Is credibility is getting harder to sell?

- Another example: Does breastfeeding increase child IQ?
- Several observational studies show higher IQs for breastfed children.
  - “The authors of this and other studies claim to find effects of breastfeeding because even once they adjust for the differences they see across women, the effects persist. But this assumes that the adjustments they do are able to remove all of the differences across women. This is extremely unlikely to be the case.”
  - “I would argue that in the case of breastfeeding, this issue is impossible to ignore and therefore any study that simply compares breast-fed to formula-fed infants is deeply flawed. That doesn’t mean the results from such studies are necessarily wrong, just that we can’t learn much from them.”

Oster (2015). http://fivethirtyeight.com/features/everybody-calm-down-about-breastfeeding/

Is credibility is getting harder to sell?

- Another example: Does breastfeeding increase child IQ?
- Several observational studies show higher IQs for breastfed children.
  - "The authors of this and other studies claim to find effects of breastfeeding because even once they adjust for the differences they see across women, the effects persist. But this assumes that the adjustments they do are able to remove all of the differences across women. This is extremely unlikely to be the case."
  - "I would argue that in the case of breastfeeding, this issue is impossible to ignore and therefore any study that simply compares breast-fed to formula-fed infants is deeply flawed. That doesn't mean the results from such studies are necessarily wrong, just that we can't learn much from them."

Oster (2015). http://fivethirtyeight.com/features/everybody-calm-down-about-breastfeeding/

Is credibility is getting harder to sell?

- Another example: Does breastfeeding increase child IQ?
- Several observational studies show higher IQs for breastfed children.
  - "The authors of this and other studies claim to find effects of breastfeeding because even once they adjust for the differences they see across women, the effects persist. But this assumes that the adjustments they do are able to remove all of the differences across women. This is extremely unlikely to be the case."
  - "I would argue that in the case of breastfeeding, this issue is impossible to ignore and therefore any study that simply compares breast-fed to formula-fed infants is deeply flawed. That doesn't mean the results from such studies are necessarily wrong, just that we can't learn much from them."

Oster (2015). http://fivethirtyeight.com/features/everybody-calm-down-about-breastfeeding/

Is credibility is getting harder to sell?

- Another example: Does breastfeeding increase child IQ?
- Several observational studies show higher IQs for breastfed children.
    - "The authors of this and other studies claim to find effects of breastfeeding because even once they adjust for the differences they see across women, the effects persist. But this assumes that the adjustments they do are able to remove all of the differences across women. This is extremely unlikely to be the case."
    - "I would argue that in the case of breastfeeding, this issue is impossible to ignore and therefore any study that simply compares breast-fed to formula-fed infants is deeply flawed. That doesn't mean the results from such studies are necessarily wrong, just that we can't learn much from them."

Oster (2015). http://fivethirtyeight.com/features/everybody-calm-down-about-breastfeeding/

Outline

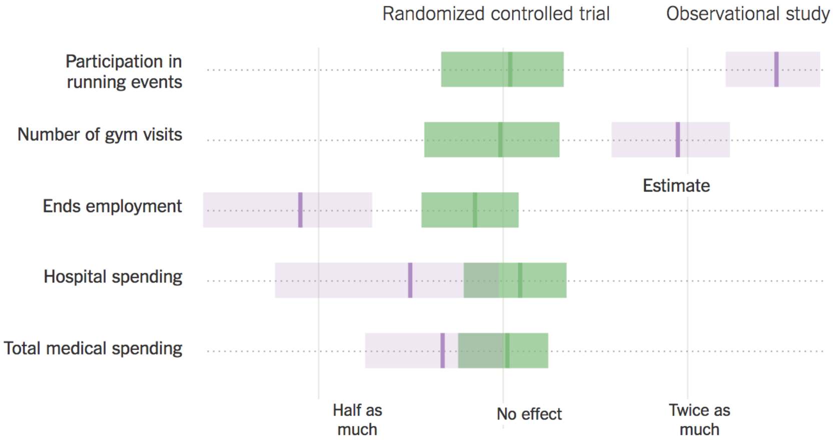
1. Problems with non-randomized studies
2. Natural and quasi-experiments
3. Example study designs
   - Difference-in-Differences
   - Instrumental variables
   - Regression discontinuity
4. Concluding thoughts
5. References

Why we worry about observational studies

- Recent evaluation of “Workplace Wellness” program in US state of Illinois
- Treatment: biometric health screening; online health risk assessment, access to a wide variety of wellness activities (e.g., smoking cessation, stress management, and recreational classes).
- Randomized evaluation:
  - 3,300 individuals assigned treated group.
  - 1,534 assigned to control (could not access the program).
- Also analyzed as an observational study:
  - comparing “participants” vs. non-participants in treated group.

Jones et al. 2018

How the Illinois Wellness Program Affected . . .



Carroll, *New York Times*, Aug 6, 2018.

How can natural experiments help?

- Natural experiments mimic RCTs.

- Usually not "natural", and they are observational studies, not experiments.

- Typically "accidents of chance" that create:
  1. Comparable treated and control units
  2. Random or "as-if" random assignment to treatment.

How can natural experiments help?

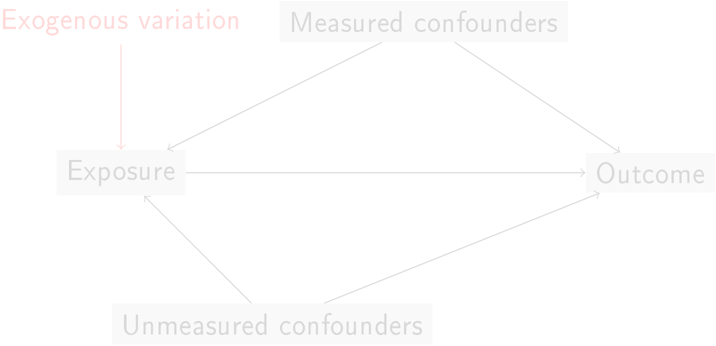
- Natural experiments mimic RCTs.
- Usually not “natural”, and they are observational studies, not experiments.
- Typically “accidents of chance” that create:
    1. Comparable treated and control units
    2. Random or “as-if” random assignment to treatment.

How can natural experiments help?

- Natural experiments mimic RCTs.
- Usually not “natural”, and they are observational studies, not experiments.
- Typically “accidents of chance” that create:
    1. Comparable treated and control units
    2. Random or “as-if” random assignment to treatment.

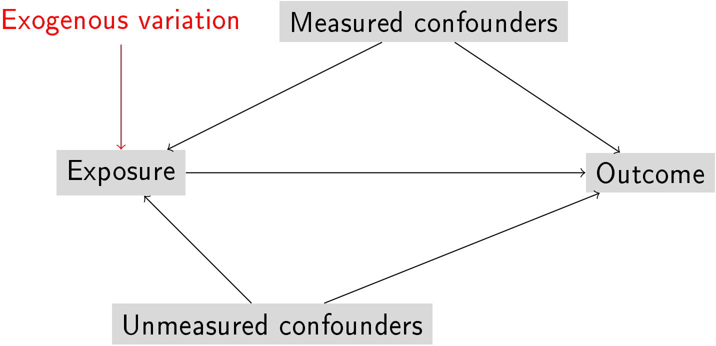
Selection on "observables" and "unobservables"

- Observables: Things you measured or can measure.
- Unobservables: Things you can't measure (e.g., innate abilities).
- Exogenous variation: predicts exposure but (**we assume**) not associated with anything else [mimicking random assignment].



Selection on “observables” and “unobservables”

- Observables: Things you measured or can measure.
- Unobservables: Things you can't measure (e.g., innate abilities).
- Exogenous variation: predicts exposure but (**we assume**) not associated with anything else [mimicking random assignment].



Strategies based on observables and unobservables

- Most observational study designs control for *measured* factors using:
  - Stratification (tabular analysis)
  - Adjustment (usually OLS regression)
  - Matching (pre-processing to create treated and control groups)

- Quasi-experimental strategies aim to control for some *unmeasured* factors using:
  - Interrupted time series (ITS)
  - Difference-in-differences (DD)
  - Synthetic controls (SC)
  - Instrumental variables (IV)
  - Regression discontinuity (RD)

- Selecting on "unobservables" = natural experiments

Strategies based on observables and unobservables

- Most observational study designs control for *measured* factors using:
  - Stratification (tabular analysis)
  - Adjustment (usually OLS regression)
  - Matching (pre-processing to create treated and control groups)

- Quasi-experimental strategies aim to control for some *unmeasured* factors using:
  - Interrupted time series (ITS)
  - Difference-in-differences (DD)
  - Synthetic controls (SC)
  - Instrumental variables (IV)
  - Regression discontinuity (RD)

- Selecting on "unobservables" = natural experiments

Strategies based on observables and unobservables

- Most observational study designs control for *measured* factors using:
  - Stratification (tabular analysis)
  - Adjustment (usually OLS regression)
  - Matching (pre-processing to create treated and control groups)

- Quasi-experimental strategies aim to control for some *unmeasured* factors using:
  - Interrupted time series (ITS)
  - Difference-in-differences (DD)
  - Synthetic controls (SC)
  - Instrumental variables (IV)
  - Regression discontinuity (RD)

- Selecting on "unobservables" = natural experiments

Natural experiments and quasi-experiments

- These lines are a little blurry, and the terms are sometimes used interchangeably. Dunning [3] makes a clear distinction:

Natural experiments

Treatment groups are random or "as if" randomly assigned, but not by the investigator.

- Ex: lotteries, arbitrary treatment discontinuities, weather shocks.

Quasi-experiments

Treatment groups are not random or "as if" random. Usually require more controls and assumptions for "as if" random.

- Assignment clearly not random, but may make a convincing case with added design features.

Natural experiments and quasi-experiments

- These lines are a little blurry, and the terms are sometimes used interchangeably. Dunning [3] makes a clear distinction:

Natural experiments

Treatment groups are random or "as if" randomly assigned, but not by the investigator.

- Ex: lotteries, arbitrary treatment discontinuities, weather shocks.

Quasi-experiments

Treatment groups are not random or "as if" random. Usually require more controls and assumptions for "as if" random.

- Assignment clearly not random, but may make a convincing case with added design features.

Natural experiments and quasi-experiments

- These lines are a little blurry, and the terms are sometimes used interchangeably. Dunning [3] makes a clear distinction:

Natural experiments

Treatment groups are random or “as if” randomly assigned, but not by the investigator.

- Ex: lotteries, arbitrary treatment discontinuities, weather shocks.

Quasi-experiments

Treatment groups are not random or “as if” random. Usually require more controls and assumptions for “as if” random.

- Assignment clearly not random, but may make a convincing case with added design features.

Some *potential* sources of natural experiments

- Law changes
- Eligibility for social programs (roll-outs)
- Lotteries
- Genes
- Weather shocks (rainfall, disasters)
- Arbitrary policy or clinical guidelines (thresholds)
- Factory or business closures
- Historical legacies (physical environment)
- Seasonality

Outline

1. Problems with non-randomized studies

2. Natural and quasi-experiments

3. Example study designs
   - Difference-in-Differences
   - Instrumental variables
   - Regression discontinuity

4. Concluding thoughts

5. References

Difference-in-Differences

Thinking about research design

- Approaches using natural or quasi-experiments focus on exploiting:
  1. A treatment group that experiences a **change** in the exposure of interest.
  2. Comparison with an appropriate control group that does not experience a change in exposure.

- In order to say something about the effect of the treatment, we need a substitute (control) population.

- Where should we get our counterfactual?

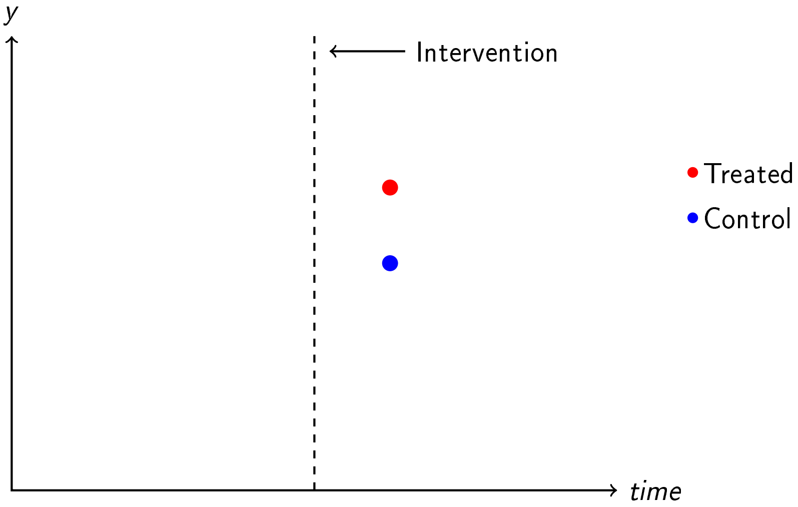
Thinking about research design

- Approaches using natural or quasi-experiments focus on exploiting:
  1. A treatment group that experiences a **change** in the exposure of interest.
  2. Comparison with an appropriate control group that does not experience a change in exposure.

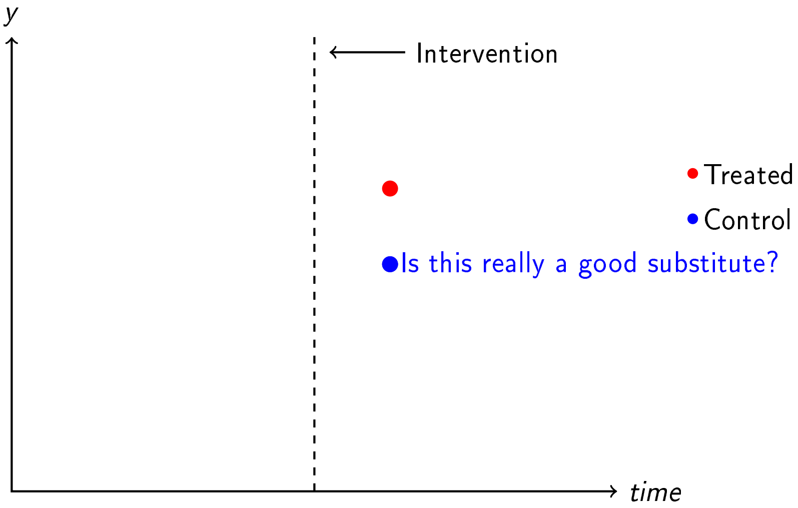
- In order to say something about the effect of the treatment, we need a substitute (control) population.

- Where should we get our counterfactual?

One-group posttest design with control group



One-group posttest design with control group



●Is this really a good substitute?

Problems comparing non-randomized treated and controls

- Treated and controls may have different characteristics and it may be those characteristics rather than the program that explain the difference in outcomes between the two groups (i.e., confounding/endogeneity).
- We could try to measure some observed characteristics that differ between the two groups.
- But we can't measure everything, and unobserved differences are often a concern (think about people who take advantage of policies).
- By definition, it is impossible for us to include unobserved differences in characteristics in the analysis.
- Could instead measure the treated group before the intervention.

Problems comparing non-randomized treated and controls

- Treated and controls may have different characteristics and it may be those characteristics rather than the program that explain the difference in outcomes between the two groups (i.e., confounding/endogeneity).
- We could try to measure some observed characteristics that differ between the two groups.
- But we can't measure everything, and unobserved differences are often a concern (think about people who take advantage of policies).
- By definition, it is impossible for us to include unobserved differences in characteristics in the analysis.
- Could instead measure the treated group before the intervention.

Problems comparing non-randomized treated and controls

- Treated and controls may have different characteristics and it may be those characteristics rather than the program that explain the difference in outcomes between the two groups (i.e., confounding/endogeneity).
- We could try to measure some observed characteristics that differ between the two groups.
- But we can't measure everything, and unobserved differences are often a concern (think about people who take advantage of policies).
- By definition, it is impossible for us to include unobserved differences in characteristics in the analysis.
- Could instead measure the treated group before the intervention.

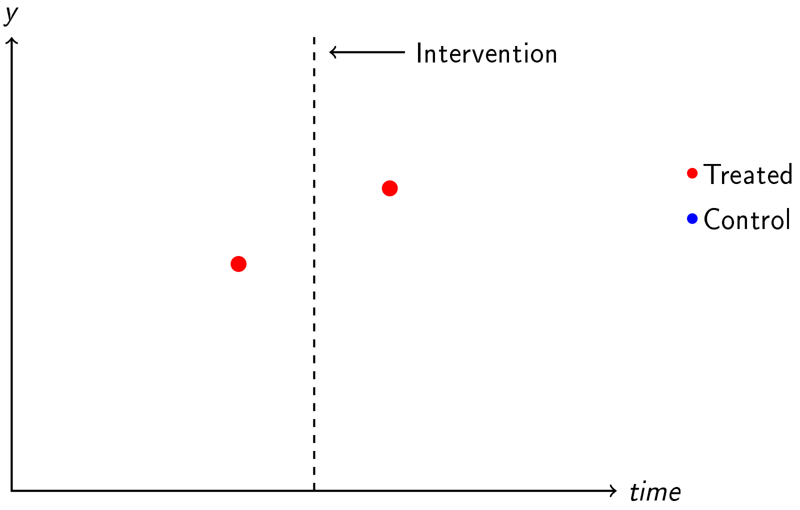
Problems comparing non-randomized treated and controls

- Treated and controls may have different characteristics and it may be those characteristics rather than the program that explain the difference in outcomes between the two groups (i.e., confounding/endogeneity).
- We could try to measure some observed characteristics that differ between the two groups.
- But we can't measure everything, and unobserved differences are often a concern (think about people who take advantage of policies).
- By definition, it is impossible for us to include unobserved differences in characteristics in the analysis.
- Could instead measure the treated group before the intervention.

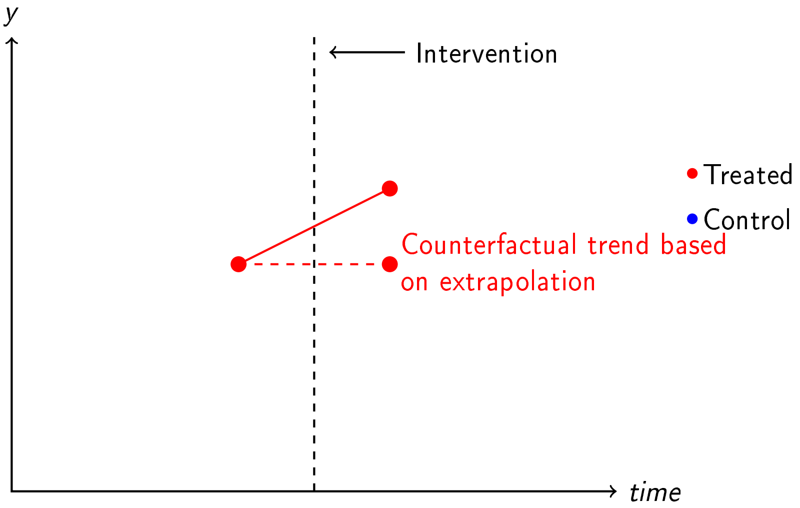
Problems comparing non-randomized treated and controls

- Treated and controls may have different characteristics and it may be those characteristics rather than the program that explain the difference in outcomes between the two groups (i.e., confounding/endogeneity).
- We could try to measure some observed characteristics that differ between the two groups.
- But we can't measure everything, and unobserved differences are often a concern (think about people who take advantage of policies).
- By definition, it is impossible for us to include unobserved differences in characteristics in the analysis.
- Could instead measure the treated group before the intervention.

One-group pretest-posttest design



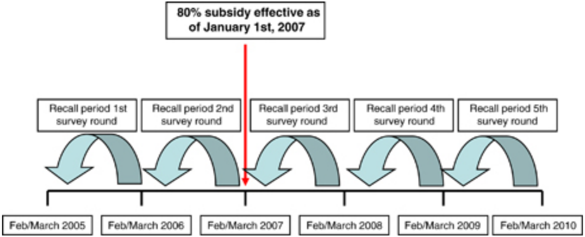
One-group pretest-posttest design



What is the impact of this program?

De Allegri et al. The **impact** of targeted subsidies for facility-based delivery on access to care and equity – Evidence from a population-based study in rural Burkina Faso. *J Public Health Policy* 2012;33:439–453

> *...the first population-based impact assessment of a financing policy introduced in Burkina Faso in 2007 on women's access to delivery services. The policy offers an 80 per cent subsidy for facility-based delivery. We collected information on delivery... from 2006 to 2010 on a representative sample of 1050 households in rural Nouna Health District. Over the 5 years, the proportion of facility-based deliveries increased from 49 to 84 per cent (P<0.001).*


80% subsidy effective as of January 1st, 2007
Recall period 1st survey round
Recall period 2nd survey round
Recall period 3rd survey round
Recall period 4th survey round
Recall period 5th survey round
Feb/March 2005
Feb/March 2006
Feb/March 2007
Feb/March 2008
Feb/March 2009
Feb/March 2010

One group pretest-posttest design

- Even a single pretest observation provides some improvement over the posttest only design.
- Now we derive a counterfactual prediction from the same group before the intervention.
- Provides weak counterfactual evidence about what would have happened in the absence of the program.
  - We know that Y_{t-1} occurs before Y_t (correct temporal ordering).
  - Could be many other reasons apart from the intervention that $Y_t \neq Y_{t-1}$.
- Stronger evidence if the outcomes can be reliably predicted and the pre-post interval is short.
- Better still to add a pretest and posttest from a control group.

One group pretest-posttest design

- Even a single pretest observation provides some improvement over the posttest only design.
- Now we derive a counterfactual prediction from the same group before the intervention.
- Provides weak counterfactual evidence about what would have happened in the absence of the program.
    - We know that Y_{t-1} occurs before Y_t (correct temporal ordering).
    - Could be many other reasons apart from the intervention that $Y_t \neq Y_{t-1}$.
- Stronger evidence if the outcomes can be reliably predicted and the pre-post interval is short.
- Better still to add a pretest and posttest from a control group.

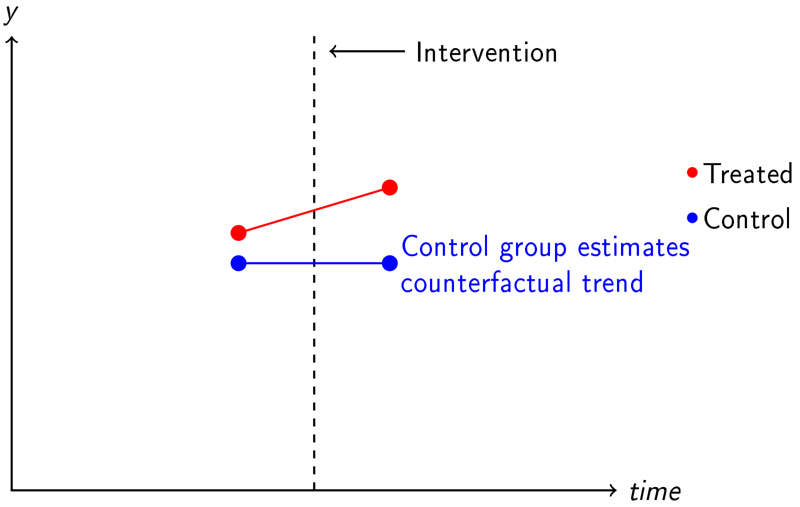
One group pretest-posttest design

- Even a single pretest observation provides some improvement over the posttest only design.
- Now we derive a counterfactual prediction from the same group before the intervention.
- Provides weak counterfactual evidence about what would have happened in the absence of the program.
  - We know that Y_{t-1} occurs before Y_t (correct temporal ordering).
  - Could be many other reasons apart from the intervention that $Y_t \neq Y_{t-1}$.
- Stronger evidence if the outcomes can be reliably predicted and the pre-post interval is short.
- Better still to add a pretest and posttest from a control group.

One group pretest-posttest design

- Even a single pretest observation provides some improvement over the posttest only design.
- Now we derive a counterfactual prediction from the same group before the intervention.
- Provides weak counterfactual evidence about what would have happened in the absence of the program.
    - We know that Y_{t-1} occurs before Y_t (correct temporal ordering).
    - Could be many other reasons apart from the intervention that $Y_t \neq Y_{t-1}$.
- Stronger evidence if the outcomes can be reliably predicted and the pre-post interval is short.
- Better still to add a pretest and posttest from a control group.

Adding pretests for both groups



How does this help?

- Pre/post in a control group helps resolve this by differencing out any **time-invariant** characteristics of both groups.
    - Many observed factors don't change over the course of an intervention (e.g., geography, parents' social class, birth cohort).
    - Any time-invariant *unobserved* factors also won't change over intervention period.
    - We can therefore effectively control for them.

- Measuring same units before and after a program cancels out any effect of all of the characteristics that are unique to units of observation and that do not change over time.

How does this help?

- Pre/post in a control group helps resolve this by differencing out any **time-invariant** characteristics of both groups.
  - Many observed factors don't change over the course of an intervention (e.g., geography, parents' social class, birth cohort).
  - Any time-invariant *unobserved* factors also won't change over intervention period.
  - We can therefore effectively control for them.
- Measuring same units before and after a program cancels out any effect of all of the characteristics that are unique to units of observation and that do not change over time.

Difference-in-Differences: Basic Idea

- The average change over time in the non-exposed (control) group is subtracted from the change over time in the exposed (treatment) group.

- Double differencing removes biases in second period comparisons between the treatment and control group that could result from:

1. Fixed (i.e., non time-varying) differences between those groups.
2. Comparisons over time in both groups that could be the result of time trends unrelated to the treatment.

Difference-in-Differences: Basic Idea

- The average change over time in the non-exposed (control) group is subtracted from the change over time in the exposed (treatment) group.

- Double differencing removes biases in second period comparisons between the treatment and control group that could result from:

1. Fixed (i.e., non time-varying) differences between those groups.
2. Comparisons over time in both groups that could be the result of time trends unrelated to the treatment.

Causal effects without regression?

Good natural experiments are also transparent. Can also be analyzed via differences in means. Let $\mu_{it} = E(Y_{it})$:

- $i = 0$ is control group, $i = 1$ is treatment.
- $t = 0$ is pre-period, $t = 1$ is post-period.
- One 'difference' estimate of causal effect is: μ_{11}—μ_{10} (pre-post in treated)
- Differences-in-Differences estimate of causal effect is: $(\mu_{11} - \mu_{10}) - (\mu_{01} - \mu_{00})$

	Policy Change		
Area	Before	After	Difference (A - B)
Treated	135	100	-35
Control	80	60	-20
T - C	55	40	-15

Key Assumption: Parallel Pre-Intervention Trends

- Basic DD controls for any time invariant characteristics of both treated and control groups.
- Does not control for any **time-varying** characteristics.
- If another policy/intervention occurs in the treated (or control) group at the same time as the intervention, we cannot cleanly identify the effect of the program.
- DD main assumption: in the absence of the intervention treated and control groups would have displayed equal **trends**.
- Impossible to verify.

Key Assumption: Parallel Pre-Intervention Trends

- Basic DD controls for any time invariant characteristics of both treated and control groups.
- Does not control for any **time-varying** characteristics.
- If another policy/intervention occurs in the treated (or control) group at the same time as the intervention, we cannot cleanly identify the effect of the program.
- DD main assumption: in the absence of the intervention treated and control groups would have displayed equal **trends**.
- Impossible to verify.

Key Assumption: Parallel Pre-Intervention Trends

- Basic DD controls for any time invariant characteristics of both treated and control groups.
- Does not control for any **time-varying** characteristics.
- If another policy/intervention occurs in the treated (or control) group at the same time as the intervention, we cannot cleanly identify the effect of the program.
- DD main assumption: in the absence of the intervention treated and control groups would have displayed equal **trends**.
- Impossible to verify.

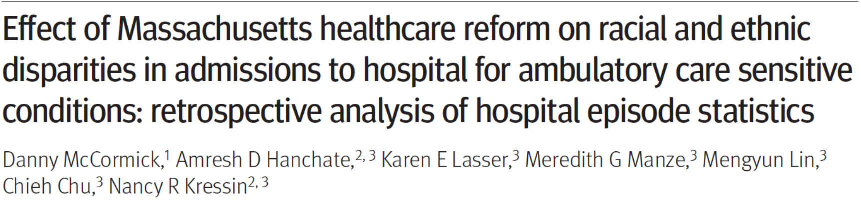
Key Assumption: Parallel Pre-Intervention Trends

- Basic DD controls for any time invariant characteristics of both treated and control groups.
- Does not control for any **time-varying** characteristics.
- If another policy/intervention occurs in the treated (or control) group at the same time as the intervention, we cannot cleanly identify the effect of the program.
- DD main assumption: in the absence of the intervention treated and control groups would have displayed equal **trends**.
- Impossible to verify.

Key Assumption: Parallel Pre-Intervention Trends

- Basic DD controls for any time invariant characteristics of both treated and control groups.
- Does not control for any **time-varying** characteristics.
- If another policy/intervention occurs in the treated (or control) group at the same time as the intervention, we cannot cleanly identify the effect of the program.
- DD main assumption: in the absence of the intervention treated and control groups would have displayed equal **trends**.
- Impossible to verify.

A social epidemiology example



Effect of Massachusetts healthcare reform on racial and ethnic disparities in admissions to hospital for ambulatory care sensitive conditions: retrospective analysis of hospital episode statistics

Danny McCormick,[1] Amresh D Hanchate,[2,3] Karen E Lasser,[3] Meredith G Manze,[3] Mengyun Lin,[3] Chieh Chu,[3] Nancy R Kressin[2,3]

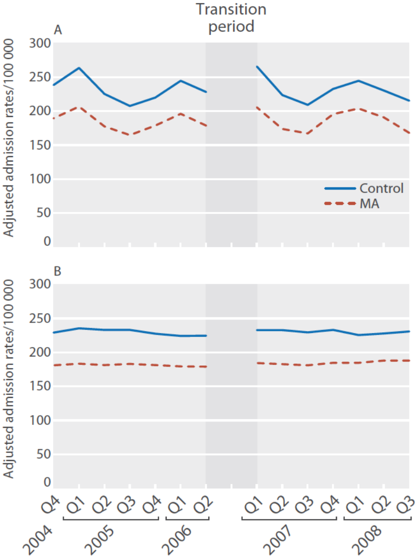
- Evaluated impact of MA reform on inequalities in hospital admissions.
- Compared MA to nearby states: NY, NJ, PA.
- Intervention “worked”: % uninsured halved (12% to 6%) from 2004-06 to 2008-09.

McCormick et al. 2015 [4]

Evaluating pre-intervention trends

- Adds credibility to assumption that post-intervention trends would have been similar in the absence of the intervention.

- “Null” results help focus on alternative mechanisms linking disadvantage to hospital admissions.



Key Considerations

- Choose an appropriate control group
  - Investigate the data in the pre-period
  - Common trends in the outcome of interest are more important than common levels
  - Verify whether the composition of the groups changes as a result of the exposure (migration)
- Investigate the exogeneity of your treatment
  - Investigate why the change occurred (qualitative research).
  - Pre-period data are important here too.

Instrumental variables

Why use instrumental variables?

- Trial may be impossible or unethical (especially for many social exposures)
- We may actually want to know the effect of T on Y.
- We are concerned about unmeasured confounding for the effect of T on Y.
- Many examples of social exposures where this is problematic:
  - Education
  - Income
  - Health behaviors
  - Policies/programs

Why use instrumental variables?

- Trial may be impossible or unethical (especially for many social exposures)
- We may actually want to know the effect of T on Y.
- We are concerned about unmeasured confounding for the effect of T on Y.
- Many examples of social exposures where this is problematic:
  - Education
  - Income
  - Health behaviors
  - Policies/programs

Why use instrumental variables?

- Trial may be impossible or unethical (especially for many social exposures)
- We may actually want to know the effect of T on Y.
- We are concerned about unmeasured confounding for the effect of T on Y.
- Many examples of social exposures where this is problematic:
  - Education
  - Income
  - Health behaviors
  - Policies/programs

Why use instrumental variables?

- Trial may be impossible or unethical (especially for many social exposures)
- We may actually want to know the effect of T on Y.
- We are concerned about unmeasured confounding for the effect of T on Y.
- Many examples of social exposures where this is problematic:
  - Education
  - Income
  - Health behaviors
  - Policies/programs

Thinking about research design

- Remember that quasi-experimental designs and natural experiments are trying to mimic an RCT as closely as possible.
- In an RCT, the randomized assignment to treatment means we know that the only reason why outcomes might differ is because of the treatment.
- Can we find some variable in our real-world data that mimics randomized treatment assignment?

Thinking about research design

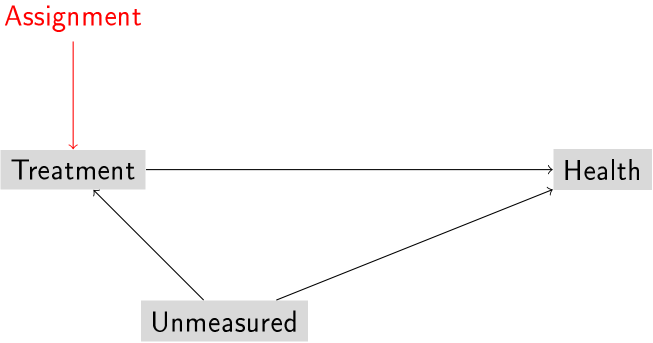
- Remember that quasi-experimental designs and natural experiments are trying to mimic an RCT as closely as possible.
- In an RCT, the randomized assignment to treatment means we know that the only reason why outcomes might differ is because of the treatment.
- Can we find some variable in our real-world data that mimics randomized treatment assignment?

Thinking about research design

- Remember that quasi-experimental designs and natural experiments are trying to mimic an RCT as closely as possible.
- In an RCT, the randomized assignment to treatment means we know that the only reason why outcomes might differ is because of the treatment.
- Can we find some variable in our real-world data that mimics randomized treatment assignment?

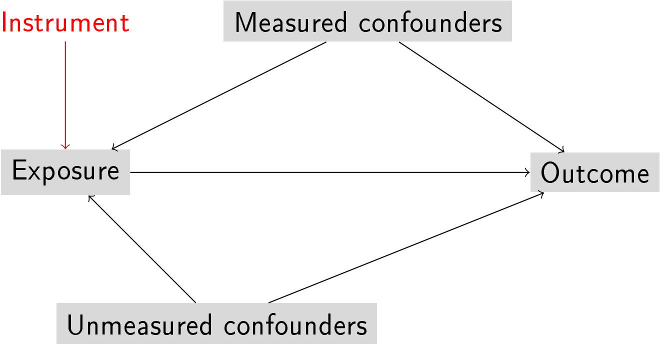
Hypothetical randomized assignment

- Does treatment (T, 1=yes, 0=no) affect health (Y)?
- “Instrumental variable”: random assignment.



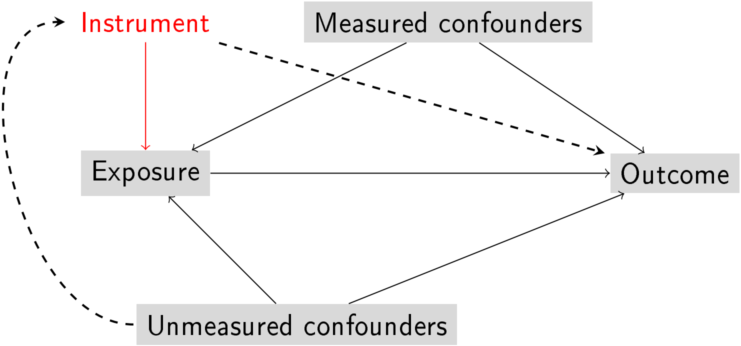
Non-randomized instrumental variable

- Does exposure (T, 1=yes, 0=no) affect health (Y)?
- "Instrumental variable": random or "as-if random" assignment, but not under investigator control.



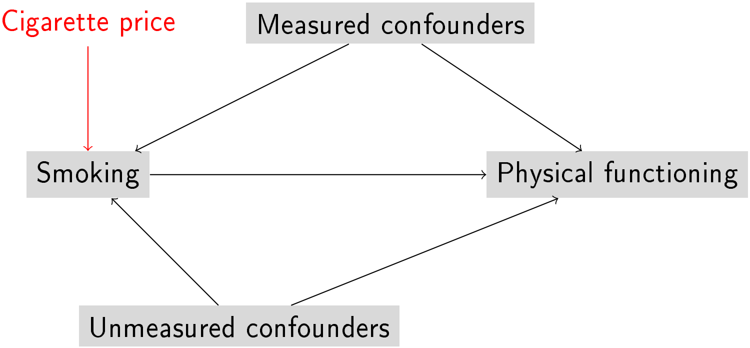
Non-randomized instrument creates additional issues

- In the RCT we know the treatment assignment is not associated directly with the outcome or with other unmeasured common causes.
- This assumption is less credible when our "instrument" is non-randomized.



Non-randomized examples of IV: Policies

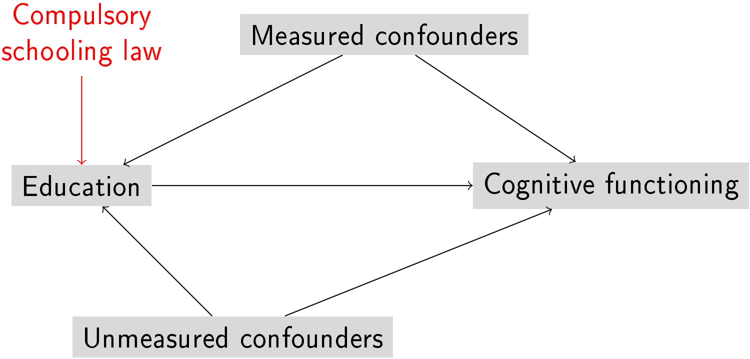
- Does smoking (T, 1=yes, 0=no) affect physical functioning (Y)?
- **Instrument**: changes in cigarette prices [mimicking random assignment].



Leigh and Schembri 2004 [5]

Non-randomized examples of IV: Policies

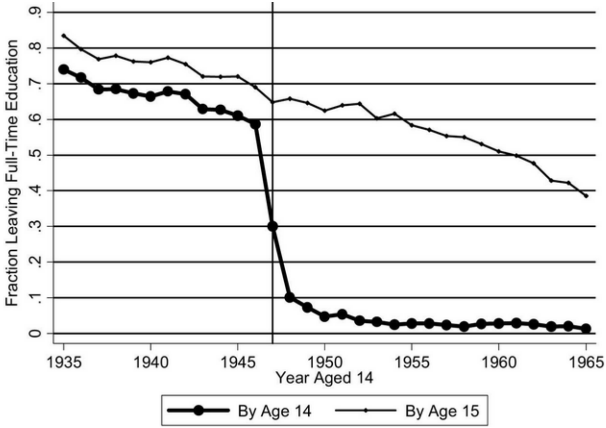
- Does education (T, 1=yes, 0=no) affect cognitive functioning (Y)?
- **Instrument**: changes in compulsory schooling laws [mimicking random assignment].



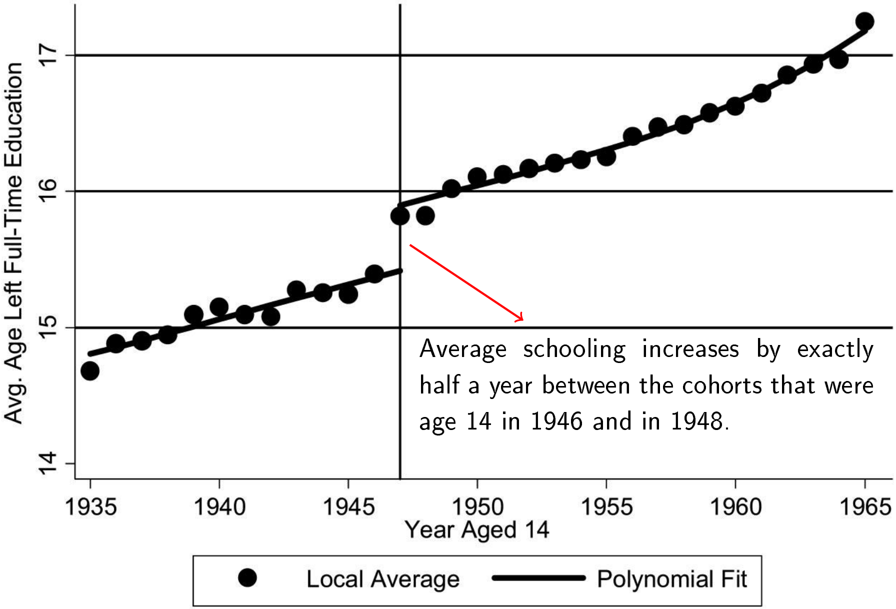
Glymour et al. 2008 [6]

What does a quasi-experiment look like?

Fraction left full-time education by year aged 14 and 15 (Great Britain)



The lower line shows the proportion of British-born adults aged 32 to 64 from the 1983 to 1998 General Household Surveys who report leaving full-time education at or before age 14 from 1935 to 1965. The upper line shows the same, but for age 15. The minimum school-leaving age in Great Britain changed in 1947 from 14 to 15 [Oreopoulos 2006].



Average schooling increases by exactly half a year between the cohorts that were age 14 in 1946 and in 1948.

Quasi Experiments: Education and Mortality

- Changes in education due to national or state-level changes in laws regarding compulsory schooling.
- Differ from the usual approach by attempting to focus on plausibly random changes in education, rather than comparing those achieving high vs. low education.
- Findings are heterogenous, in contrast to much of the evidence from observational studies:
  - USA (Lleras-Muney, 2005): IV (yes), RD (no)
  - UK (Oreopolous 2008, Clark 2010): RD (no)
  - France (Albouy 2009): RD (no)
  - Also positive and negative evidence for other health outcomes in Denmark, Sweden, Germany, Italy, Netherlands
- Importance of explicitly trying to mimic an RCT for education

Quasi Experiments: Education and Mortality

- Changes in education due to national or state-level changes in laws regarding compulsory schooling.
- Differ from the usual approach by attempting to focus on plausibly random changes in education, rather than comparing those achieving high vs. low education.
- Findings are heterogenous, in contrast to much of the evidence from observational studies:
  - USA (Lleras-Muney, 2005): IV (yes), RD (no)
  - UK (Oreopolous 2008, Clark 2010): RD (no)
  - France (Albouy 2009): RD (no)
  - Also positive and negative evidence for other health outcomes in Denmark, Sweden, Germany, Italy, Netherlands
- Importance of explicitly trying to mimic an RCT for education

Contents lists available at ScienceDirect

Social Science & Medicine

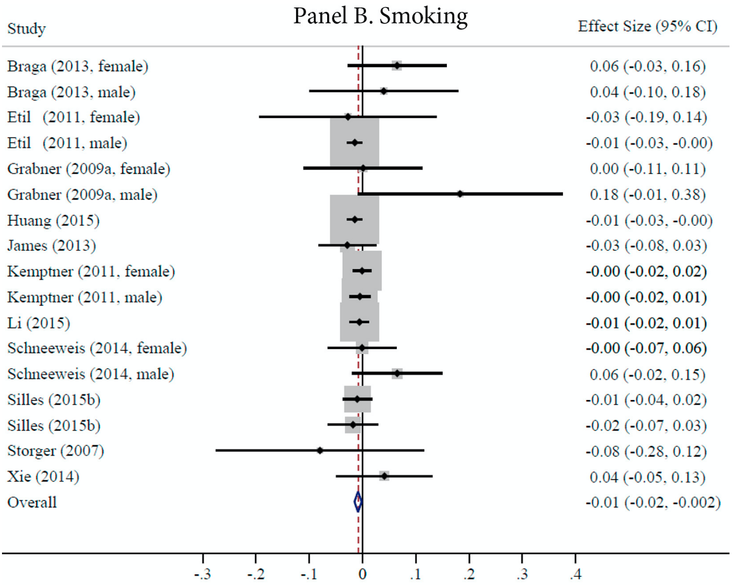
journal homepage: www.elsevier.com/locate/socscimed

Review article

How and why studies disagree about the effects of education on health: A systematic review and meta-analysis of studies of compulsory schooling laws

Rita Hamad[a,*], Holly Elser[b,c], Duy C. Tran[c], David H. Rehkopf[c], Steven N. Goodman[c]

[a] University of California San Francisco, Philip R. Lee Institute for Health Policy Studies, Department of Family & Community Medicine, 995 Potrero Avenue, Building 80, Ward 83, San Francisco, CA, 94110, USA
[b] University of California Berkeley, School of Public Health, Division of Epidemiology, Berkeley, CA, USA
[c] Stanford University, School of Medicine, Stanford, CA, USA



Panel B. Smoking

Questions to ask yourself if you want to use IV

- Is the exclusion restriction believable?
  - Would you expect a direct effect of Z on Y? Are there unobserved common causes of Z and Y?
  - Not directly testable
- What effect is being estimated?
  - Is this the one you would want?
  - Is it a quantity of theoretical interest?
  - Is it applicable in other contexts (generalizable)?

Regression Discontinuity

RD: Basic Idea

- Take advantage of arbitrary thresholds that sometimes assign treatment to individuals.
- When an administrative or rule-based cutoff in a continuous variable (present in your data) predicts treatment assignment, being on one side or the other of this cutoff determines, or predicts, treatment received.
- The continuous variable is called the “assignment” or “forcing” variable.
- Groups just on either side are the threshold considered “as good as randomly” assigned to treatment.

RD: Basic Idea

- Take advantage of arbitrary thresholds that sometimes assign treatment to individuals.
- When an administrative or rule-based cutoff in a continuous variable (present in your data) predicts treatment assignment, being on one side or the other of this cutoff determines, or predicts, treatment received.
- The continuous variable is called the "assignment" or "forcing" variable.
- Groups just on either side are the threshold considered "as good as randomly" assigned to treatment.

RD: Basic Idea

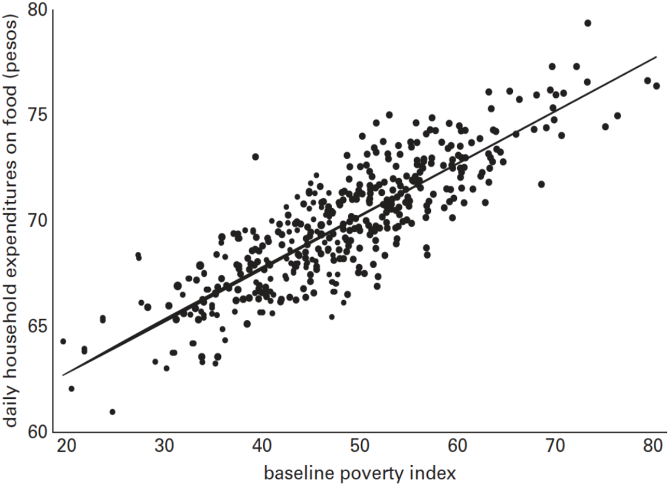
- Take advantage of arbitrary thresholds that sometimes assign treatment to individuals.
- When an administrative or rule-based cutoff in a continuous variable (present in your data) predicts treatment assignment, being on one side or the other of this cutoff determines, or predicts, treatment received.
- The continuous variable is called the “assignment” or “forcing” variable.
- Groups just on either side are the threshold considered “as good as randomly” assigned to treatment.

RD: Motivating example

- Suppose we want to estimate the impact of a cash transfer program on daily food expenditure of poor households.
- Poverty is measured by a continuous score between 0 and 100 that is used to rank households from poorest to richest.
- Poverty is the assignment variable, Z, that determines eligibility for the cash transfer program.
- The outcome of interest, daily food expenditure, is denoted by Y.

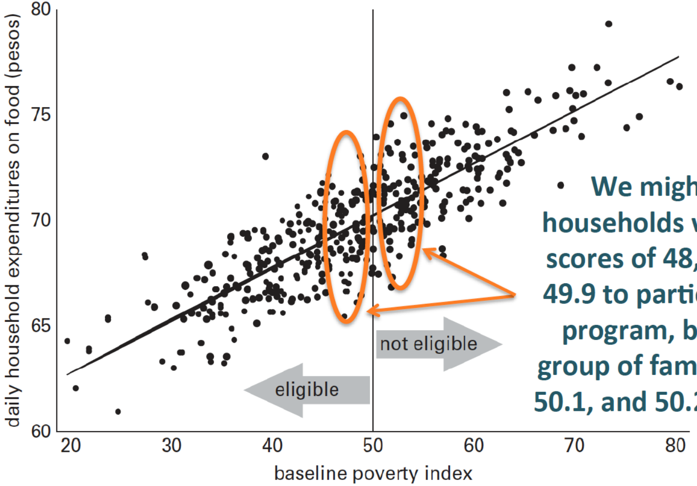
Source: Gertler, 2011[7]

At baseline, you might expect poorer households to spend less on food, on average, than richer ones, which might look like:



Source: Gertler, 2011[7]

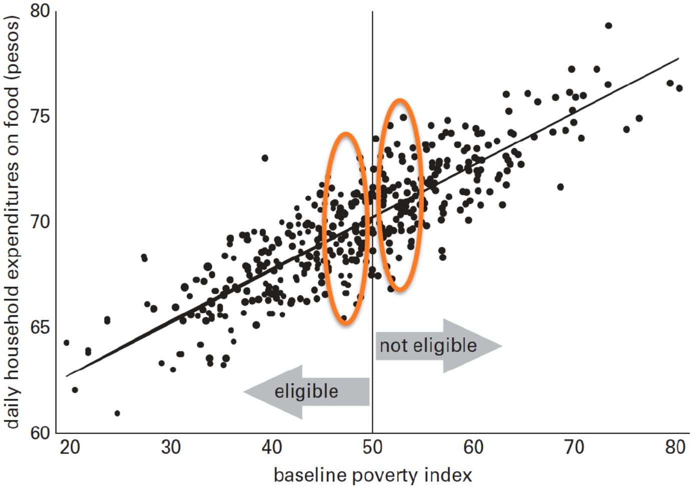
Under the program's rules, only households with a poverty score, Z, below 50 are eligible for the cash payment:



- We might expect households with poverty scores of 48, 49, or even 49.9 to participate in the program, but another group of families with 50, 50.1, and 50.2 acres won't

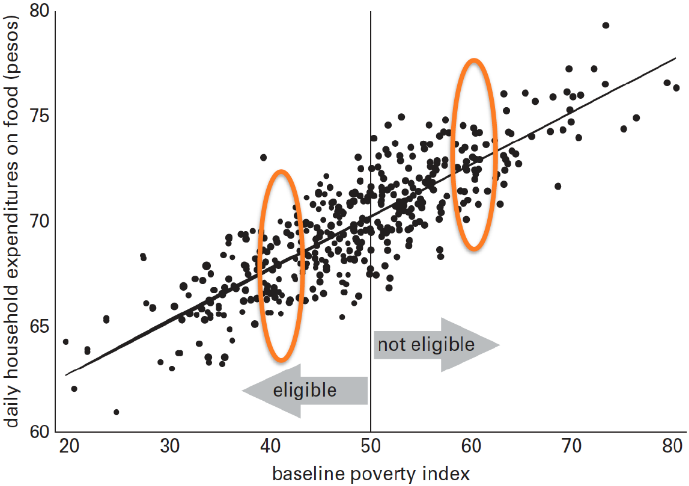
Source: Gertler, 2011[7]

Would you expect these two groups of families to be, on average, very different from one another? Why or why not?



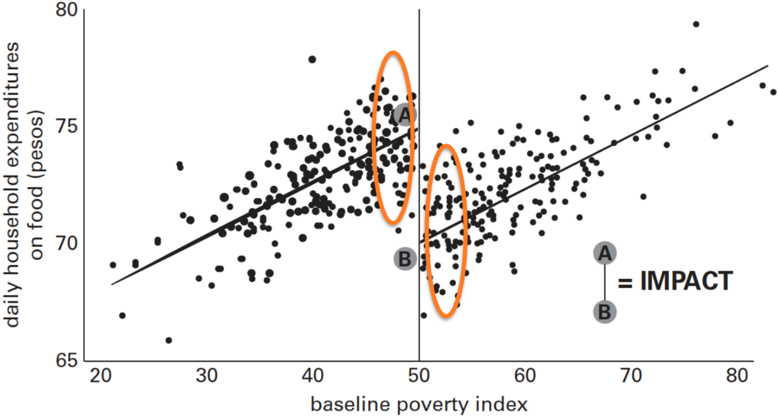
Source: Gertler, 2011[7]

How about these families?



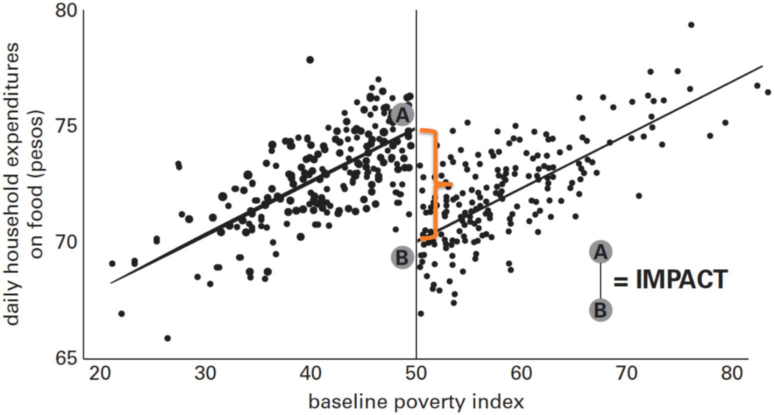
Source: Gertler, 2011[7]

As we approach the cutoff value from above and below, the individuals in both groups become more and more alike, on both measured and unobserved characteristics—in a small area around the threshold, the only difference is in treatment assignment



Source: Gertler, 2011[7]

RD measures the difference in post-intervention outcomes between units near the cutoff—those units that were just above the threshold and did not receive cash payments serve as the counterfactual comparison group



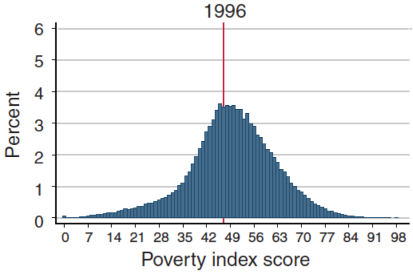
Source: Gertler, 2011[7]

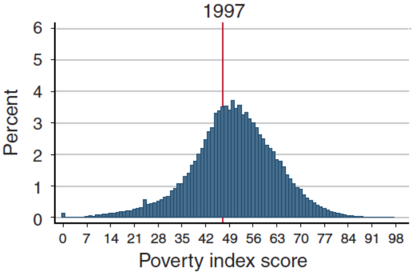
Assignment should be continuous at the cutoff

- In the simplest case, individuals have no control (e.g., birth date) and cannot manipulate the treatment assignment
- We must assume that units cannot manipulate the assignment variable to influence whether they receive treatment or not—the presence of manipulation can be assessed by examining the density of the assignment variable at the cutoff
- If individuals can modify their characteristics, such as household income, in order to qualify for the program, then groups on either side of the threshold may not be exchangeable
- Using a histogram of the assignment variable Z we can confirm that there is no "bunching", which would indicate manipulation.

Manipulation example (poverty threshold in Colombia)

- Colombian census collected comprehensive information on dwelling characteristics, demographics, income, and employment to assign a poverty index score to each family.
- Eligibility rules for several social welfare programs use specific thresholds (score=47) from the poverty index score.
- Prior to 1997, the precise algorithm was confidential:

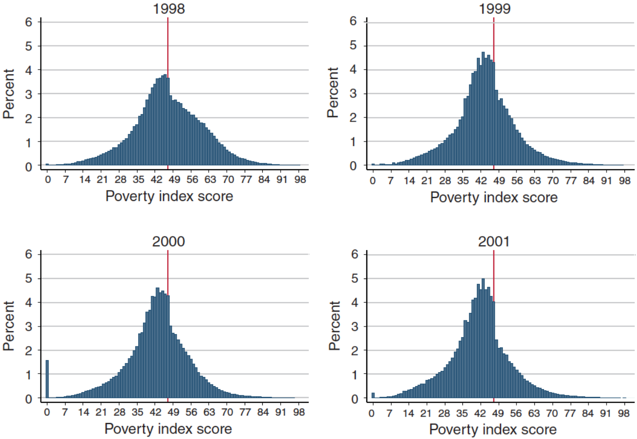




Camacho and Conover 2011[8]

Manipulation example (poverty threshold in Colombia)

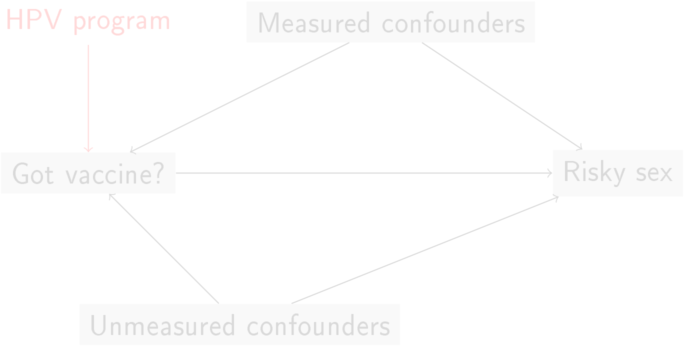
- After 1997, the algorithm was provided to municipal administrators, leading to evidence of manipulation:



Camacho and Conover 2011[8]

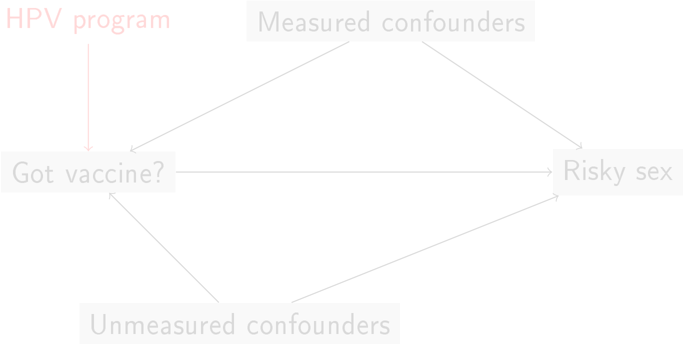
Applied example: HPV vaccine and sexual behaviors

- Does getting the HPV vaccine affect sexual behaviors?
- Vaccine policy: predicts vaccine receipt but (**we assume**) not associated with anything else [mimicking random assignment].



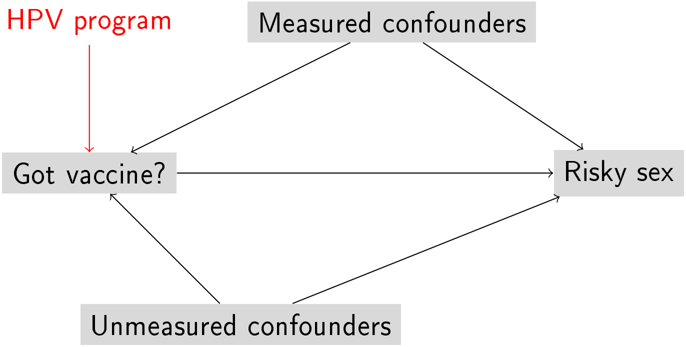
Applied example: HPV vaccine and sexual behaviors

- Does getting the HPV vaccine affect sexual behaviors?
- Vaccine policy: predicts vaccine receipt but (**we assume**) not associated with anything else [mimicking random assignment].



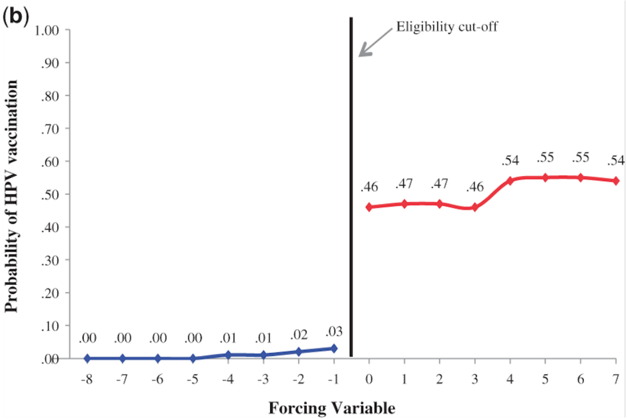
Applied example: HPV vaccine and sexual behaviors

- Does getting the HPV vaccine affect sexual behaviors?
- Vaccine policy: predicts vaccine receipt but (**we assume**) not associated with anything else [mimicking random assignment].



Does the cutoff predict treatment?

- Girls “assigned” to HPV program by quarter of birth.
- The probability of receiving the vaccine jumps discontinuously between eligibility groups at the eligibility cut-off.



Smith et al., 2015[9]

What does a credible natural experiment look like?

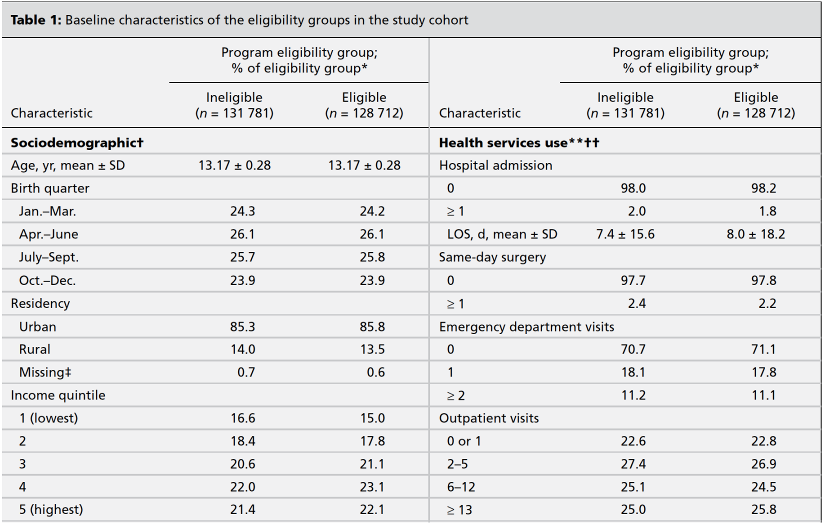


Table 1: Baseline characteristics of the eligibility groups in the study cohort

	Program eligibility group; % of eligibility group*			Program eligibility group; % of eligibility group*	
Characteristic	Ineligible (*n* = 131 781)	Eligible (*n* = 128 712)	Characteristic	Ineligible (*n* = 131 781)	Eligible (*n* = 128 712)
Sociodemographic†			**Health services use**††**		
Age, yr, mean ± SD	13.17 ± 0.28	13.17 ± 0.28	Hospital admission		
Birth quarter			0	98.0	98.2
Jan.–Mar.	24.3	24.2	≥ 1	2.0	1.8
Apr.–June	26.1	26.1	LOS, d, mean ± SD	7.4 ± 15.6	8.0 ± 18.2
July–Sept.	25.7	25.8	Same-day surgery		
Oct.–Dec.	23.9	23.9	0	97.7	97.8
Residency			≥ 1	2.4	2.2
Urban	85.3	85.8	Emergency department visits		
Rural	14.0	13.5	0	70.7	71.1
Missing‡	0.7	0.6	1	18.1	17.8
Income quintile			≥ 2	11.2	11.1
1 (lowest)	16.6	15.0	Outpatient visits		
2	18.4	17.8	0 or 1	22.6	22.8
3	20.6	21.1	2–5	27.4	26.9
4	22.0	23.1	6–12	25.1	24.5
5 (highest)	21.4	22.1	≥ 13	25.0	25.8

Smith et al., 2015[9]

Note little impact of adjustment

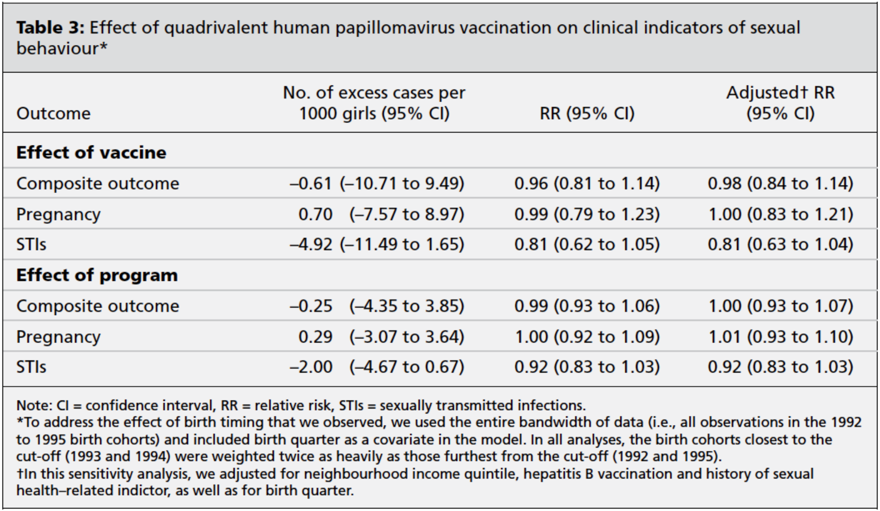


Table 3: Effect of quadrivalent human papillomavirus vaccination on clinical indicators of sexual behaviour*

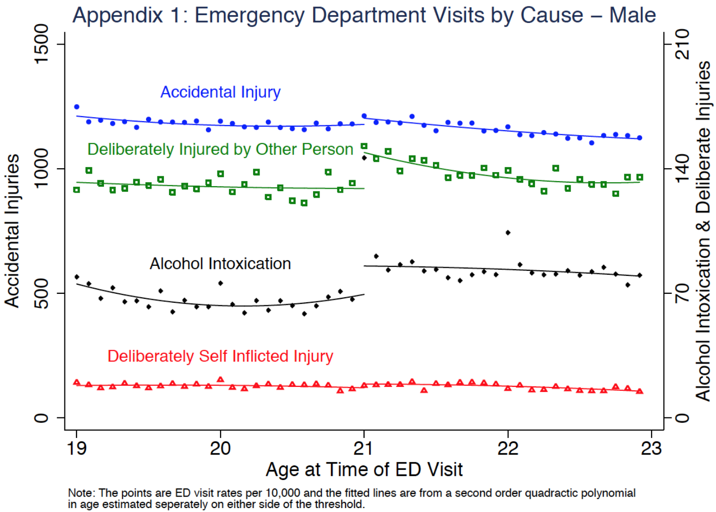
Outcome	No. of excess cases per 1000 girls (95% CI)	RR (95% CI)	Adjusted† RR (95% CI)
Effect of vaccine			
Composite outcome	–0.61 (–10.71 to 9.49)	0.96 (0.81 to 1.14)	0.98 (0.84 to 1.14)
Pregnancy	0.70 (–7.57 to 8.97)	0.99 (0.79 to 1.23)	1.00 (0.83 to 1.21)
STIs	–4.92 (–11.49 to 1.65)	0.81 (0.62 to 1.05)	0.81 (0.63 to 1.04)
Effect of program			
Composite outcome	–0.25 (–4.35 to 3.85)	0.99 (0.93 to 1.06)	1.00 (0.93 to 1.07)
Pregnancy	0.29 (–3.07 to 3.64)	1.00 (0.92 to 1.09)	1.01 (0.93 to 1.10)
STIs	–2.00 (–4.67 to 0.67)	0.92 (0.83 to 1.03)	0.92 (0.83 to 1.03)

Note: CI = confidence interval, RR = relative risk, STIs = sexually transmitted infections.
*To address the effect of birth timing that we observed, we used the entire bandwidth of data (i.e., all observations in the 1992 to 1995 birth cohorts) and included birth quarter as a covariate in the model. In all analyses, the birth cohorts closest to the cut-off (1993 and 1994) were weighted twice as heavily as those furthest from the cut-off (1992 and 1995).
†In this sensitivity analysis, we adjusted for neighbourhood income quintile, hepatitis B vaccination and history of sexual health–related indictor, as well as for birth quarter.

Smith et al.,2015[9]

Another recent example: US drinking age

- Minimum legal drinking age and non-fatal injuries:



Note: The points are ED visit rates per 10,000 and the fitted lines are from a second order quadractic polynomial in age estimated seperately on either side of the threshold.

Carpenter, 2017[10]

Issues related to generalizability

- RD estimates local average impacts around the eligibility cutoff where treated and control units are most similar and results cannot be generalized to units whose scores are further away from the cutoff (unless we assume treatment homogeneity).
- If the goal is to answer whether the program should exist or not, then RD is likely not the appropriate methodology.
- However, if the question is whether the program should be cut or expanded at the margin, then it produces the local estimate of interest to inform this policy decision

Best practices for RD study design

- Need to show convincingly that:
- Treatment changes discontinuously at the cutpoint.
  - Outcomes change discontinuously at the cutpoint.
  - Other covariates do not change discontinuously at the cutpoint.
  - There is no manipulation of the assignment variable.
- Need to argue that:
  - Unobserved factors don't change discontinuously at the cutoff.
  - Cases near the cutpoint are interesting to someone.

Best practices for RD study design

- Need to show convincingly that:
- Treatment changes discontinuously at the cutpoint.
    - Outcomes change discontinuously at the cutpoint.
    - Other covariates do not change discontinuously at the cutpoint.
    - There is no manipulation of the assignment variable.
- Need to argue that:
    - Unobserved factors don't change discontinuously at the cutoff.
    - Cases near the cutpoint are interesting to someone.

Outline

1. Problems with non-randomized studies
2. Natural and quasi-experiments
3. Example study designs
   - Difference-in-Differences
   - Instrumental variables
   - Regression discontinuity
4. Concluding thoughts
5. References

What are natural experiments good for?

1. To understand the effect of treatments *induced by policies* on outcomes, e.g., Policy → Treatment → Outcome:
   - Environmental exposures.
   - Education/income/financial resources.
   - Access to health care.
   - Health behaviors.

2. To understand the effect of policies on outcomes, e.g., Policy → Outcome:
   - Taxes, wages.
   - Environmental legislation.
   - Food policy.
   - Employment policy.
   - Civil rights legislation.

Glymour 2014 [11]

What are natural experiments good for?

1. To understand the effect of treatments *induced by policies* on outcomes, e.g., Policy → Treatment → Outcome:
   - Environmental exposures.
   - Education/income/financial resources.
   - Access to health care.
   - Health behaviors.

2. To understand the effect of policies on outcomes, e.g., Policy → Outcome:
   - Taxes, wages.
   - Environmental legislation.
   - Food policy.
   - Employment policy.
   - Civil rights legislation.

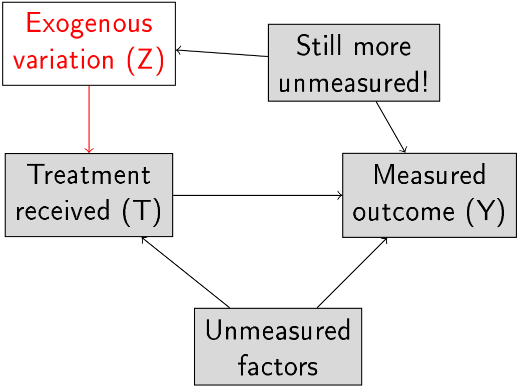
Glymour 2014 [11]

Are natural experiments always more credible?

- Not necessarily, but probably.
- Key is "as-if" randomization of treatment:
  - If this is credible, it is a much stronger **design** than most observational studies.
  - Should eliminate self-selection in to exposure groups.
- Allows for simple, transparent analysis of average differences between groups.
- Allows us to rely on weaker assumptions.

Assumptions still matter!

- Quasi-experimental studies are still observational.
- Most credible if they create unconditional randomized treatment groups (e.g., lottery).
- Credibility is continuous, not binary.
- I worry about the cognitive impact of the "quasi-experimental" label.



Potential drawbacks of quasi-experimental approaches

- How good is “as-if” random? (need “shoe-leather”)
- Credibility of additional (modeling) assumptions.
- Relevance of the intervention.
- Relevance of population.

How can we capitalize on natural experiments?

- Take “as-if random” seriously in all study designs.
- Find them.
- Teach them.
- Create them (aka increase dialogue with policymakers):
  - Challenges of observational evidence.
  - Great value of (“as-if”) randomization.
  - Policy roll-out with evaluation in mind.

Back to basics: assumptions and costs

- Major benefit of randomized evaluations are that few assumptions are needed to estimate a causal effect.
- Necessary assumptions can often be checked.
- Non-randomization means more assumptions, more possibility for assumptions to be violated.
- Should lead us to spend lots of time trying to test the credibility of these assumptions.
  - How good is "as-if random"?
  - Are there compelling non-causal alternative explanations for the observed results?
- All non-randomized designs are not created equal.

References I

[1] Michael D Keall, et al., “Home modifications to reduce injuries from falls in the home injury prevention intervention (HIPI) study: a cluster-randomised controlled trial”, *Lancet*, 385(9964), 2015, pp. 231–8.

[2] Frank J van Lenthe, et al., “Investigating explanations of socio-economic inequalities in health: the Dutch GLOBE study”, *Eur J Public Health*, 14(1), 2004, pp. 63–70.

[3] Thad Dunning, *Natural experiments in the social sciences: a design-based approach*, Strategies for social inquiry, Cambridge University Press, Cambridge, 2012.

[4] Danny McCormick, et al., “Effect of Massachusetts healthcare reform on racial and ethnic disparities in admissions to hospital for ambulatory care sensitive conditions: retrospective analysis of hospital episode statistics”, *BMJ*, 350, 2015, p. h1480.

[5] J Paul Leigh and Michael Schembri, “Instrumental variables technique: cigarette price provided better estimate of effects of smoking on SF-12”, *J Clin Epidemiol*, 57(3), 2004, pp. 284–93.

[6] M M Glymour, et al., “Does childhood schooling affect old age memory or mental status? Using state schooling laws as natural experiments”, *J Epidemiol Community Health*, 62(6), 2008, pp. 532–7.

[7] Paul J Gertler, et al., *Impact evaluation in practice*, World Bank Publications, 2011.

[8] Adriana Camacho and Emily Conover, “Manipulation of social program eligibility”, *American Economic Journal: Economic Policy*, 3(2), 2011, pp. 41–65.

[9] Leah M Smith, et al., “Effect of human papillomavirus (HPV) vaccination on clinical indicators of sexual behaviour among adolescent girls: the Ontario Grade 8 HPV Vaccine Cohort Study”, *CMAJ*, 187(2), 2015, pp. E74–81.

[10] Christopher Carpenter and Carlos Dobkin, “The Minimum Legal Drinking Age and Morbidity in the United States”, *Review of Economics and Statistics*, 99(1), 2017, pp. 95–104.

[11] M Maria Glymour, *Social epidemiology*, Oxford University Press, chap. Policies as tools for research and translation in social epidemiology, 2014, pp. 452–77.