Intervention studies in public health – 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

PhD course: Advanced Social Epidemiology, 19th Aug – 23rd Aug, 2019, University of Copenhagen

What question do most studies in social epidemiology answer?

- Do individuals who are disadvantaged with respect to social position have worse health than those who are advantaged?

Other kinds of questions that could be asked:

- Would individuals who are disadvantaged with respect to social position have better health if they were to become advantaged?
- Would individuals who are advantaged with respect to social position have worse health if they were to become disadvantaged?

These are **causal** questions.

"Normal" etiological science in social epidemiology

1. Follow-up of individuals in different social groups for various health outcomes (incidence, mortality, risk factors)

2. Adjustment for various confounders/mediators (are inequalities "explained" by....A, B, C?).

- "Our results demonstrate that"...we should:
  - *raise* education levels
  - *increase* economic assistance to the poor
  - *remove* noxious exposures from home environments/neighborhoods
  - *reduce* psychosocial workplace hazards
  - *eliminate* hierarchies, and the like.

- These statements are based on making **causal** inferences.

What's the problem?

- We are mainly (though not exclusively) interested in causal effects.

- We want to know:
  - Should we intervene to reduce exposure to X?, or
  - Did the program work? If so, for whom? If not, why not?, or
  - 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 treatment have better outcomes?

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

- **Association**: Do individuals who happen to be treated have better outcomes?

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

- **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).

- Recall that randomization means that we can generally estimate the causal effect without bias.
- Randomization guarantees exchangeability on measured and unmeasured factors.
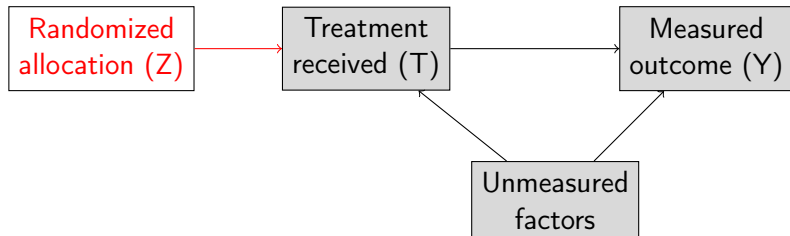
- Howden-Chapman et al. (2007)
- Cluster-RCT to retrofit houses with insulation.

- Randomization leads to:
  - balance on measured factors.
  - balance on unmeasured factors.

- Unmeasured factors cannot bias the estimate of the exposure effect.

Table 1
Baseline information for each group

	Control group	Intervention group
Household factors at baseline		
Number recruited	672	680
Number returned questionnaire	652	658
Dwelling reported in "poor" or "very poor" condition (%)	116/644 (18)	118/653 (18)
Condensation reported (%)	566/633 (89)	577/640 (90)
Non-condensation dampness reported (%)	413/613 (66)	437/641 (68)
Mould reported (%)	481/643 (75)	490/651 (75)
Dwelling cold "always" or "most of the time" (%)	452/647 (70)	473/651 (73)
Individual factors at baseline		
Number	2152	2261
Female (%)	1115/2152 (52)	1187/2261 (53)
Ethnicity		
NZE (%)	794/2121 (37)	835/2231 (37)
Maori (%)	1005/2121 (47)	1109/2231 (50)
Pacific (%)	578/2121 (27)	501/2231 (22)

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 ($Z \rightarrow Y$)

- RCTs have serious limitations.
  - Non-compliance.
  - Attrition.
  - Spillovers.
  - Blinding (esp. in clinical trials).

Problem of Social Exposures

- Many social exposures cannot be randomized by investigators:
    - Unethical (poverty, parental social class, job loss)
    - Impossible (ethnic background, place of birth)
    - Expensive (neighborhood environments)

- Some exposures are hypothesized to have long latency periods (many years before outcomes are observable).

- Effects may be produced by complex, intermediate pathways.

- We need alternatives to RCTs.

- We have lots of statistical associations between social exposures and health.

$$X - - - Y$$

- Some possible situations *consistent* with statistical associations:

1. Causal $X \to Y$
2. Heterogeneity $X_a \ Y_a$ vs. $X_b \to Y_b$
3. Reverse causation $Y \to X$
4. Confounding $X \leftarrow C \to Y$
5. Selection bias $X \to S \leftarrow Y$

Unmeasured confounding is a serious challenge

- We often compare socially advantaged and disadvantaged on health.

- 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...**adjust.**
  - Measure and adjust (regression) for C confounding factors.
  - Conditional on C, we are supposed to believe assignment is "as good as random" = causal.

- How credible is this assumption?

Ex: SEP and CVD in Australia

Many low p-values. Is "no other unmeasured differences" credible?

Table 1 Characteristics of 38 355 subjects in the Melbourne Collaborative Cohort Study at baseline (1990−1994)

|		Highest level of education attained				
		Completed tertiary* n=8588	Completed secondary† n=7882	Some secondary‡ n=14543	Primary only§ n=7342	p Value for trend¶
Male	n (%)	4025 (47%)	3776 (48%)	4680 (32%)	2780 (38%)	<0.001
Female	n (%)	4563 (53%)	4106 (52%)	9863 (68%)	4562 (62%)	<0.001
Age (years)	(Mean, SD)	51.6 (8.4)	54.5 (8.8)	55.7 (8.5)	57.8 (7.1)	<0.001
Country of birth, n (%)	Australia, New Zealand or northern Europe (n=28 835)	8263 (96%)	6814 (86%)	12696 (87%)	1062 (14%)	<0.001
	Southern Europe (n=9520)	325 (4%)	1068 (14%)	1847 (13%)	6280 (86%)	<0.001
Behavioural risk factors						
Current smoker	n (%)	574 (7%)	960 (12%)	1828 (13%)	947 (13%)	<0.001
Vegetable intake (times/day)	Mean (SD)	5.7 (3)	5.3 (3)	5.2 (3)	5.8 (4)	1.000
Fruit intake (times/day)	Mean (SD)	4.4 (3)	4.0 (3)	3.9 (3)	4.7 (4)	0.007
Saturated fat intake (g/day)	Mean (SD)	35.0 (15)	34.3 (16)	33.7 (16)	30.3 (18)	<0.001
Current drinker	n (%)	7061 (82%)	5883 (75%)	9397 (65%)	3666 (50%)	<0.001
Alcohol intake, current drinkers (g/day)	Median (IQR)	14 (5,26)	13 (4,26)	10 (3, 23)	15 (4,30)	<0.001
Physical activity (% inactive)	n (%)	1224 (14%)	1520 (19%)	3238 (22%)	2546 (35%)	<0.001
Social connection						
Living alone	n (%)	1514 (18%)	1274 (16%)	2250 (15%)	498 (7%)	<0.001
No social activity per week (%)	n (%)	733 (9%)	1069 (14%)	2479 (17%)	1889 (26%)	<0.001
Physiological risk factors						
Prior diabetes	n (%)	105 (1%)	183 (2%)	386 (3%)	598 (8%)	<0.001
Systolic blood pressure (mm Hg)	Mean (SD)	130.9 (17)	135.7 (19)	137.4 (19)	142.3 (20)	<0.001
Waist circumference (cm)	Mean (SD)	82.6 (12)	85.1 (13)	84.2 (13)	90.0 (12)	<0.001
Total cholesterol (mmol/l)	Mean (SD)	5.3 (1)	5.5 (1)	5.6 (1)	5.7 (1)	<0.001

Beauchamp et al. (2010)

Is observational 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 breastfed 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/

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 . . .
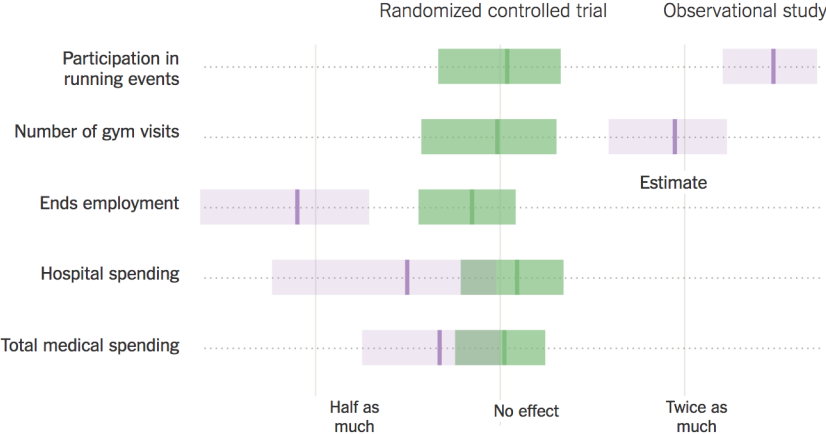
- Quasi-experiments aim to mimic RCTs.

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

- Control for (some) sources of bias that cannot be adequately controlled using regression adjustment.

- More credible designs also help us to understand the relevance of other factors that may be implicated in generating inequalities.

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].

- Most observational study designs control for *measured* factors using:
  - Stratification
  - Regression adjustment
  - Matching (propensity scores, etc.)

- 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)

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)
- Business / factory closures
- Historical legacies (physical environment)
- Seasonality

Difference-in-Differences

What's missing here?

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).
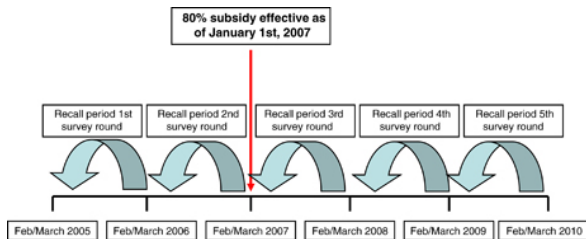
Difference-in-Differences: Basic Idea

- The simplest DD setting:
  - outcomes are observed for units observed in one of two groups
  - in one of two time periods.
- Treated: only units in one of the two groups are exposed to a treatment, in the second time period.
- Control: Never observed to be exposed to the treatment.

- The average change over time in the non-exposed (control) group is subtracted from the gain over time in the exposed (treatment) group.
- Double differencing removes biases in second period comparisons between the treatment and control group that could be the result from permanent differences between those groups, as well as biases from comparisons over time in the treatment group that could be the result of time trends unrelated to the treatment.

Visual Intuition of DD
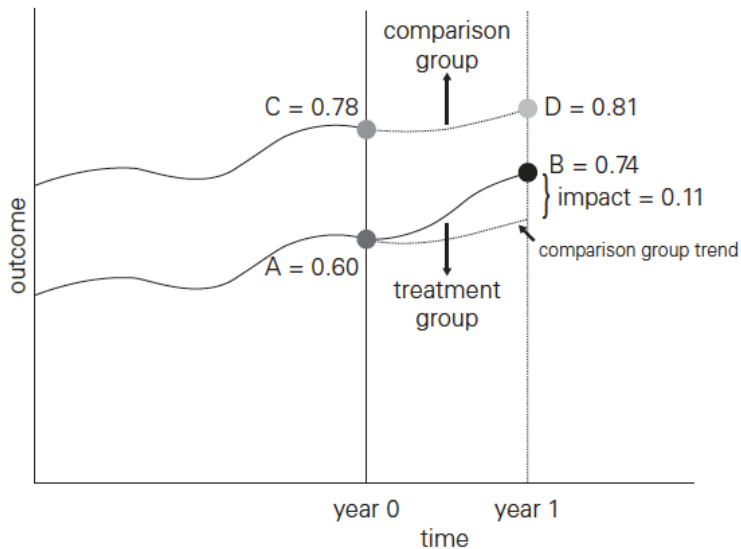
Difference-in-Differences without Regression

DD is just 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

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.

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.
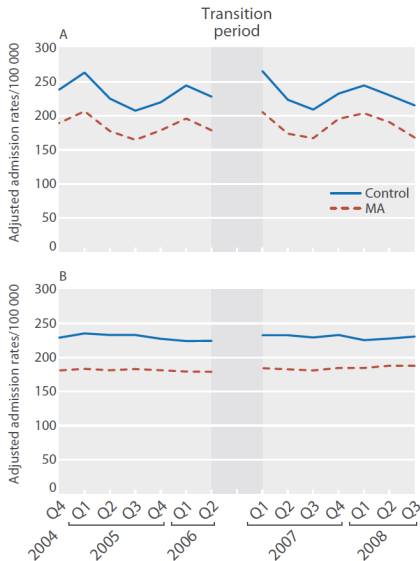
Instrumental variables

Challenge of conventional observation study (again)

- WHO: "Educational attainment is linked to improved health outcomes."
- But what about unmeasured confounding? Unmeasured factors such as personality traits, cognitive ability, etc. may be predictive of both eduction and disease.
- Failure to measure such factors will falsely attribute their effects to education.
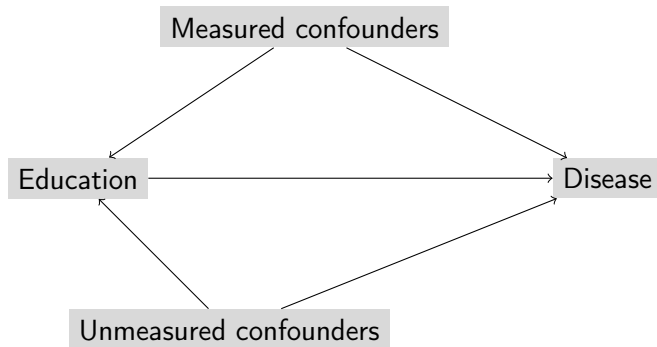
- "Instrumental variable": predicts education but not associated with anything else [mimicking random assignment].


Observational study

Quasi-experimental study
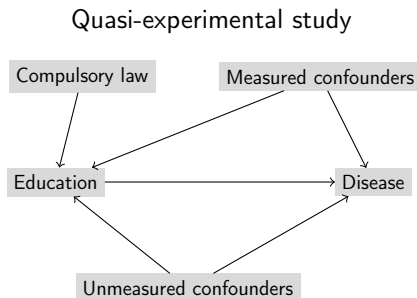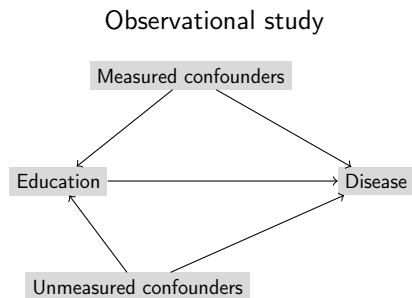
Non-randomized instrument creates additional issues

- In an 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.
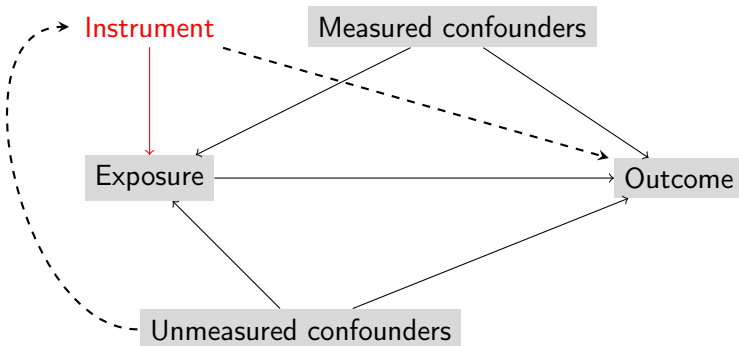
Non-randomized examples of IV: Policies

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

Leigh and Schembri 2004 [2]
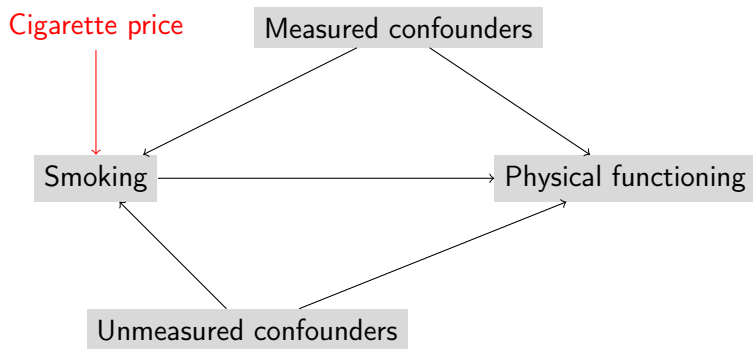
Non-randomized examples of IV: Policies

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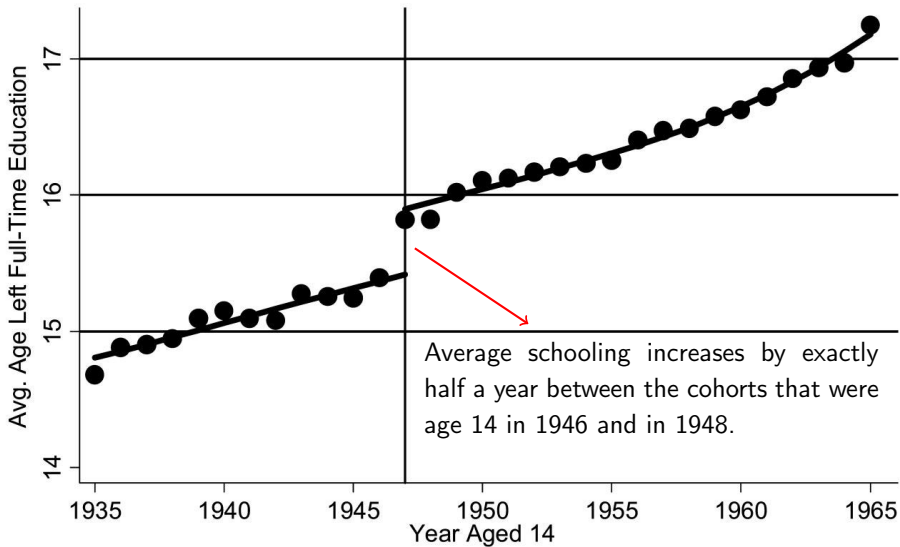

Glymour et al. 2008 [3]
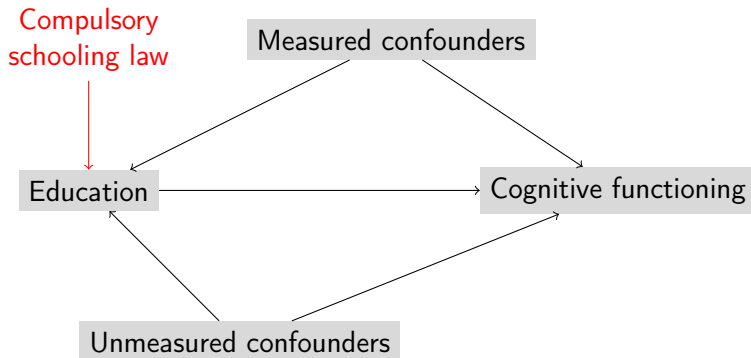
What does a quasi-experiment look like?

Fraction left full-time education by year aged 14 and 15 (Great Britain)
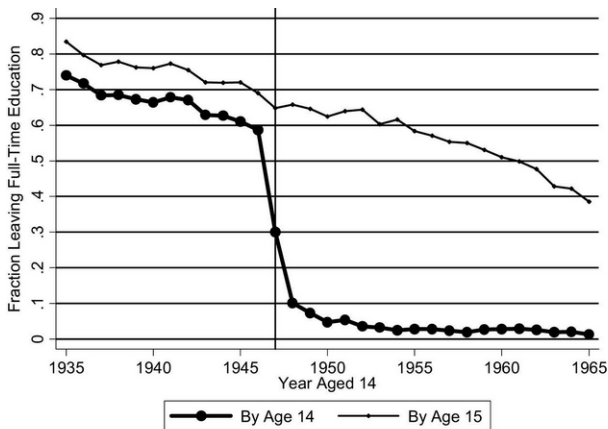
The lower line shows the proportion of British-born adults aged 32 to 64 from the 1983 to 1998 General Household Surveys who report leaving full-time education at or before age 14 from 1935 to 1965. The upper line shows the same, but for age 15. The minimum school-leaving age in Great Britain changed 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

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

Study	Effect Size (95% CI)
Braga (2013, female)	0.06 (-0.03, 0.16)
Braga (2013, male)	0.04 (-0.10, 0.18)
Etil (2011, female)	-0.03 (-0.19, 0.14)
Etil (2011, male)	-0.01 (-0.03, -0.00)
Grabner (2009a, female)	0.00 (-0.11, 0.11)
Grabner (2009a, male)	0.18 (-0.01, 0.38)
Huang (2015)	-0.01 (-0.03, -0.00)
James (2013)	-0.03 (-0.08, 0.03)
Kemptner (2011, female)	-0.00 (-0.02, 0.02)
Kemptner (2011, male)	-0.00 (-0.02, 0.01)
Li (2015)	-0.01 (-0.02, 0.01)
Schneeweis (2014, female)	-0.00 (-0.07, 0.06)
Schneeweis (2014, male)	0.06 (-0.02, 0.15)
Silles (2015b)	-0.01 (-0.04, 0.02)
Silles (2015b)	-0.02 (-0.07, 0.03)
Storger (2007)	-0.08 (-0.28, 0.12)
Xie (2014)	0.04 (-0.05, 0.13)
Overall	-0.01 (-0.02, -0.002)

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.

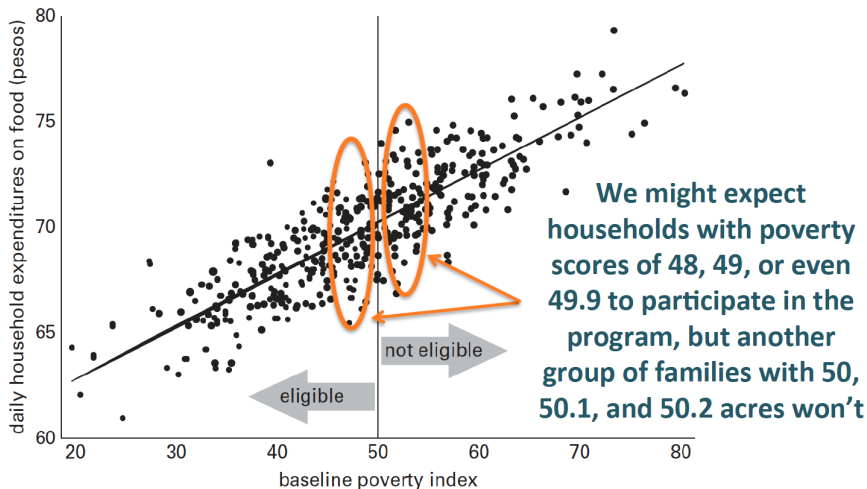- Groups just on either side are the threshold considered "as good as randomly" assigned to treatment.

RD: 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[4]

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

Source: Gertler, 2011[4]
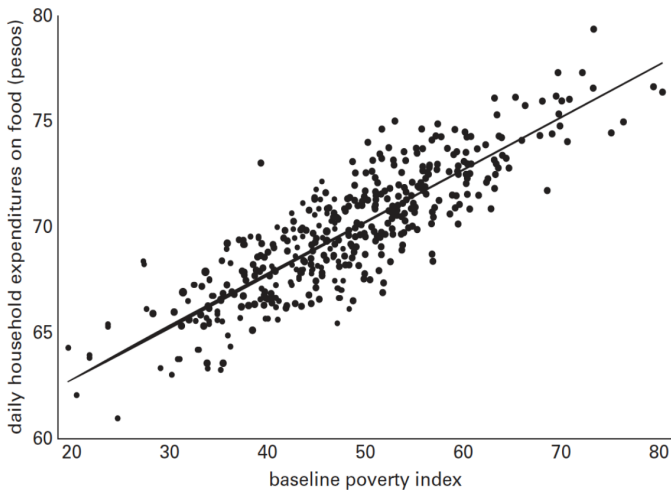
Under the program's rules, only households with a poverty score, Z, below 50 are eligible for the cash payment:

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[4]

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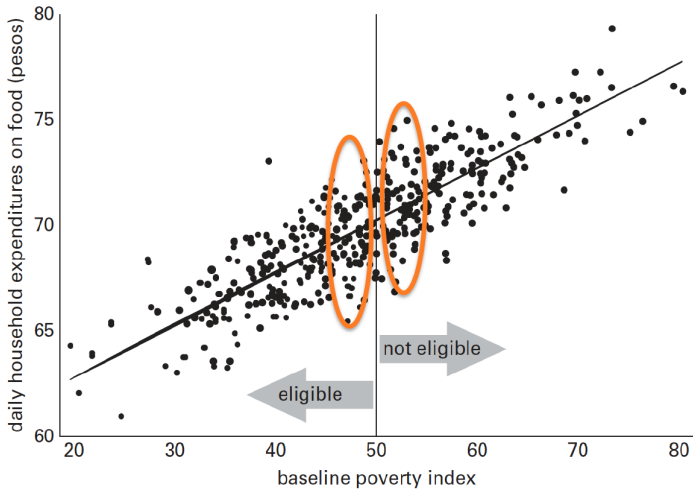
How about these families?


Source: Gertler, 2011[4]
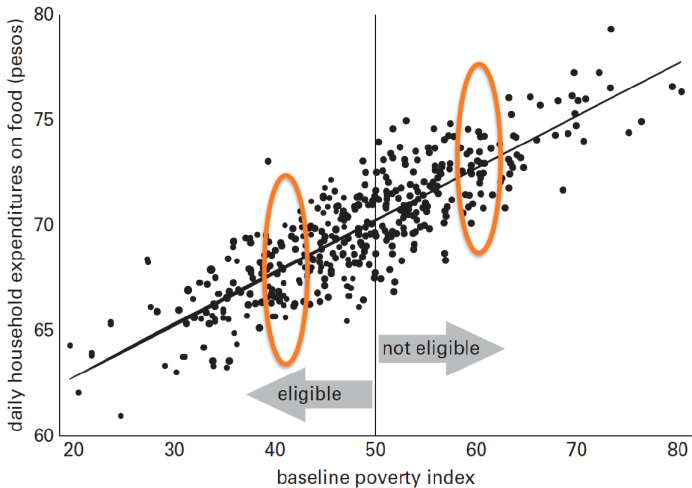
As we approach the cutoff value from above and below, the individuals in both groups become more and more alike, on both measured and unobserved characteristics—in a small area around the threshold, the only difference is in treatment assignment
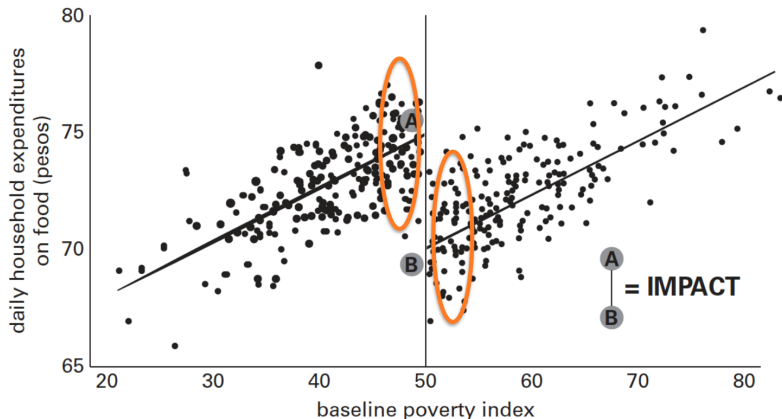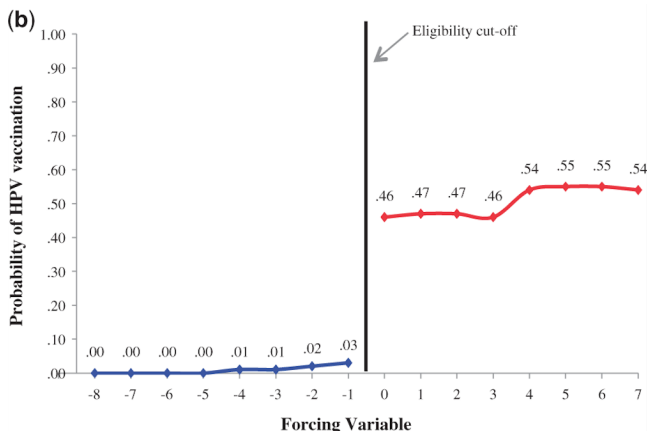
- Does getting the HPV vaccine affect sexual behaviors?

- Vaccine policy: predicts vaccine receipt but (**we assume**) not associated with anything else [mimicking random assignment].
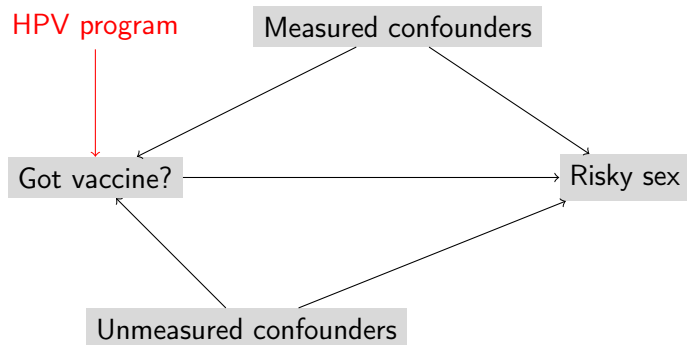
Does the cutoff predict treatment?

- 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.

(b)

Probability of HPV vaccination vs Forcing Variable

.46 .47 .47 .46 .54 .55 .55 .54

.00 .00 .00 .00 .01 .01 .02 .03

Eligibility cut-off

Smith et al., 2015[5]

What does a credible natural experiment look like?

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

| Characteristic | Program eligibility group; % of eligibility group* | | Characteristic | Program eligibility group; % of eligibility group* | |
	Ineligible (n = 131 781)	Eligible (n = 128 712)		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[5]

Note little impact of adjustment

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

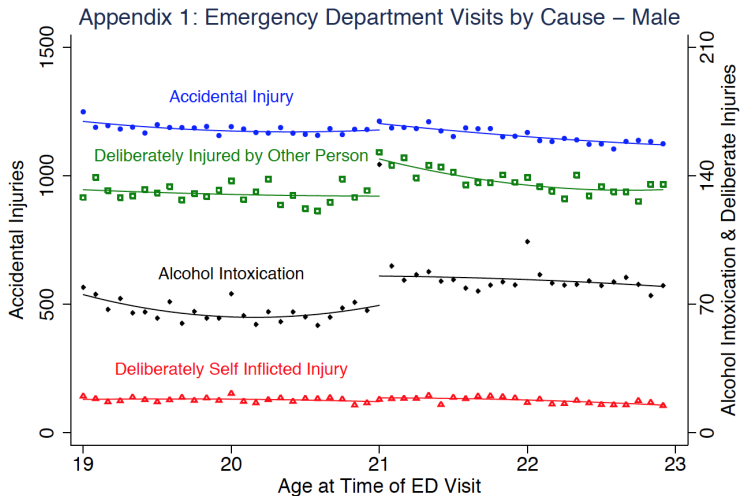| Outcome	No. of excess cases per 1000 girls (95% CI)	RR (95% CI)	Adjusted† RR (95% CI)
Effect of vaccine			
Composite outcome	–0.61 (–10.71 to 9.49)	0.96 (0.81 to 1.14)	0.98 (0.84 to 1.14)
Pregnancy	0.70 (–7.57 to 8.97)	0.99 (0.79 to 1.23)	1.00 (0.83 to 1.21)
STIs	–4.92 (–11.49 to 1.65)	0.81 (0.62 to 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[5]

Another recent example: US drinking age

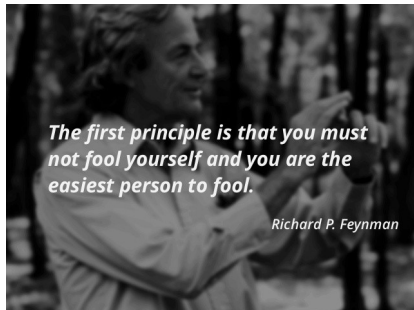- Minimum legal drinking age and non-fatal injuries:

Appendix 1: Emergency Department Visits by Cause – Male

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[6]

- 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 heterogeneity).

- 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

Be careful, and skeptical

- Correlations between social factors and health are easy to find.

- They do not necessarily reflect **causal** relationships.

- Need to search hard for alternative explanations.

- Important to consider the strength of evidence in considering interventions.

The first principle is that you must not fool yourself and you are the easiest person to fool.

Richard P. Feynman

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.
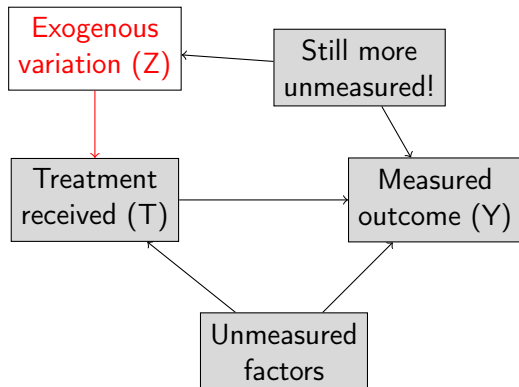
- How good is "as-if" random? (need "shoe-leather")

- Credibility of additional (modeling) assumptions.

- Relevance of the intervention.

- Relevance of population.

- 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.

References I

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.

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.

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.

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

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.

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.

Acknowledgements

- Canadian Institutes for Health Research

  - Salary award from Fonds de recherche du Québec – Santé

  - Public Policy and Population Health Observatory (www.3po.ca)

  - Smarter Choices for Better Health Initiative, Erasmus University

Thank you!

sam.harper@mcgill.ca   @sbh4th