Introduction to Impact Evaluation in Social Epidemiology and Public Health

Sam Harper Arijit Nandi

McGill University

Epi Tools Workshop Series, UCSF, 13 April 2017

Outline

This is an "Epi Tools" workshop

- What's in your toolbox?

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?

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

- **Confounding [Omitted Variable Bias]** :

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

Effect of what?

- Questions about effects are expressed as counterfactual contrasts.
- We can write the average causal effect (ACE) or average treatment effect (ATE) in terms of potential outcomes:

$$E\left(Y^{a=1}\right) - E\left(Y^{a=0}\right)$$

  - indicating potential outcomes for an individual under two possible treatments.
- Consistency assumption: $Y^a = a$ for every individual with $A = a$.
- Problematic when there are multiple versions of the treatment, or when we do not have control over treatment assignment.
  - We need well-defined treatments.

Hernan 2016[1], among others.

Here's a question: Does obesity shorten life?

- The increased emphasis on causal inference in epidemiology has generated some acrimonious debate:*
- Vandenbroucke, Pearce, Broadbent, Krieger, Davey Smith, Ebrahim, Schwartz, etc...
  - "You members of the potential outcomes methodological lynch mob are running a totalitarian police state. There is no room left for creative thinking, radical ideas, or pluralism when it comes to causal inference."
- Hernan, VanderWeele, Robins, Kaufman, Daniel, De Stavola, etc.
  - "we don't claim that counterfactul causal models subsume all of causal inference, but if you want to talk about causal effects, you need to specify well-defined hypothetical interventions."

*To paraphrase Sayre: Academic politics are so vicious precisely because the stakes are so low.

Debates about causality in **social** epidemiology

- Recent debate: Are race and sex causes of disease?
- Kaufman: "The causal inference field has given precise definitions to terms like 'effect' and 'confounding', and at least since Robins and Greenland's influential article on the topic, these fundamental epidemiologic quantities have been expressed in terms of hypothetical interventions."[5]
- Schwartz: "forcing investigators to carefully define real-world interventions, the potential outcomes approach takes the attention away from philosophical debates about the definition of cause toward the consideration of potentially implementable policies that can alleviate poverty, reduce discrimination, and increase economic opportunities for disadvantaged populations."

See Kaufman, 1999[2]; Vanderweele, 2014[3]; Hernan, 2016[1] Glymour, 2017[4]

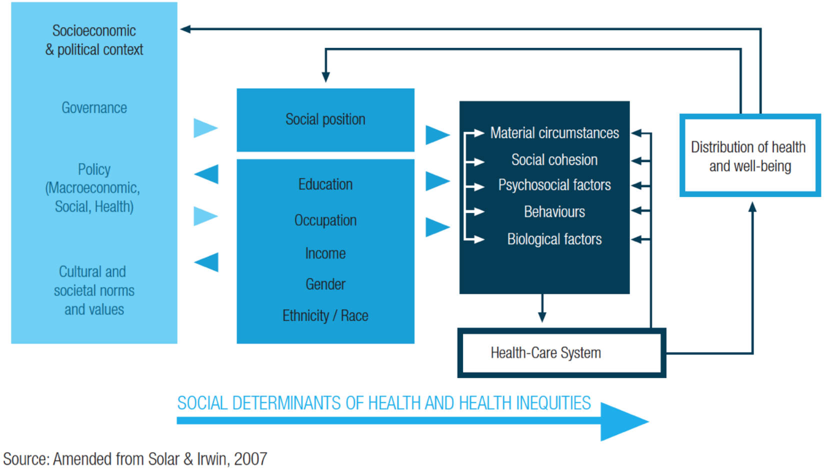
Relevance for **social** epidemiology

- Recent debate: Are race and sex causes of disease?
- Glymour: it's difficult, but possible, to think of them as causes, and good for social epidemiologists to struggle with it.
- Hernan: interventions for race/sex are ill-defined, and so it is not useful to consider estimating their causal effects.
  - "[a] sufficiently well-defined intervention needs to specify the start and end of the intervention and the implementation of its different components over time."
- Everyone: if we want to reduce health inequalities, it is pragmatic and useful to identify well-defined interventions that may lead to meaningful changes in differentially distributed risks **between** social groups.

See Kaufman, 1999[2]; Vanderweele, 2014[3]; Hernan, 2016[1] Glymour, 2017[4]

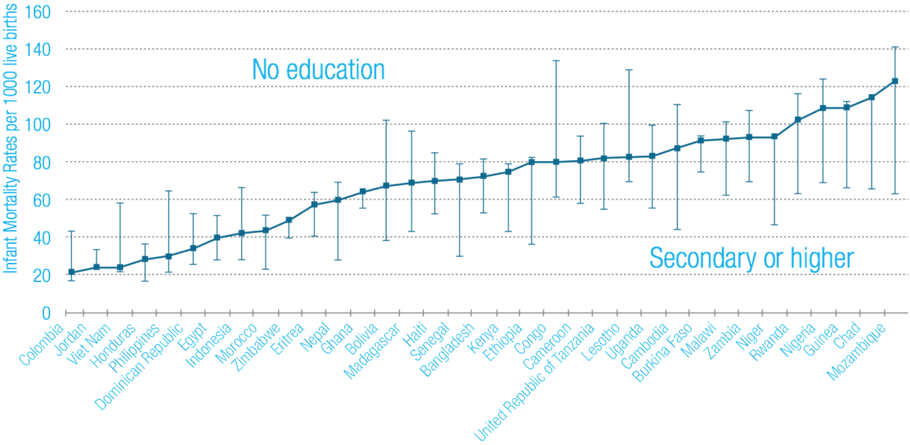
But which interventions are the **right** ones?

Social determinants of health



WHO CSDH, 2008[6]

Motivating idea: educational inequalities in infant mortality



WHO CSDH, 2008[6]

WHO Committee on Social Determinants of Health

Three principles of action

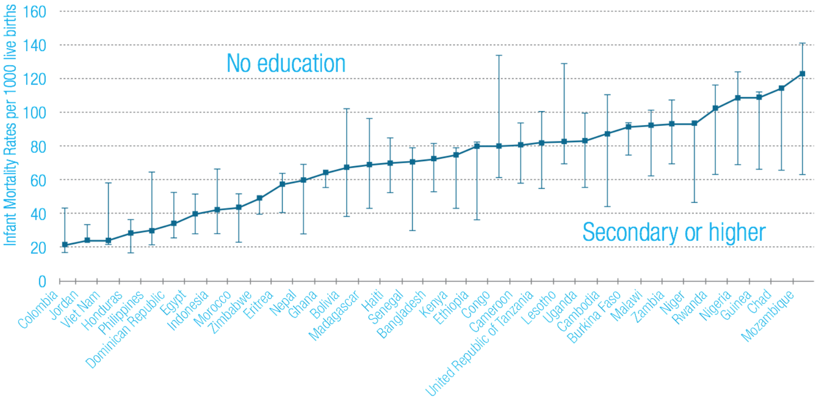
1. Improve the conditions of daily life – the circumstances in which people are born, grow, live, work, and age.
2. Tackle the inequitable distribution of power, money, and resources – the structural drivers of those conditions of daily life – globally, nationally, and locally.
3. Measure the problem, evaluate action, expand the knowledge base, develop a workforce that is trained in the social determinants of health, and raise public awareness about the social determinants of health.

- How should we design interventions based on these principles?

WHO CSDH, 2008[6]

A thought experiment

Let's assume that the education-based gradients in infant mortality, perhaps counter to fact, reflect a causal effect, and you were charged with eliminating these inequalities. . . what would you do?



WHO CSDH, 2008[6]

A thought experiment

How should we intervene to reduce educational-based inequalities in infant mortality? Should we:

- Increase secondary or higher education by making it free?
- Increase secondary education by making it compulsory?
- Increase secondary education by increasing school quality?
- Build more secondary schools?

- Increase access to maternal care among less-educated women?
- Increase immunization among kids of less-educated mothers?
- Increase access to family planning?
- Increase access to household resources among less-educated mothers?

- All of the above?
- Some of the above
- None of the above?

Relevance for social epidemiology

- Why this matters for social epidemiology:
- "The branch of epidemiology that studies the social distribution and social determinants of health states"
    - Social factors as exposures.
    - Social groups as populations.
    - Interventions on social determinants or consequences of interventions for different social groups.
- How much do we know about interventions in social epidemiology?

Berkman et al. 2016[7]

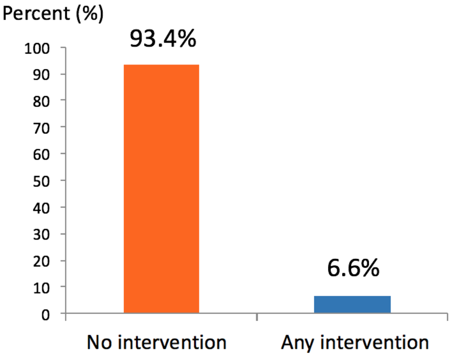
Are social epidemiologists focused on interventions?

- Certainly lots of talk, but not an easy question to answer.
- One perspective: we identified all "social epidemiology" abstracts from 2009-2013 SER meetings (n=619).
- Assessed whether study evaluated or simulated the impact of a specific (not necessarily well-defined) intervention, e.g.,
  - "Yes" examples: Earned-income tax credit; smoke-free legislation; conditional-cash transfer
  - "No" examples: Neighborhood SES; mediation of inequalities by risk factors

Nandi and Harper, 2015[8]

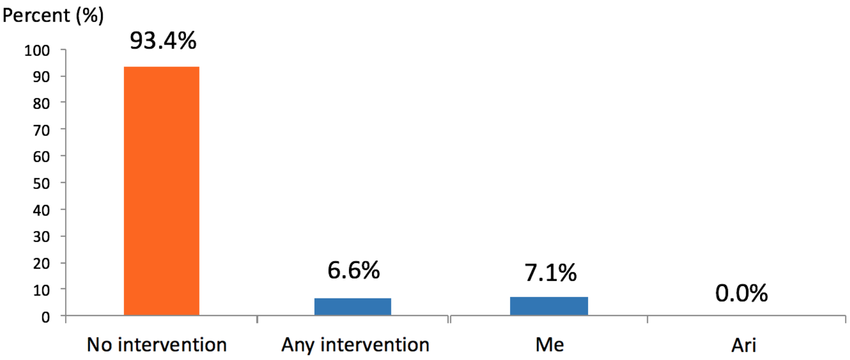
How are we doing?

- 41/619 social epidemiology studies (6.6%) evaluated or simulated the impact of a specific intervention.



Nandi and Harper, 2015[8]

- Our own work was no better. Sigh.



Nandi and Harper, 2015[8]

Policymakers' Context for Health Inequalities

- Interviews with UK health policymakers in the early 2000s were disappointing for those wanting their research to have "impact".
- The "inverse evidence law" (Petticrew 2004[9]): "...relatively little [evidence] about some of the wider social economic and environmental determinants of health, so that with respect to health inequalities we too often have the right answers to the wrong questions."
- Problem of "policy-free evidence": an abundance of research that does not answer clear, or policy relevant questions.

Policymakers' Context for Health Inequalities

- Policymakers desire for research on plausible causal pathways, but...
- Much of the available evidence on health inequalities concerned with health behaviors and clinical issues (mediation), rather than broader social determinants of health (i.e., the "total" effects)
- Research in social epidemiology is often explanatory rather than evaluative (i.e., obsessed with "explaining" away gradients and "independent" effects that do not correspond to any kind of intervention)

Petticrew, 2004[9]

How to make social epidemiology relevant to policy?

..."researchers may improve the likelihood of their research having a wider policy impact by focusing less on describing the problem and more on ways to solve it, working closely with those who are charged with the task of tackling health inequalities, and others who can contribute to the creation of a climate in which reducing health inequalities is perceived to be not only politically possible but publicly desirable."

Bambra et al., 2011 "A labour of Sisyphus?"[10]

How can impact evaluation help?

What is impact evaluation?

"Evaluations are periodic, objective assessments of a planned, ongoing, or completed project, program, or policy [that] can address three types of questions"(Gertler, 2016 [11]):

1. Descriptive questions, which seek "to determine what is taking place and describes processes, conditions, organizational relationships, and stakeholder views"
2. Normative questions, which compare "what is taking place to what should be taking place. . . .normative questions can apply to inputs, activities, and outputs"
3. **Cause-and-effect questions**, which examine outcomes and try to assess what difference the intervention makes in outcomes

What is impact evaluation?

- An impact evaluation "assesses the changes in the well-being of individuals that can be attributed to a particular project, program, or policy"(Gertler, 2016 [11]).
- "Impact evaluation asks about the difference between what happened with the program and what would have happened without it (referred to as the counterfactual)."(Savedoff, 2006[12]).
- The "impact" can be defined as the change in the outcome that can be causally attributed to the program (Ravallion, 2008 [13])
- Impact evaluation studies are among a range of complementary techniques for supporting evidence-based policymaking.

What is impact evaluation?

- By definition, impact evaluations focus on estimating impact: changes in outcomes directly attributable to the intervention.
- These are obviously based on counterfactual contrasts of (hopefully) well-defined interventions or programs.
- Specific type of evaluation design largely depends on the program under consideration.
- Impact evaluations show how programs work and, as evidence builds incrementally, can generate synthesized evidence on how effective a particular intervention is at changing an outcome.

Answering specific questions **and** testing theories

- Impact evaluations can serve a number of purposes, notably:
- Impact evaluations are public goods.
- Answer important pragmatic questions for implementing organizations:
  - Will more families use insecticide-treated bednets if the price declines by 50 cents?
- Contribute evidence to broader questions about mechanisms:
  - Do small costs prevent people from taking up beneficial interventions?
  - Similar results in different settings inform us about generalizability.

Converging evidence is more convincing

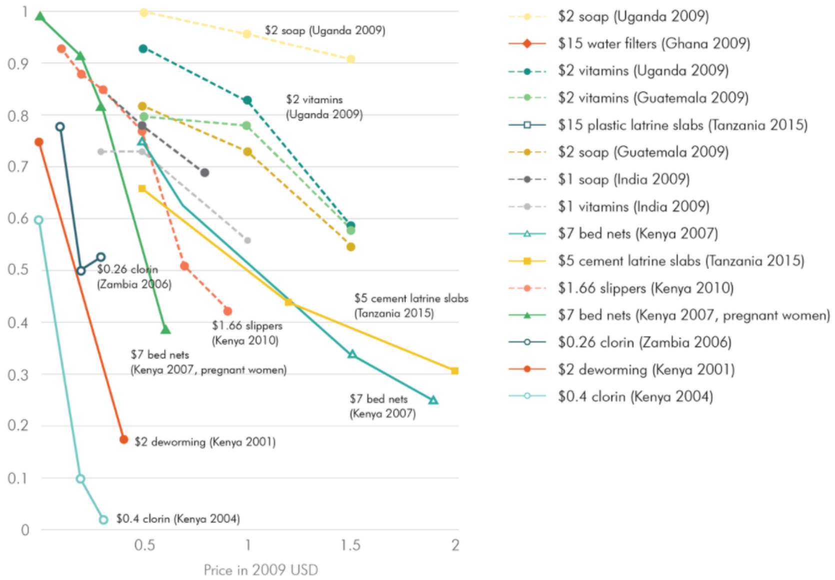
- Specific case of the costs of preventive health products (bednets, vitamins, soap).
- Charging small user fees to the poor for health products and services has been promoted for theoretical reasons.

1. Greater efficiency by allocating resources to those who value it most.
2. Increasing accountability for social programs.
3. Fairness by only charging those who use goods/services.

- A number of impact evaluations (randomized) tested these ideas for preventive health products.

http://runningres.com/blog/2016/5/27/not-so-small

Demand for Preventive Healthcare Products Based on Price



http://runningres.com/blog/2016/5/27/not-so-small

Potential for large impact with scale-up

"Coverage of ITNs in sub-Saharan Africa (the region with the highest burden of malaria) has improved dramatically with the vast majority of coverage accounted for by free mass distribution (43 out of 47 countries had mass free programs). As the great maps from Giving What We Can illustrate, malaria cases have fallen dramatically. A recent article in Nature estimates that 2/3 to 3/4 of the decline in malaria cases between 2000 and 2015 can be attributed to increased net coverage: 450 million cases of malaria and 4 million deaths averted from ITN distribution."

Glennerster, 2016

Prospective and retrospective evaluations

- Impact evaluations are always planned, but can be conducted prospectively or retrospectively.
- Prospective:
  - Developed alongside the intervention.
  - Shared definitions of intended outcomes and results.
  - Baseline data usually collected (pre-intervention), quality control.
  - Clear definitions of treatment and comparison groups.
  - Flexibility of designs (randomized or non-randomized), tailored to provide best counterfactual.
- Retrospective
  - Likely fewer resources needed.
  - Data availability can be problematic.
  - Typically require stronger assumptions.

Natural complementary activities

- Monitoring:
  - Verification of intervention activities, treatment integrity.
- Simulated impacts:
  - Expected impacts in another population, future impacts, scaled effects.
- Qualitative methods (focus groups, life histories):
  - Pre-intervention: informing hypotheses, tailoring aspects of the intervention or questions.
  - Post-intervention: insights into mechanisms, unexpected results.
- Process evaluations:
  - Program fidelity, operations, implementation details.

Potential impact on policy decisions

- Impact evaluations often estimate the both the costs and benefits of programs.
- Cost-benefit and cost-effectiveness analysis:
  - CBA: compares costs and benefits of a program.
  - CEA: relative costs of different programs for the same outcome.
- The comparative effectiveness (and cost-effectiveness) of various policy levers for affecting the same outcome can be compared and used to promote evidence based decision-making.

INCENTIVES FOR IMMUNIZATION

Small incentives for parents, coupled with reliable services at convenient mobile clinics, increased full immunization rates sixfold. This approach was twice as cost-effective as improving service reliability without incentives.

FIGURE 2: PERCENTAGE OF CHILDREN AGED 1-3 YEARS FULLY IMMUNIZED BY TREATMENT STATUS

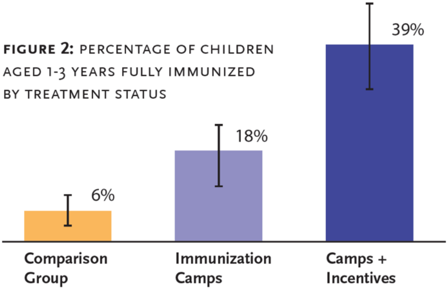
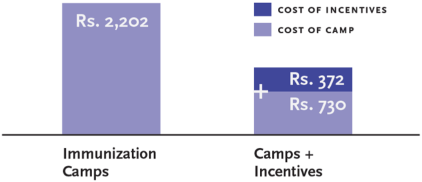
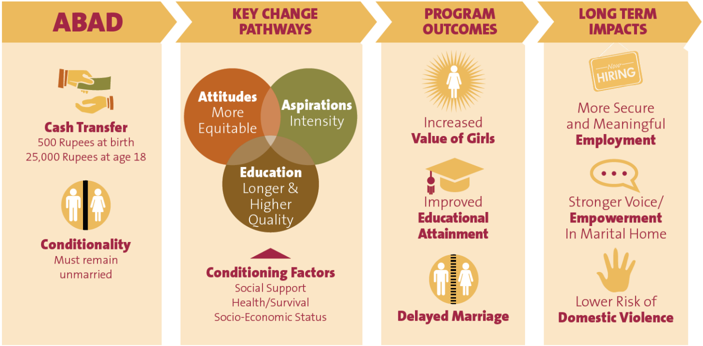


FIGURE 3: COSTS PER FULLY IMMUNIZED CHILD



Impact evaluation also shows us aspects of interventions that **don't** work—we have an imperfect understanding of how the local context works (our DAGs are wrong!) and well-intentioned interventions may have perverse effects.

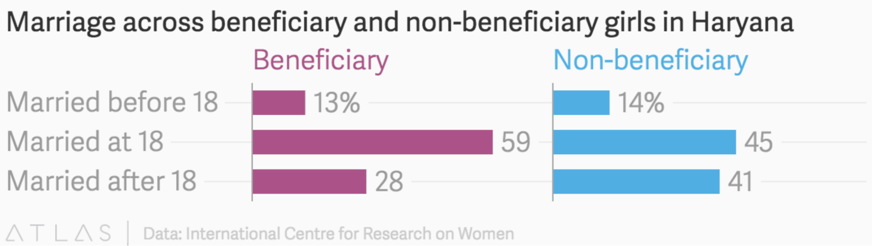
The Apni Beti Apna Dhan Program



- Implemented in 1994-1998, girls began turning 18 in 2012.
- Evaluation compared more than 10,000 eligible beneficiaries and non-beneficiaries from 300 villages between 2012 and 2015.

Wrong model? Wrong incentives? Wrong DAG?

- Overall, the survey recorded, fewer girls are marrying before 18 years of age in Haryana, regardless of whether they are supported by the government's cash scheme. But, among beneficiaries, the scheme "may have actually encouraged marriages at 18, and that parents who desired to have their daughters married early did so immediately upon receiving the cash benefit"



A scheme to end child marriage in patriarchal Haryana has totally backfired — Quartz: http://bit.ly/2o3D7wj.

Seems great! Why don't I know more about this?

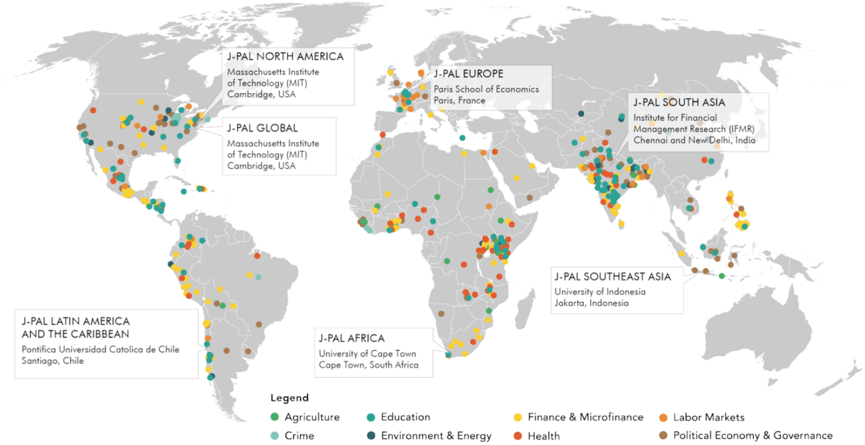
- Potential benefits of impact evaluations are high, evaluations are often feasible, but still relatively rare. Why?
- Evaluations can be expensive and the costs are immediate, but the benefits are not.
- Governments, donors, and other funders may have different priorities and rarely demand impact evaluations.
- Ignorance is bliss?

“Other incentives exist at the institutional level to discourage impact evaluations. . . Since impact evaluations can go any way—demonstrating positive, zero, or negative impact–a government or organization that conducts such research runs the risk of findings that undercut its ability to raise funds.”

“Policymakers and managers also have more discretion to pick and choose strategic directions when less is known about what does or does not work. This can even lead organizations to pressure researchers to soften or modify unfavorable studies or simply to suppress the results–despite the fact that knowledge of what does not work is as useful as knowledge of what does.”(CGD, 2006)

Center for Global Development, 2006[12]

Impact evaluations are becoming more widespread



Source: J-PAL

When to evaluate?

- We might opt not to evaluate interventions (1) for which the stakes are not high and the policy relevance is dubious or, conversely, (2) those for which the effectiveness is established.
- Smaller scale pilot studies or process evaluations rather than a full impact evaluation may be more practical for interventions lacking any evaluation, descriptive or causal.
- Better for programs that are: innovative, replicable, strategically relevant, untested, potentially influential.
- Ethical considerations:
  - It is unethical to withhold programs known to be beneficial.
  - Conversely, "the implicit corollary—that programs of unknown impact should not be widely replicated without proper evaluation—is frequently dismissed"(Savedoff, 2006[12])

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.

Randomized Trials vs. Observational Studies

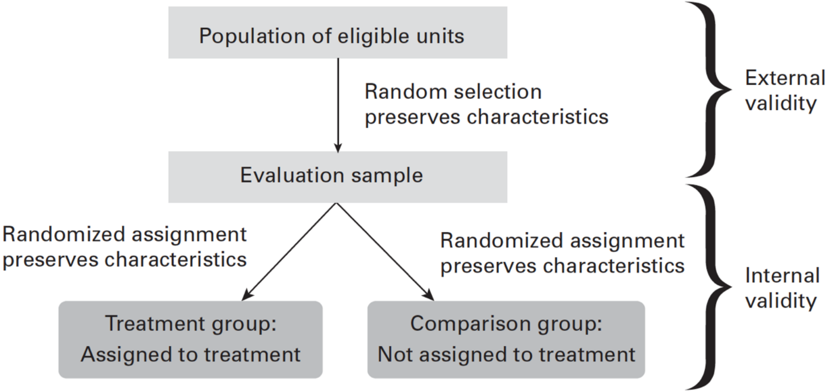
RCTs, Defined

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).

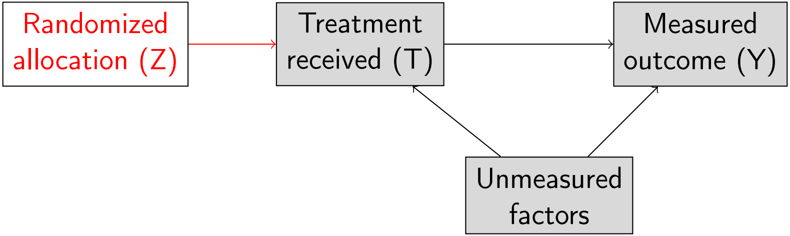
Random selection and assignment



Gertler, 2016[11]

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). The benefits are large.

- Randomization leads to: 1) balance on measured factors, and 2) balance on **unmeasured** factors.
- Unmeasured factors cannot bias the estimate of the exposure effect.
- Example: Randomized daycare program in Rajasthan

Table 1 Baseline characteristics for the total sample and stratified by treatment arm, presented at the individual ($n = 3177$) and cluster ($n = 160$) levels

	Individual-level analysis								
	Total sample			Control hamlets			Treated hamlets		
Variable[a]	No.	Mean	SD	No.	Mean	SD	No.	Mean	SD
Age (years)	3169	29.87	6.86	1517	29.87	6.90	1652	29.86	6.83
Any schooling	3175	0.26	0.44	1519	0.27	0.44	1656	0.25	0.43
Married	3177	0.98	0.12	1521	0.99	0.11	1656	0.98	0.14
Age married (years)	3070	17.47	2.85	1467	17.49	2.70	1603	17.44	2.98
No. sons	3177	1.62	1.16	1521	1.62	1.16	1656	1.61	1.16
No. daughters	3177	1.65	1.27	1521	1.60	1.25	1656	1.69	1.28
Hindu religion	3175	0.72	0.45	1519	0.73	0.44	1656	0.72	0.45
Worked in past 7 d	3177	0.59	0.49	1521	0.59	0.49	1656	0.59	0.49
Worked in past 12 mo	3177	0.95	0.22	1521	0.93	0.25	1656	0.96	0.18
Paid cash for work	3016	0.09	0.28	1420	0.09	0.29	1596	0.09	0.28
Days childcare prevents work	3016	1.53	4.66	1420	1.64	4.97	1596	1.42	4.36
Below poverty line	3168	0.50	0.50	1517	0.51	0.50	1651	0.50	0.50

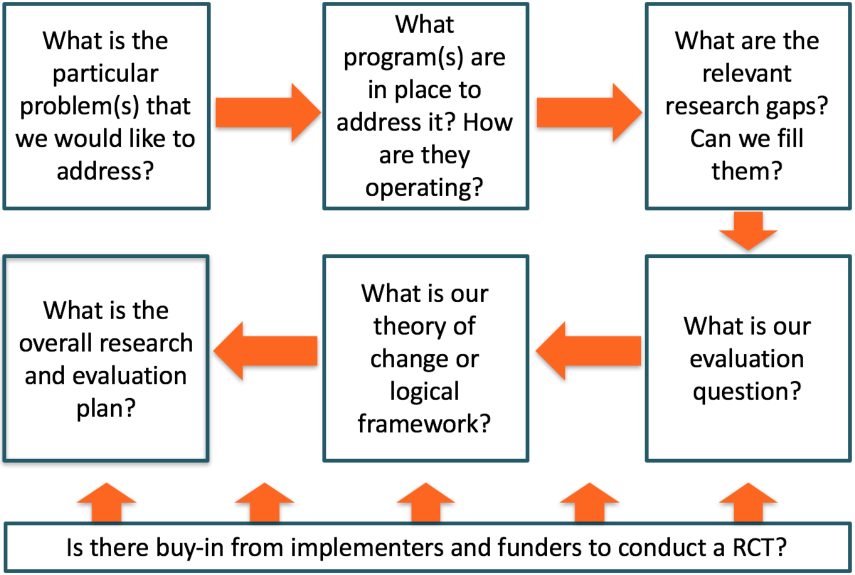
Nandi et al. 2016[14]

Pre-randomization phase

RCTs should facilitate creative collaboration

- By allowing us to vary individual elements of the treatment we can work with implementers to design our own interventions and answer questions that could not be answered in any other way.
- Comparing different variations of a program (e.g., how to deliver vaccines better?).
- Examination of process indicators can be used to understand mechanisms underpinning results.
- Feedback to implementers.
- Subgroup analyses can identify which groups are most affected.
- We can examine cost-effectiveness and increase policy-relevance.

Precursors to an evaluation plan



Pre-evaluation planning can reduce research waste

- "In 2009, Chalmers and Glasziou identified some key sources of avoidable waste in biomedical research. They estimated that the cumulative effect was that about 85% of research investment—equating to $200 billion of the investment in 2010—is wasted."

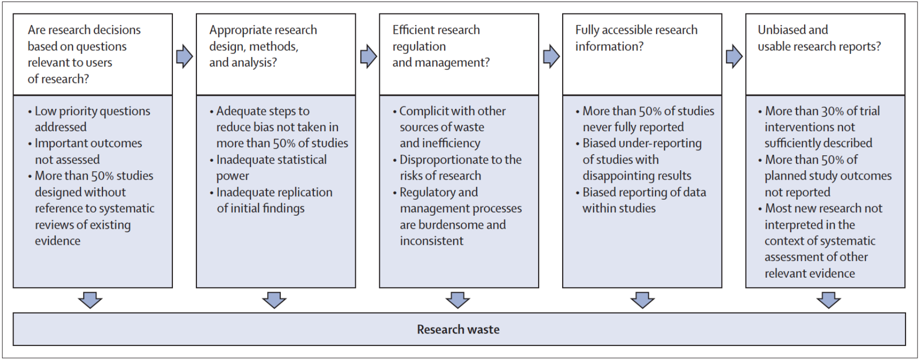


Figure: **Avoidable waste or inefficiency in biomedical research**

Macleod et al. 2014 [15]

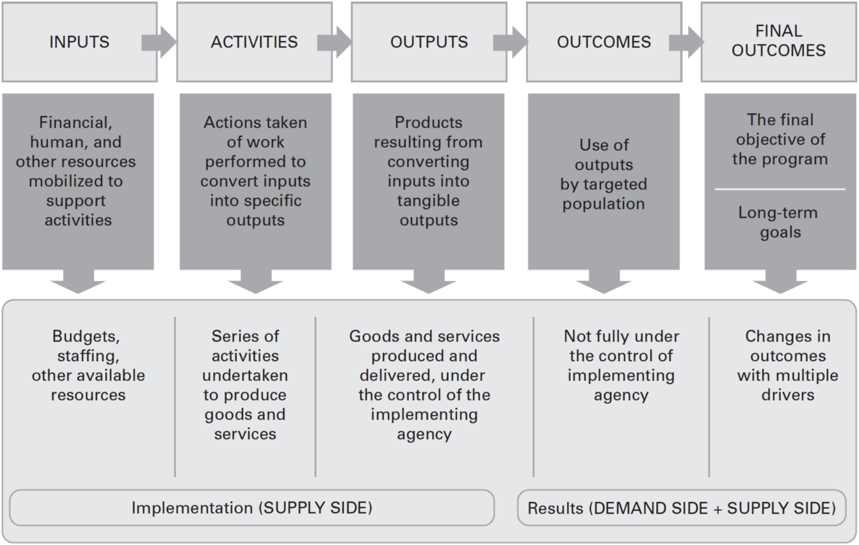
Needs assessment

- A needs assessment can be used to understand which groups in should be targeted, what problems they face, the reasons for these issues, which programs are already in place to address them, and what challenges remain unaddressed.
- A combination of qualitative and quantitative methods are useful for understanding the context of the intervention:
  - Qualitative methods, including focus group discussions, can help understand problems and opportunities and clarify why existing programs are inadequate (e.g., why take-up rates are low).
  - Representative surveys, including existing ones, can also be used.
- The needs assessment is critical for defining a **theory of change** and informing the design of the impact evaluation study.

Theory of change or "logical framework"

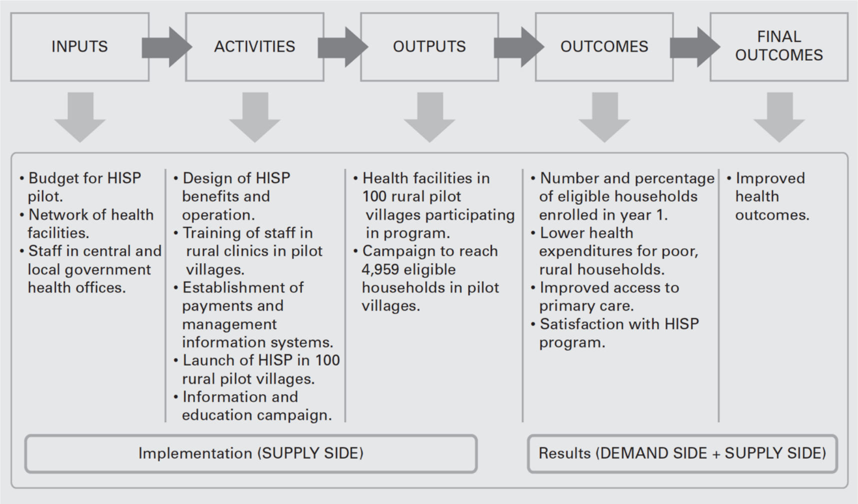
- A "structured approach used in the design and evaluation of social programs"(Glennerster, 2014[16]) that describes "how an intervention is supposed to deliver the desired results"(Gertler, 2016[11]).
- Like a directed acyclic graph (DAG), the theory of change describes the causal sequence of events leading to outcomes, as well as the conditions and assumptions needed for the change to occur.
- At the design stage, it can be developed iteratively with stakeholders to foster a shared vision for how the intervention should be designed.
- It also helps to specify the specific outcomes and indicators we will measure to help us understand whether the program has been implemented (monitoring) and how it has worked (impact evaluation).

Logical framework



Gertler, 2016[11]

Hypothetical example: Health Insurance Subsidy Program



Gertler, 2016[11]

Research plan and study protocol

- Research plan should be described a priori.
- The protocol should provide the intuition for conducting the evaluation and a detailed account of the overall study design, sample selection, survey procedures, randomization, intervention/program, measures, and ethical considerations.
- Study protocols should be registered prior to randomization; outlets include 3ie's registry, the ISRCTN clinical trial registry, and the American Economic Association' s registry for RCTs
- Detailed statistical plan should be published, including measures, hypotheses, detectable effects/power, statistical models, missing data plans, heterogeneity, etc...

Randomization phase

What can we randomize?

- We can control aspects of programs/policies to experimentally increase the probability of exposure in one group vs. another:
  - **Access**: we can randomly select which people are offered access to a program (most common).
  - **Timing**: we can randomly select when people are offered access to a program.
  - **Encouragement**: we can randomly select which people are given encouragement or incentive to participate.
- Each of these aspects can be varied for individuals or groups.

When can we randomize?

OPPORTUNITY	DESCRIPTION	EXAMPLE
Program design	We can work with implementers to design a program when a problem has been identified, but we lack consensus about the design of the solution	NGO wants to tackle teacher absenteeism but not sure what the program should look like
New service	When an existing program offers a new service	A microfinance bank begins to offer savings accounts in addition to credit services
New people and/or new location	When a program is being expanded to a new group of people	Expanding a remedial education program to a new city
Oversubscription	When there are more interested people than the program can serve	More families sign up for private school vouchers than the government can afford

Glennerster, 2013[16]

When can we randomize?

OPPORTUNITY	DESCRIPTION	EXAMPLE
Undersubscription	When not everyone who is eligible for the program takes it up; or to increase uptake and shared burden of unpopular programs	Vietnam War draft lottery
Rotation	Inadequate (fixed) resources are available so a program is randomly cycled through the population	Gender quotas for village councils in India
Admission cutoff	When a program has a merit cutoff and those just below the cutoff can be randomly admitted	Medicaid expansion in Oregon to those just above the usual cutoff
Admission in phases	When logistical or resource constraints prevent us from enrolling all potential beneficiaries at once	A program is building 200 new secondary schools, but can only build 25 per year

Glennerster, 2013[16]

Randomization at what level?

- Usually randomization is at the level at which the program is implemented—it is hard to prevent access to a program that is implemented at the community level.
- Therefore, community health programs are usually randomized by community, school improvement program are usually randomized by school, and so on...
- Randomizing below this level means changing how the program is implemented; for example, a nutritional program normally done by school but randomized by individual has to keep track of which children were treated, adding to cost of program

Methods of introducing randomization

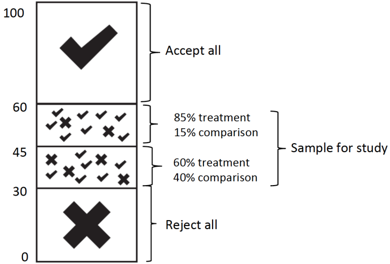
1. **Lottery**: using a simple lottery (e.g., flip of a coin) to decide which units receive the treatment among those eligible.
2. **Phase-in**: randomly select which units receive the intervention first and which receive it later.
3. **Rotation**: randomly assigns treatment to sub-groups in each area and then rotates the treatment assignments.
4. **Encouragement**: when eligibility is universal, a randomly selected group can be encouraged to take up the program.

Simple lottery example

- Oregon Health Plan (OHP) standard is a plan that serves low-income adults NOT eligible for Medicaid.
- At its peak in 2002, 110,000 people were enrolled, but by 2004 it was closed to new enrollment due to budgetary shortfalls.
- By 2008, attrition had reduced enrollment to about 19,000 and the state determined it could enroll another 10,000 adults.
- The state conducted a lottery since they anticipated that demand for the program would far exceed supply.
- There was an extensive public campaign, followed by a 5 week window during which >89,000 signed up.
- Nearly 30,000 households were selected by the lottery, and about 30% of these households were successfully enrolled.

Lottery around a cutoff

- Karlan et al.[17] worked with First Macro Bank, a for-profit bank in the Philippines, that provided small, 3-month loans at 60% annualized interest rates to micro-entrepreneurs outside Manila.
- Credit-scoring software to identify 1601 marginally creditworthy applicants loans. Two treatment groups:
  - Those with credit scores between 31-45 and 46-59; low scores (0-30) were rejected and high scores (60-100) were approved automatically.
  - 1272 of the 1601 participants with scores between 31-59 were randomly selected for a loan; the rest (329) were controls.



Phase-in design (aka "stepped wedge")

- Cross-over design in which the intervention is rolled out in a staggered manner such that, as time goes on, more and more units experience the intervention and eventually **all are treated**.
- There is an initial baseline period when units do not experience the intervention, after which the intervention is introduced; the time when each unit receives treatment is randomly assigned.
- Units are typically facilities, schools, villages, or communities; thus, this design is a special case of a cluster RCT.
- The staggered roll-out allows for time-specific between-cluster comparisons, as well as within-cluster comparisons across time.

Phase-in example

- In Baja California Sur, Mexico, most households rely on wells or springs to collect water, which is often stored in open containers
- This study evaluated the impact of UV-based household water treatment and safe storage intervention on water contamination and diarrhea in 444 households from 24 communities

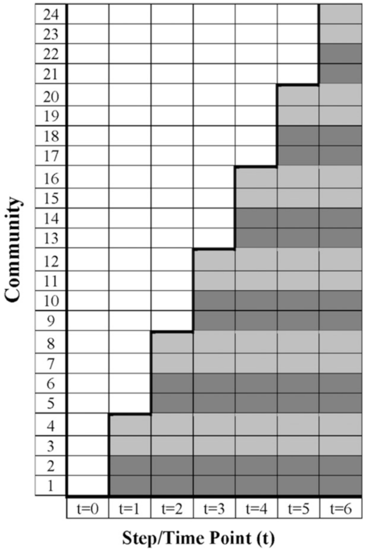


FIGURE 1. Stepped Wedge Schematic for the Mesita Azul Intervention Study. Twenty-four clusters were enrolled at baseline (t = 0) and randomly ordered. All communities started in the control group (white squares). The first four randomly ordered communities (crossover group 1) received the intervention (gray squares) at Step-1 (t = 1). The next four communities (crossover group 2) received the intervention in Step-2, and so on. Within each crossover group two communities were randomized to the Basic Program (dark gray squares) and two to the Enhanced Program (light gray squares). Once a community crossed-over it remained in the intervention group for the remainder of the study. Randomized rollout balances covariates between control- (white squares) and intervention- (all gray squares) periods and creates two comparison groups.

Encouragement design

- In urban Morocco, households that rely on public taps spend a substantial amount of time collecting water.
- A private energy company offered residents lacking an in-home connection the opportunity to a buy a connection to the water and sanitation network at full price, but on interest-free credit.
- Because both treatment and comparison groups were eligible for the loan program, the investigators used a randomized encouragement design to evaluate the impact of the intervention.
- Households that were randomized to the treatment group were pre-approved for the loan and were given assistance with applying for the program from a project officer.
- The treated group (69%) was more likely to buy a connection than controls (10%) 6 mos. after the awareness campaign.

Devoto, 2012[18]

Potential issues with randomized evaluations

- The great benefit of randomization is that it insures that treated and control groups are, in expectation, balanced on all measured and unmeasured covariates.
- This is extraordinarily powerful for achieving high internal validity.
- What could possibly go wrong?
  - Non-compliance.
  - Attrition.
  - Spillovers.
  - Blinding (esp. in clinical trials).

Strategies to limit non-compliance in impact evaluations

- Reduce barriers to take up of the program.
  - Consider how difficult it may be for participants to access program.
- Create incentives to take up the program.
  - Even small incentives/gifts can make a difference.
  - Too large, however, and risk altering outcomes.
- Simplify the program delivery.
  - Staff training of intervention delivery.
  - Try to reduce number of decisions made by implementers.
- Include a "basic" version of the program, if possible.
  - At least something for everyone may reduce treatment seeking by those assigned to control.

Problems with generalizability

- Like any study, results from experiments cannot be generalized beyond their context, without making additional assumptions.

- Implementer effects: The more unique and small-scale the implementer, the less the program can be replicated elsewhere; moreover, RCTs are often done on convenience samples.
- Small-scale trials are environmentally dependent: the program is unlikely to have the same effect if it were implemented in a different context, with different social and cultural factors at play.
- New methodological developments around transportability of effects may help.

Administrative data example: Paying your tax on time

- The UK Behavioural Insights Team evaluated whether social norms (i.e., shaming) could increase delinquent tax payments[19].
- Verified outcomes using administrative tax data (it worked).

www.hmrc.gov.uk

Dear Sir/Madam

Date of issue 4 August 2011
Reference REFERENCE NUMBER

Please pay £999999999999.99

Our records show that your Self Assessment tax payment is overdue.

It is easy to pay. Please call the phone number above to pay by debit card, credit card, or Direct Debit.

You can also pay using internet and telephone banking. For more information on when and how to pay, go to www.hmrc.gov.uk/payinghmrc

If you don't believe that this payment is overdue, please contact us on the number above.

If you have already paid, thank you. If not, please act now.

Yours faithfully

Officer of Revenue and Customs

www.hmrc.gov.uk

Dear Sir/Madam

Date of issue 4 August 2011
Reference REFERENCE NUMBER

BEHAVIORAL INSIGHTS TEAM

Please pay £999999999999.99

Our records show that your Self Assessment tax payment is overdue.

The great majority of people in your local area pay their tax on time. Most people with a debt like yours have paid it by now. **

It is easy to pay. Please call the phone number above to pay by debit card, credit card, or Direct Debit.

You can also pay using internet and telephone banking. For more information on when and how to pay, go to www.hmrc.gov.uk/payinghmrc

If you don't believe that this payment is overdue, please contact us on the number above.

If you have already paid, thank you. If not, please act now.

Yours faithfully

Officer of Revenue and Customs

US not moved by similar shaming

- Design not generalizable to US Medicare (or prescribers).

By Adam Sacarny, David Yokum, Amy Finkelstein, and Shantanu Agrawal

Medicare Letters To Curb Overprescribing Of Controlled Substances Had No Detectable Effect On Providers

DOI: 10.1377/hlthaff.2015.1025
HEALTH AFFAIRS 35,
NO. 3 (2016): 471–479

ABSTRACT Inappropriate prescribing is a rising threat to the health of Medicare beneficiaries and a drain on Medicare's finances. In this study we used a randomized controlled trial approach to evaluate a low-cost, light-touch intervention aimed at reducing the inappropriate provision of Schedule II controlled substances in the Medicare Part D program. Potential overprescribers were sent a letter explaining that their practice patterns were highly unlike those of their peers. Using rich administrative data, we were unable to detect an effect of these letters on prescribing. We describe ongoing efforts to build on this null result with alternative interventions. Learning about the potential of light-touch interventions, both effective and ineffective, will help produce a better toolkit for policy makers to improve the value and safety of health care.

Adam Sacarny (ajs2102@columbia.edu) is an assistant professor in the Department of Health Policy and Management at the Mailman School of Public Health, Columbia University, in New York City.

David Yokum is a fellow at the White House Social and Behavioral Sciences Team and director of the Office of Evaluation Sciences in the General Services Administration, both in Washington, D.C.

Amy Finkelstein is the Ford

Sacarny 2016[20]

A common critique of randomized evaluations

..."RCTs are of limited value since they focus on very small interventions that by definition only work in certain contexts. It's like designing a better lawnmower—and who wouldn't want that? —unless you're in a country with no grass, or where the government dumps waste on your lawn."

Nobel Laureate Angus Deaton
"Debates in Development" (2012)

https://nyudri.wordpress.com/initiatives/deaton-v-banerjee/

What can we randomize?

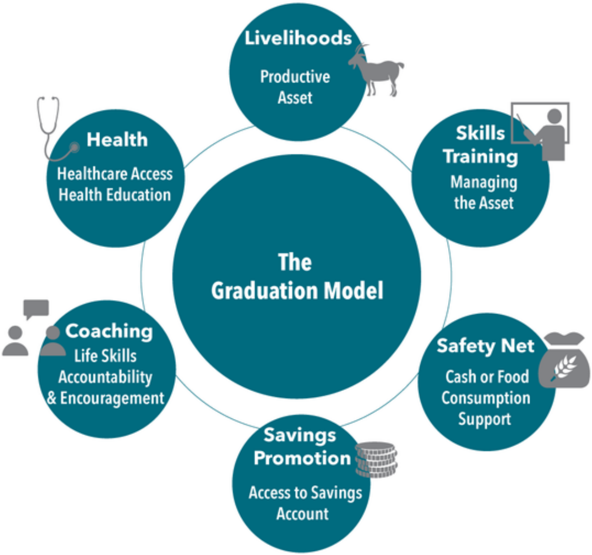
A multifaceted program causes lasting progress for the very poor: Evidence from six countries

Abhijit Banerjee,[1,2,3,4] Esther Duflo,[1,2,3,4] Nathanael Goldberg,[5] Dean Karlan,[2,3,4,5,6*]
Robert Osei,[7] William Parienté,[4,8] Jeremy Shapiro,[9]
Bram Thuysbaert,[5,10] Christopher Udry[2,3,4,6]

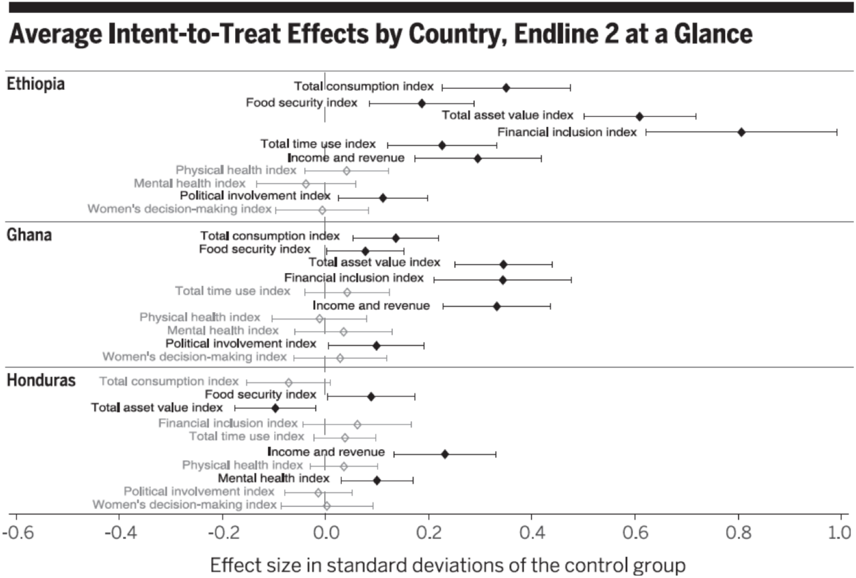
We present results from six randomized control trials of an integrated approach to improve livelihoods among the very poor. The approach combines the transfer of a productive asset with consumption support, training, and coaching plus savings encouragement and health education and/or services. Results from the implementation of the same basic program, adapted to a wide variety of geographic and institutional contexts and with multiple implementing partners, show statistically significant cost-effective impacts on consumption (fueled mostly by increases in self-employment income) and psychosocial status of the targeted households. The impact on the poor households lasted at least a year after all implementation ended. It is possible to make sustainable improvements in the economic status of the poor with a relatively short-term intervention.

Banerjee, 2015[21]

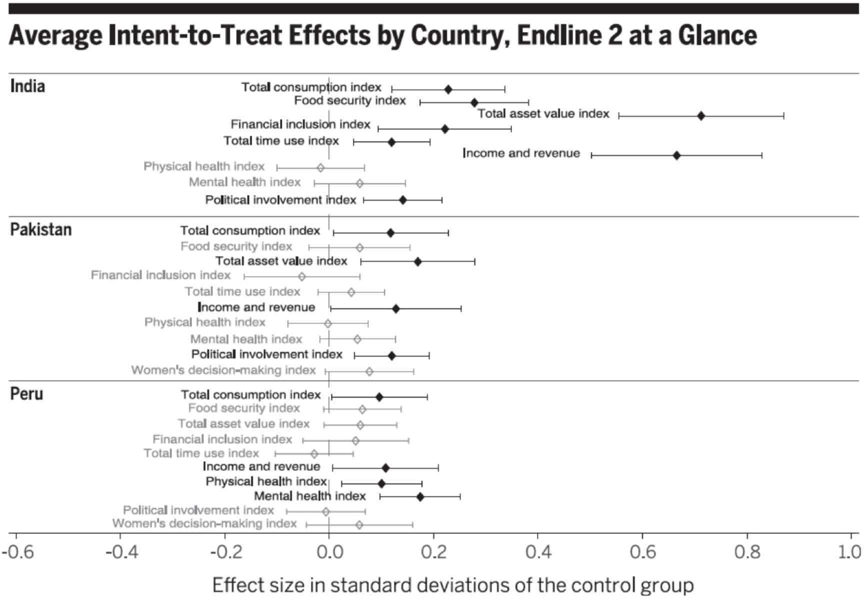
This is a complex, tailored intervention



Banerjee, 2015[21]



Banerjee, 2015[21]



Banerjee, 2015[21]

Benefit of prospective RCTs: pushing innovation

- Gates: we should give chickens, esp. to the extreme poor in sub-Saharan Africa
  - Established benefits to ultrapoor (graduation approach)
  - Inexpensive and easy to care for.
  - Good investment and promotes empowerment.
- Blattman: Maybe but...
  - Inexpensive relative to what? Graduation approach is expensive ($1700 per participant).
  - Cites evidence that giving cash is cheaper and more flexible.
  - "by betting on either cash or chickens, you and I are gambling with poor people's lives. We don't actually know who is right."
- "It would be straightforward to run a study with a few thousand people in six countries, and eight or 12 variations, to understand which combination works best, where, and with whom."

---

http://www.vox.com/the-big-idea/2017/3/14/14914996/bill-gates-chickens-cash-africa-poor-development

Summary: Randomized interventions

- The great benefit of randomization is that it insures that treated and control groups are, in expectation, balanced on all measured and unmeasured covariates.
- Allow you to focus prospectively on designing an evaluation to answer a clear and precise question about a well-defined intervention.
- Expensive, but robust evidence and transparent methods usually worth the investment.
- Not useful for all situations, many opportunities to fail.

New York to make state college tuition free for middle class

THE ASSOCIATED PRESS
Sunday, April 9, 2017, 11:10 PM

ALBANY, N.Y. (AP) — New York will be the first state to make tuition at public colleges and universities free for middle-class students under a state budget approved by lawmakers Sunday.

The plan crafted by Democratic Gov. Andrew Cuomo will apply to any New York student whose family has an annual income of $125,000 or less. To qualify the student would have to meet certain class load and grade point average restrictions, and room and board would not be covered.

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.

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.

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!"

Free primary education and infant mortality

- Many observed differences between treatment groups.
- Is assuming "no other unmeasured differences" credible?

Are tuition-free primary education policies associated with lower infant and neonatal mortality in low- and middle-income countries?

Amm Quamruzzaman [a], José M. Mendoza Rodríguez [b], Jody Heymann [c], Jay S. Kaufman [d], Arijit Nandi [e, *]

[a] Department of Sociology & Institute for Health and Social Policy, McGill University, Montreal, QC, Canada

Table 1
Sample characteristics for the full sample and stratified by exposure to primary education policies; 37 LMICs included in the Demographic and Health Surveys (DHS), 2003–2011.

	Total		Primary education			
			Not free		Free	
Infant mortality sample	33,735		12,443 (36.9%)		21,292 (63.1%)	
Variables	Mean[a]	SD[b]	Mean	SD	Mean	SD
Death before age of 1 year (Yes = 1)	0.08	0.27	0.09	0.29	0.07	0.25
Gender of child (female = 1)	0.49	0.50	0.49	0.50	0.49	0.50
Urban residence (Yes = 1)	0.26	0.44	0.23	0.42	0.28	0.45
SES1: 1st (lowest) wealth quintile (Yes = 1)	0.26	0.44	0.26	0.44	0.26	0.44
SES2: 2nd wealth quintile (Yes = 1)	0.25	0.43	0.24	0.43	0.25	0.44
SES3: 3rd wealth quintile (Yes = 1)	0.22	0.41	0.21	0.41	0.22	0.42
SES4: 4th wealth quintile (Yes = 1)	0.17	0.37	0.18	0.38	0.16	0.37
SES5: 5th (highest) wealth quintile (Yes = 1)	0.10	0.30	0.11	0.31	0.10	0.30
Household head has completed primary education (Yes = 1)	0.27	0.45	0.20	0.40	0.31	0.46
Female household head (Yes = 1)	0.16	0.37	0.14	0.35	0.17	0.38
GDP per capita (PPP, constant 2005 USD)[c]	1.62	1.31	1.10	0.54	1.91	1.50
Percentage of population in urban centers[d]	3.12	1.52	2.73	0.91	3.33	1.72
Health expenditure per capita (PPP, constant 2005 USD)[e]	0.61	0.54	0.38	0.15	0.74	0.63

Quamruzzaman et al., 2014 [22]

Ex: Neighborhood block parties and health in Philly

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

Characteristics of Philadelphia neighborhoods (census tract averages)	Block parties (n = 293)		No block parties (n = 88)		p-value
	Mean	Std. Dev.	Mean	Std. Dev.	
Population demographics					
Total population (Census 2000)	4287.66	2235.73	2968.92	2696.39	<0.001
Percent White	39.79	35.39	61.83	34.95	<0.001
Percent Black	48.06	37.44	21.80	29.26	<0.001
Percent male	45.98	5.65	45.36	16.61	0.58
Percent female	52.99	6.25	47.82	17.12	<0.001
Mean age	33.96	6.75	35.95	12.82	0.06
Percent high school graduates	69.54	15.15	67.30	26.89	0.32
Percent below 200% poverty line	42.98	19.01	26.24	20.95	<0.001
Percent below 100% poverty line	23.17	13.94	13.86	14.89	<0.001
Percent commuting more than 30 min	50.93	13.82	42.16	20.69	<0.001
Neighborhood characteristics					
Number of households	1664.68	914.02	1162.73	1132.40	<0.001
Number of families	984.13	571.76	727.06	709.38	<0.001
Number of housing units	1879.77	994.29	1263.49	1215.52	<0.001
Percent owner occupied	56.64	20.09	54.99	30.81	0.56
Percent renter occupied	42.33	19.72	35.92	27.82	0.02
Segregation index, Whites from Blacks	0.43	0.35	0.66	0.34	<0.001
Segregation index, Blacks from Whites	0.51	0.35	0.22	0.28	<0.001
Residential blocks per square mile	208.64	104.82	114.42	129.01	<0.001
Percent overcrowded	0.63	0.99	0.32	0.85	0.008

Dean et al. (2015)

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.

Why are quasi-experiments useful?

- Under various sets of assumptions, they provide valid estimates of causal effects.
- So what? Can't regression do that? ("We adjusted for everything!").
- Yes. So what makes quasi-experiments special?
  - Plausibly random treatment assignment.
  - Stronger design for "identifying" causal effects.
- May also expand the kinds of interventions that are possible to evaluate:
  - Rare outcomes (e.g., pesticides and suicide).
  - Infeasible/unethical RCTs.
- Can avoid selection mechanisms that lead some weird people to participate in trials.

Policy Question Example: Will More Police on the Streets Reduce Crime?

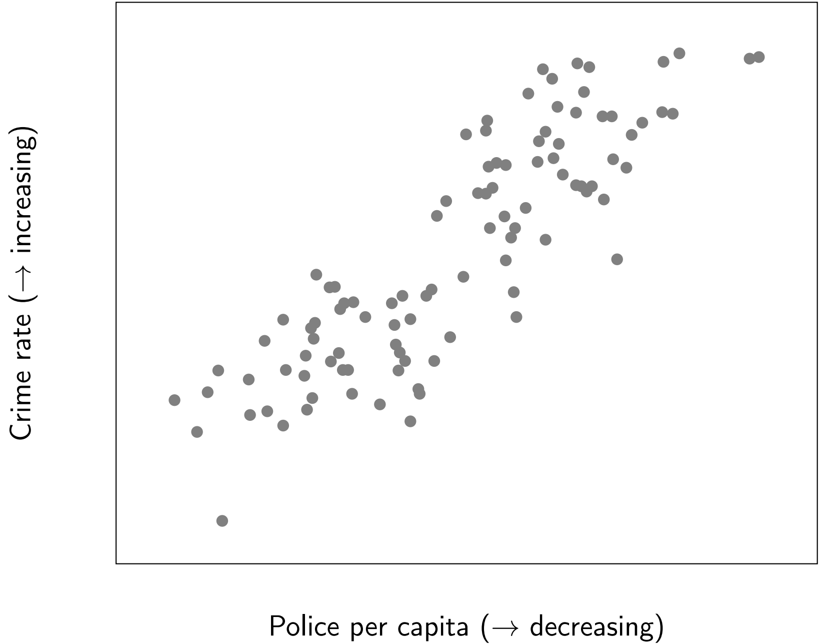
One strategy:

- Gather as much data as you can on police, crime, and possible confounders for c cities.
- Estimate some regression model, e.g.:

$$Crime_c = \alpha + \beta Police_c + \gamma Confounders_c + \varepsilon_c$$

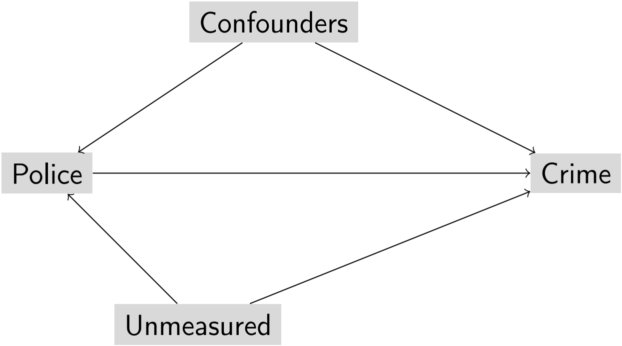
- Hope that assumptions hold (no unmeasured confounding, temporal ordering).
- Interpret β as the "effect" of police on crime.

Why might this not be causal?



Challenge of unmeasured confounding

- Levels of police and crime are likely to have common causes that cannot all be measured.
- Failure to account for such factors will falsely attribute their effects to police presence.



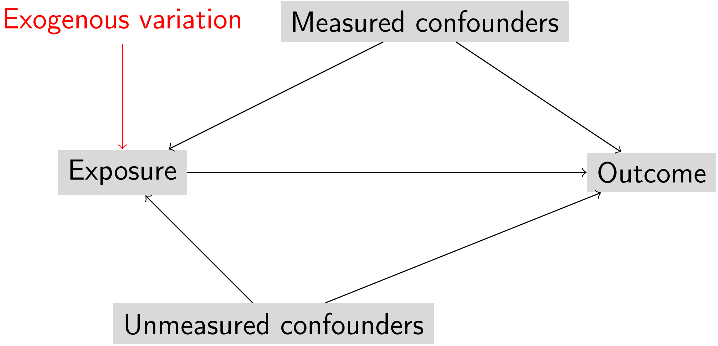
Do More Police on the Streets Reduce Crime?

- Time to get clever: What might prompt an exogeneous change in the quantity of police?
- US Homeland Security Advisory Terror Alert System:
  - Low
  - Guarded
  - Elevated
  - High $\rightsquigarrow$ More cops!
  - Severe $\rightsquigarrow$ More cops!
- In DC, during heightened alert periods, effective police presence increases by 50 percent.
- High alert days associated with fewer crimes.
- Not confounded by changes in tourism on high alert days.

Klick and Tabarrok, 2005[23]

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

- Selection on observables:
  - Stratification (tabular analysis)
  - Adjustment (usually OLS regression)
  - Matching (pre-processing to create treated and control groups)
- Selection on unobservables:
  - Difference-in-differences
  - Interrupted time series
  - Instrumental variables
  - Regression discontinuity
- Selecting on "unobservables" = natural experiments

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

Instrumental variables

Why use instrumental variables?

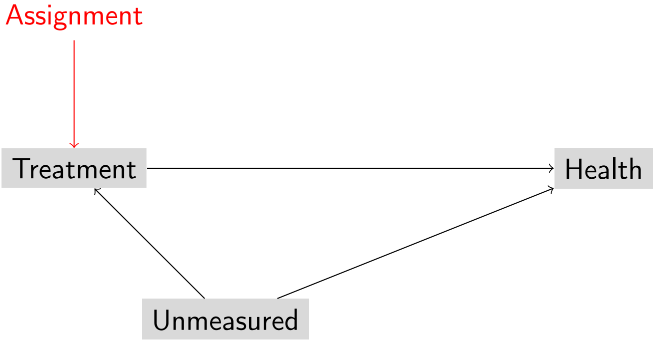
- Trial may be impossible or unethical (especially for many social exposures)
- We may actually want to know the effect of T on Y.
- We are concerned about unmeasured confounding for the effect of T on Y.
- Many examples of social exposures where this is problematic:
    - Education
    - Income
   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

Thinking about basic research design

- Remember that quasi-experimental designs and natural experiments are trying to mimic an RCT as closely as possible.
- 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?

Hypothetical randomized assignment

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



Randomized natural experiment example

- For example, Angrist et al. (2002) use a lottery that assigned school-choice vouchers in Colombia as an instrumental variable for using a school-choice voucher.
- Vouchers were assigned randomly because of excess demand.
- However, not all winners used them.
- Is this a "natural experiment" or an RCT?

Angrist et al. 2002

Randomization increases credibility

- Voucher winners look similar to losers on measured characteristics:

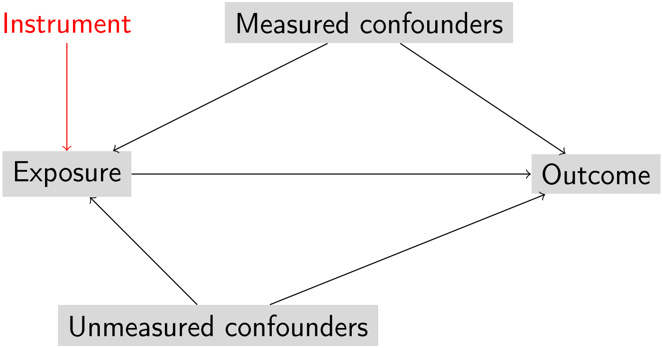
TABLE 2—PERSONAL CHARACTERISTICS AND VOUCHER STATUS

	Bogotá 1995		Bogotá 1997		Jamundi 1993		Combined sample	
Dependent variable	Loser means	Won voucher	Loser means	Won voucher	Loser means	Won voucher	Loser means	Won voucher
A. *Data from PACES Application:*								
Has phone	0.882	0.009	0.828	0.029	0.301	0.068	0.825	0.017
		(0.011)		(0.025)		(0.052)		(0.010)
Age at time of application	12.7	−0.086	12.7	−0.227	12.7	−0.383	12.7	−0.133
	(1.3)	(0.045)	(1.5)	(0.102)	(1.5)	(0.162)	(1.4)	(0.040)
Male	0.493	0.013	0.484	0.007	0.386	0.114	0.483	0.019
		(0.017)		(0.044)		(0.055)		(0.015)
N	1,519	3,661	256	1,736	166	334	1,941	5,731

Angrist et al 2002

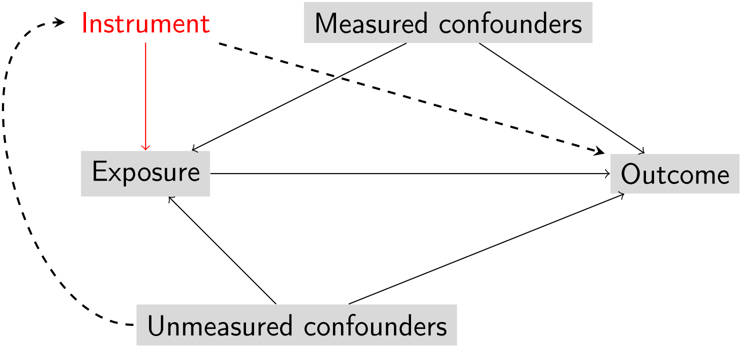
Non-randomized instrumental variable

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



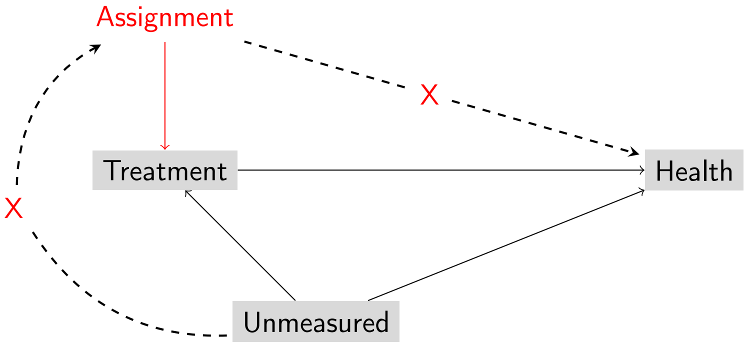
Non-randomized instrument creates additional issues

- 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.



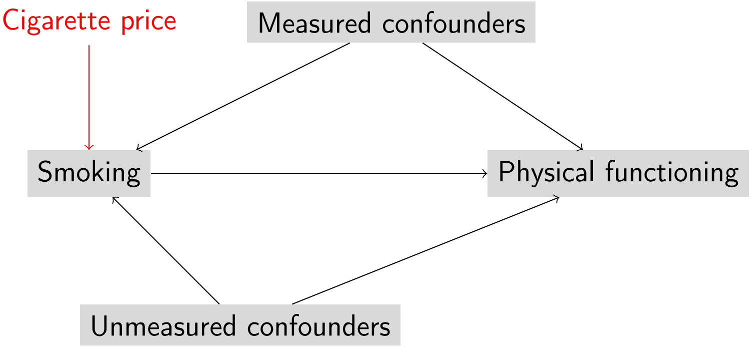
Key IV Assumptions

1) Instrument affects treatment; 2) Instrument only affects outcome via treatment; 3) No common causes of instrument and outcome.



Non-randomized examples of IV: Policies

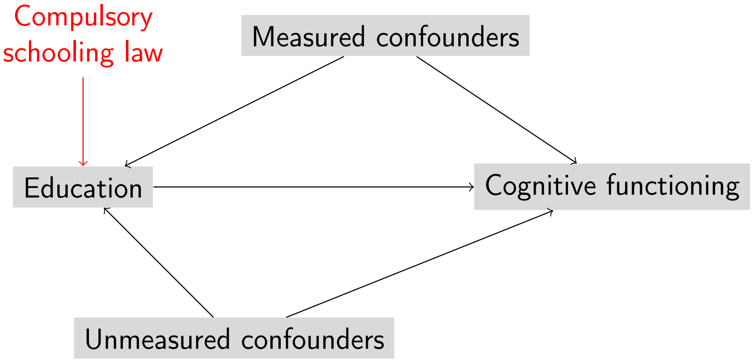
- Does smoking (T, 1=yes, 0=no) affect physical functioning (Y)?
- **Instrument**: changes in cigarette prices [mimicking random assignment].



Leigh and Schembri 2004 [24]

Non-randomized examples of IV: Policies

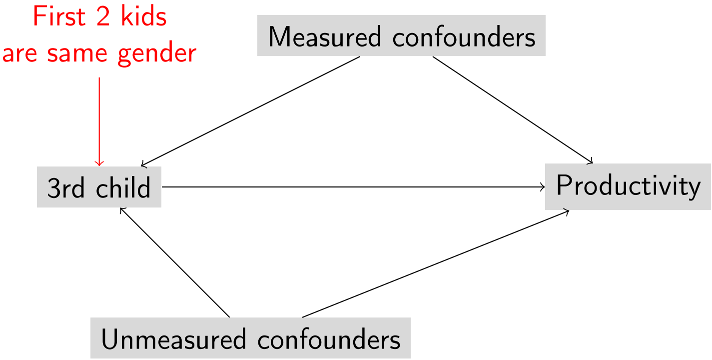
- Does education (T, 1=yes, 0=no) affect cognitive functioning (Y)?
- **Instrument**: changes in compulsory schooling laws [mimicking random assignment].



Glymour et al. 2008 [25]

Non-randomized examples of IV: Gender

- Does a third child (T, 1=yes, 0=no) affect work productivity (Y)?
- **Instrument**: gender concordance of first 2 children [mimicking random assignment].



Angrist et al. 1998 [26]

Intuition for the last example

- The substantive question: we want to know how additional children affect work among parents.
- Could just compare people with 2 vs. 3 kids and adjust, but they may differ in lots of ways that are hard to measure (careerism, assets, etc.).
- RCT is (likely) impossible.

Intuition for the last example

- Would be great to find an instrument, but that requires assumptions:
- Relevance:
  - Most families like gender diversity among children.
  - If first 2 are same gender, more likely to have a 3rd.
- Exclusion restriction:
  - In most countries people have little control over their child's gender.
  - Gender concordance unlikely to affect work productivity apart from its impact on having another child.
- Anyone buying this?

Impact Evaluation Example: Child Labor and Schooling

- Ravallion and Wodon [27] used Bangladesh's Food-For-Education (FFE) program to study whether child labor affects kids' schooling.
- Problem:
  - Parents largely choose either to send their kids to work or to school.
  - Concerns that child labor takes the place of schooling.
  - Households where kids work more vs. less are likely to be different in lots of ways that are correlated with schooling.
- IV idea:
  - FFE provides food subsidies to keep poor rural children in school.
  - More food reduces the "price" for parents to send kids to school.
  - Used geographic targeting as an instrument for individual participation.

Impact Evaluation Example: Child Labor and Schooling

- Found that the FFE program increased schooling but had little effect on child labour.
- An extra 100 kg of rice increased the probability of going to school by:
  - 0.17 for boys.
  - 0.16 for girls.
- If parents are substituting labour for school, then the program should also reduce child labour.
- Found little evidence and weak impact of FFE program on child labor.

Not all instruments are created equal

Food Price Spikes Are Associated with Increased Malnutrition among Children in Andhra Pradesh, India[1–3]

Sukumar Vellakkal,[4,6]* Jasmine Fledderjohann,[4] Sanjay Basu,[5] Sutapa Agrawal,[6] Shah Ebrahim,[7] Oona Campbell,[7] Pat Doyle,[7] and David Stuckler[4,6]

[4]Department of Sociology, University of Oxford, Oxford, United Kingdom; [5]Department of Medicine, Stanford University, Stanford, CA; [6]Public Health Foundation of India, New Delhi, India; and [7]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom

Abstract

Background: Global food prices have risen sharply since 2007. The impact of food price spikes on the risk of malnutrition in children is not well understood.

Objective: We investigated the associations between food price spikes and childhood malnutrition in Andhra Pradesh, one of India's largest states, with >85 million people. Because wasting (thinness) indicates in most cases a recent and severe process of weight loss that is often associated with acute food shortage, we tested the hypothesis that the escalating prices of rice, legumes, eggs, and other staples of Indian diets significantly increased the risk of wasting (weight-for-height *z* scores) in children.

Methods: We studied periods before (2006) and directly after (2009) India's food price spikes with the use of the Young Lives longitudinal cohort of 1918 children in Andhra Pradesh linked to food price data from the National Sample Survey Office. Two-stage least squares instrumental variable models assessed the relation of food price changes to food consumption and wasting prevalence (weight-for-height *z* scores).

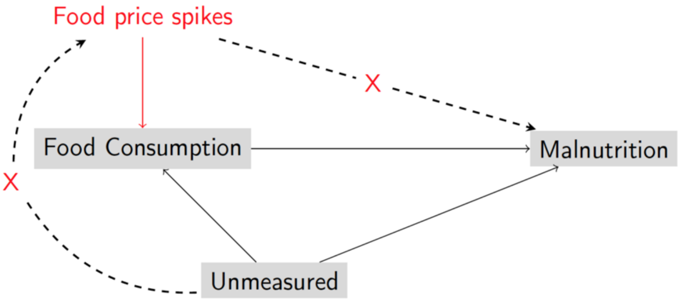
Results: Before the 2007 food price spike, wasting prevalence fell from 19.4% in 2002 to 18.8% in 2006. Coinciding with India's escalating food prices, wasting increased significantly to 28.0% in 2009. These increases were concentrated among low- (χ^2: 21.6, $P < 0.001$) and middle- (χ^2: 25.9, $P < 0.001$) income groups, but not among high-income groups (χ^2: 3.08, $P = 0.079$). Each 10.0 rupee ($0.170) increase in the price of rice/kg was associated with a drop in child-level rice consumption of 73.0 g/d (β: −7.30; 95% CI: −10.5, −3.90). Correspondingly, lower rice consumption was significantly associated with lower weight-for-height *z* scores (i.e., wasting) by 0.005 (95% CI: 0.001, 0.008), as seen with most other food categories.

Conclusion: Rising food prices were associated with an increased risk of malnutrition among children in India. Policies to help ensure the affordability of food in the context of economic growth are likely critical for promoting children's nutrition. *J Nutr* doi: 10.3945/jn.115.211250

Vellakkal et al 2015[28]

Don't forget the assumptions!

- Instrument is district variations in food prices.
- For valid IV analysis, recall we must satisfy:
  - Relevance: The instrument must affect the treatment.
  - Exclusion restriction: No effect on outcome except via treatment.
  - No common causes of instrument and outcome.



Vellakkal et al 2015[28]

Challenges

- The biggest challenge in an IV analysis is finding a valid instrument; i.e., a Z that is correlated with T but not Y (other than via T).
- Finding a good IV is based on deep substantive knowledge of the processes shaping T and Y:
  - Institutional knowledge.
  - Ideas about exposure process.
  - A well-developed DAG.
- Without that knowledge, fancy methods won't help you.

How to find an IV

Common sources of instruments include:

- Nature: geography, weather (rainfall), biology in which a truly random source of variation influences T (no possible reverse causation)
- History: things determined a long time ago, which were possibly endogenous contemporaneously, but which no longer plausibly influence Y
- Institutions: formal or informal rules that influence the assignment of T in a way unrelated to Y.
- In health care, clinical practice patterns and guidelines can be useful (because they are often arbitrarily defined and/or applied).
- Trials and Policies
  - Randomized encouragement designs
  - Public policy changes (DD is often used in the first stage)

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

Remember basic 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?

RD: Basic Idea

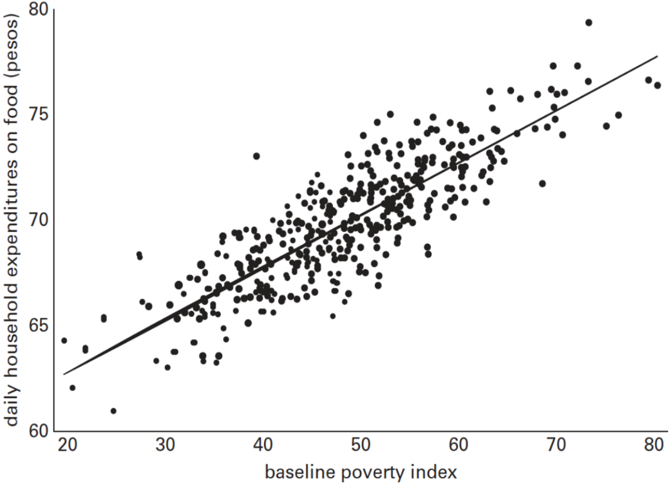
- Take advantage of arbitrary thresholds that sometimes assign treatment to individuals.
- When an administrative or rule-based cutoff in a continuous variable (present in your data) predicts treatment assignment, being on one side or the other of this cutoff determines, or predicts, treatment received.
- 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.

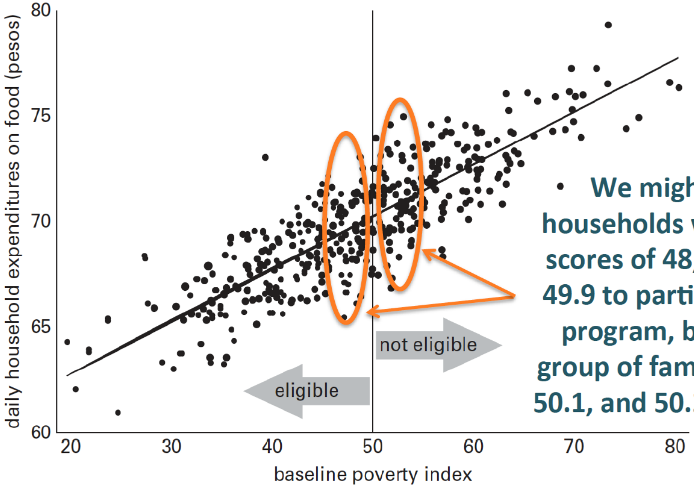
Source: Gertler, 2011[29]

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



Source: Gertler, 2011[29]

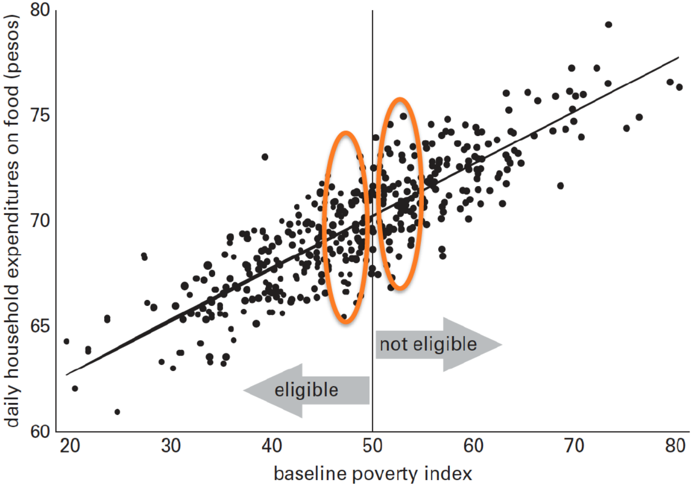
Under the program's rules, only households with a poverty score, Z, below 50 are eligible for the cash payment:



- 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

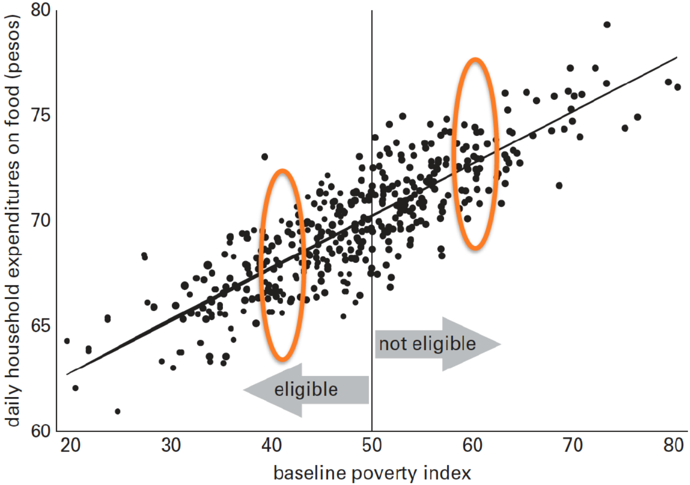
Source: Gertler, 2011[29]

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



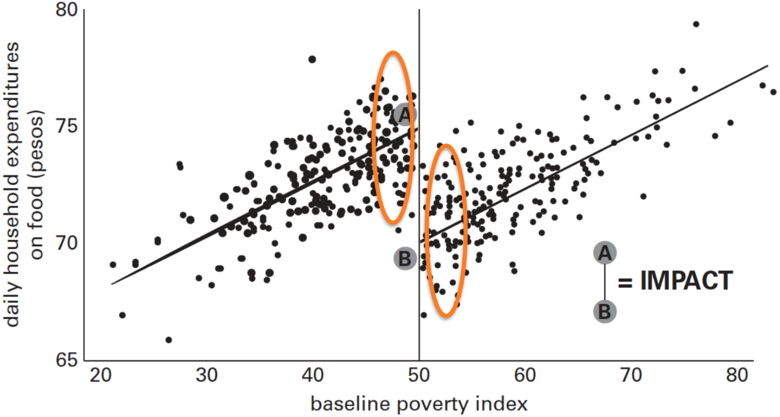
Source: Gertler, 2011[29]

How about these families?



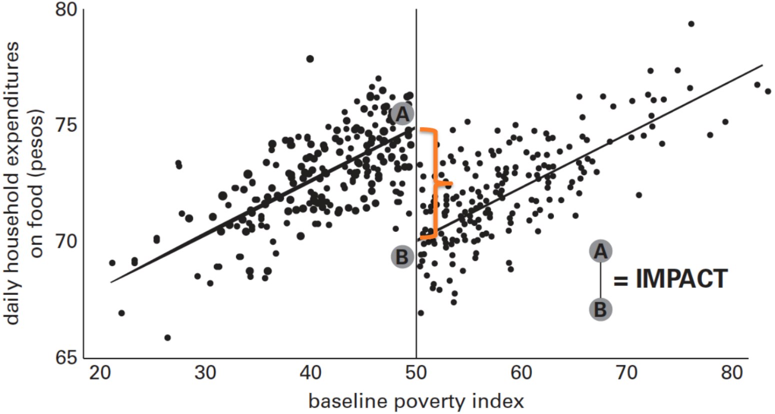
Source: Gertler, 2011[29]

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[29]

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



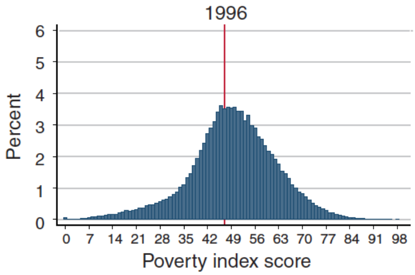
Source: Gertler, 2011[29]

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)

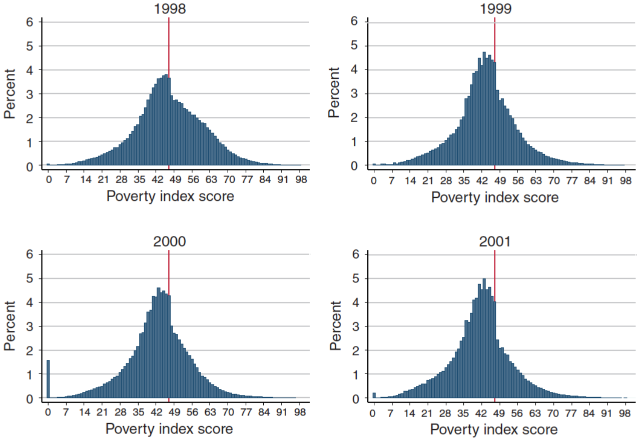
- Colombian census collected comprehensive information on dwelling characteristics, demographics, income, and employment to assign a poverty index score to each family.
- Eligibility rules for several social welfare programs use specific thresholds (score=47) from the poverty index score.
- Prior to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



Camacho and Conover 2011[30]

Manipulation example (poverty threshold in Colombia)

- After 1997, the algorithm was provided to municipal administrators, leading to evidence of manipulation.
- Reduces exchangeability between treatment groups at threshold (bias).



Recent RD studies in health

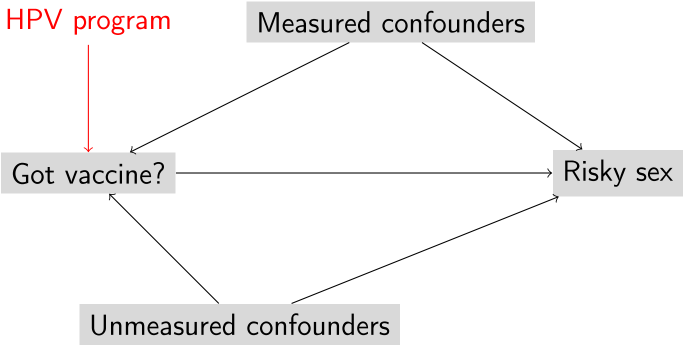
Table 1. PubMed articles with health outcomes using regression discontinuity designs

Authors	Year	Journal	Study topic
Albouy and Lequien [23]	2009	Journal Health Economics	Effect of education on mortality
Almond et al. [15]	2010	Quarterly Journal of Economics	Returns to treatment of low-birth-weight newborns
Andalón [24]	2011	Health Economics	Effect of Oportunidades on obesity
Anderson et al. [25]	2011	Journal of Health Economics	Effect of schooling on children's BMI
Arcand and Wouabe [26]	2010	Health Economics	Effect of teacher training on HIV prevention
Banks and Mazzonna [27]	2012	Economics Journal	Effect of education on old-age cognitive ability
Behrman [28]	2014	Social Science and Medicine	Effect of primary schooling on HIV status
Bor et al. [1]	2014	Epidemiology	Effect of early vs. deferred HIV treatment on mortality
Callaghan et al. [29]	2014	Drug and Alcohol Dependence	Effect of legal drinking age on mortality
Callaghan et al. [30]	2013	American Journal of Public Health	Effect of legal drinking age on alcohol-related morbidity
Callaghan et al. [31]	2013	Addiction	Effect of legal drinking age on inpatient morbidity
Carpenter and Dobkin [16]	2009	AEJ: Applied Economics	Effect of alcohol consumption on mortality
Carpenter and Dobkin [32]	2011	Journal of Economic Perspectives	Minimum legal drinking age and public health
Chen et al. [33]	2013	PNAS	Effect of air pollution on mortality
Conover and Scrimgeour [34]	2013	Journal of Health Economics	Health effects of minimum legal drinking age
De La Mata [35]	2012	Health Economics	Effect of Medicaid eligibility on coverage, utilization, and health
Deza [36]	2014	Health Economics	Effect of alcohol use on drug consumption
Flam-Zalcman et al. [37]	2012	Intl J Psych Research	Effect of criterion-based increase in alcohol treatment
Fletcher [38]	2014	Biodemography and Social Biology	Effect of genetics on stress response
Glance et al. [39]	2014	JAMA Surgery	Effect of hospital report cards on mortality
Gormley et al. [40]	2005	Developmental Psychology	Effect of universal pre-kindergarten on cognitive development
Huang and Zhou [41]	2013	Social Science and Medicine	Effect of education of cognition
Jensen and Wust [42]	2014	Journal of Health Economics	Effect of Caesarean section on maternal and child health
McFarlane et al. [43]	2014	Schizophrenia Bulletin	Effect of treatment program on psychosis onset
Miller et al. [44]	2013	AEJ: Applied Economics	Effect of insurance on health spending, utilization, and health
Nishi et al. [45]	2012	Bulletin of the WHO	Health effects of patient cost-sharing
Pierce et al. [46]	2012	Pers Soc Psych Bulletin	Effect of income disparity in marriage
Sloan and Hanrahan [47]	2014	JAMA Ophthalmology	Effect of new therapies on vision loss among elderly patients
Smith et al. [48]	2014	Canadian Medical Association Journal	Effect of HPV vaccine on sexual behavior
Sood et al. [49]	2014	BMJ	Effect of health insurance on mortality
Weaver et al. [50]	2010	Journal of Traumatic Stress	Effect of cognitive-behavioral therapy on trauma symptoms
Yörük and Yörük [51]	2012	Social Science and Medicine	Effect of alcohol on psychological well-being

Source: Moscoe, 2015[31]

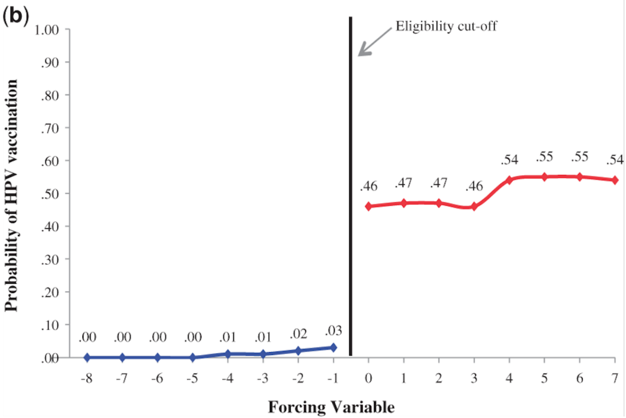
Applied example: HPV vaccine and sexual behaviors

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



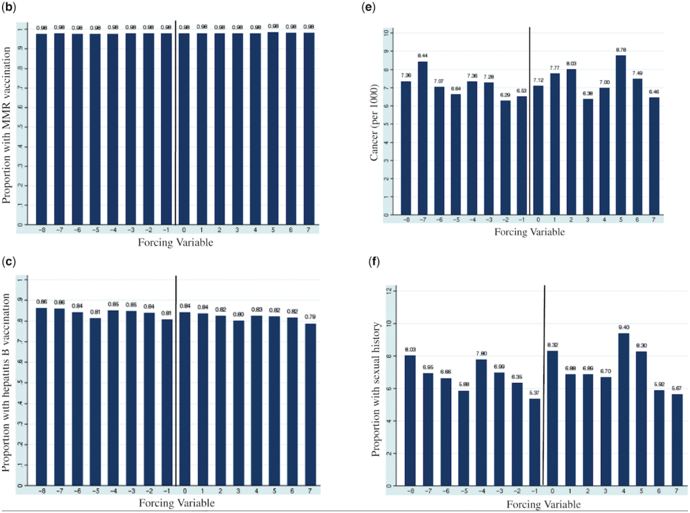
Does the cutoff predict treatment?

- 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[32]

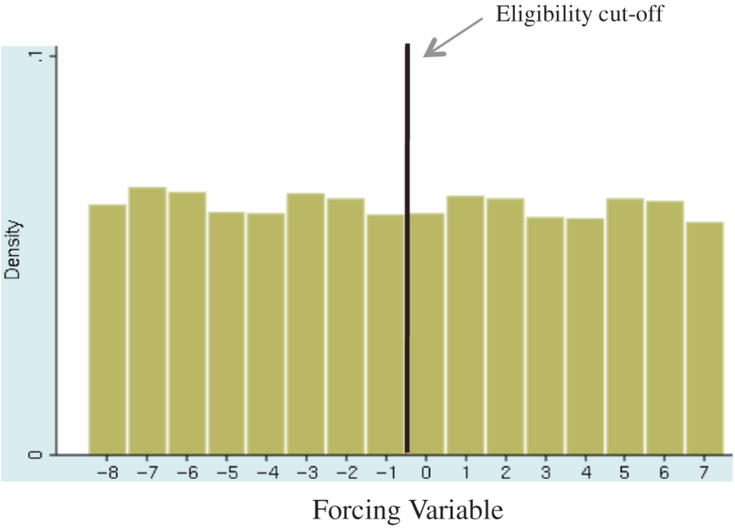
What about confounders?



Smith et al., 2016[33]

Any evidence of manipulating the cutoff?

- Probably not likely here, but an essential diagnostic.



Smith et al., 2016[33]

What does a credible natural experiment look like?

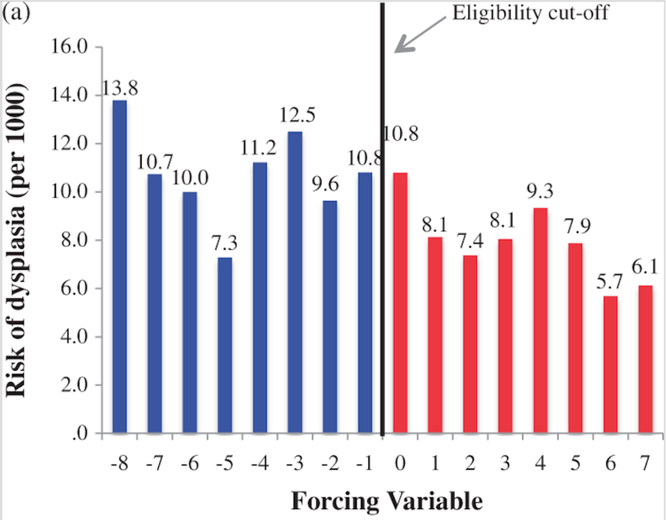
Table 1: Baseline characteristics of the eligibility groups in the study cohort

Characteristic	Program eligibility group; % of eligibility group* Ineligible (*n* = 131 781)	Program eligibility group; % of eligibility group* Eligible (*n* = 128 712)	Characteristic	Program eligibility group; % of eligibility group* Ineligible (*n* = 131 781)	Program eligibility group; % of eligibility group* 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[32]

Graphical results on outcomes

- Treatment reduces cervical displasia.



Smith et al., 2016[33]

Note little impact of adjustment

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

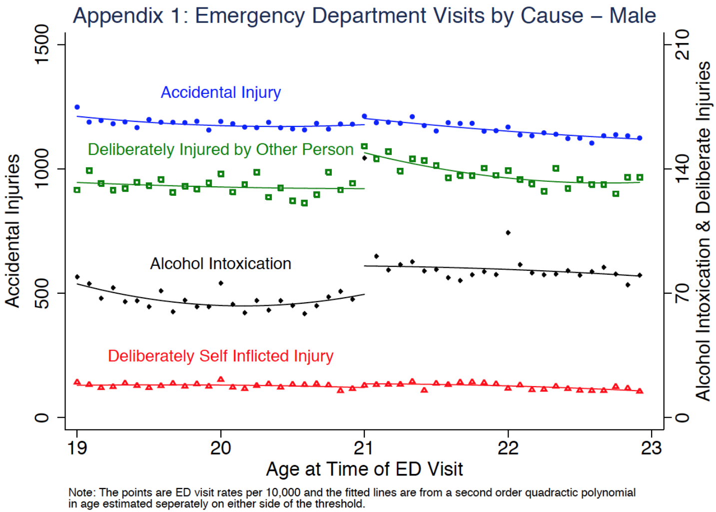
Outcome	No. of excess cases per 1000 girls (95% CI)	RR (95% CI)	Adjusted† RR (95% CI)
Effect of vaccine			
Composite outcome	–0.61 (–10.71 to 9.49)	0.96 (0.81 to 1.14)	0.98 (0.84 to 1.14)
Pregnancy	0.70 (–7.57 to 8.97)	0.99 (0.79 to 1.23)	1.00 (0.83 to 1.21)
STIs	–4.92 (–11.49 to 1.65)	0.81 (0.62 to 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[32]

Another recent example: US drinking age

- Minimum legal drinking age and non-fatal injuries:



Carpenter, 2017[34]

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 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

Best practices for RD study design

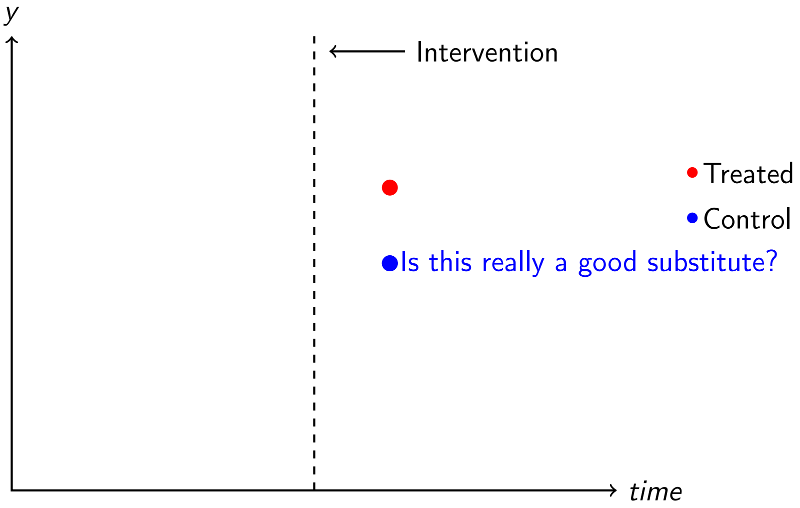
- Need to show convincingly that:
- Treatment changes discontinuously at the cutpoint.
  - Outcomes change discontinuously at the cutpoint.
  - Other covariates do not change discontinuously at the cutpoint.
  - There is no manipulation of the assignment variable.
- Need to argue that:
  - Unobserved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

ITS and Difference-in-Differences

Focus (again) on 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?

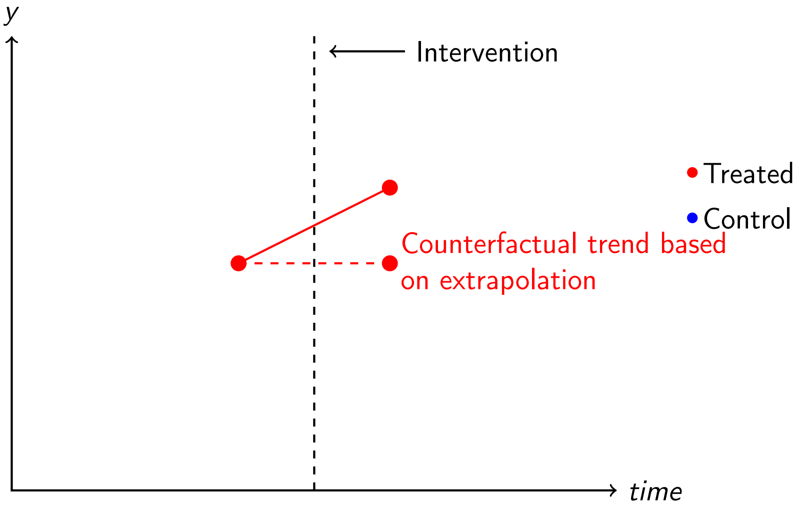
One-group posttest design with control group



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.

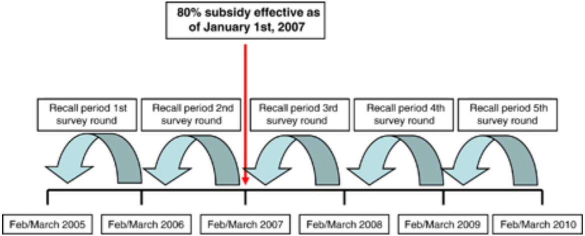
One-group pretest-posttest design



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).*

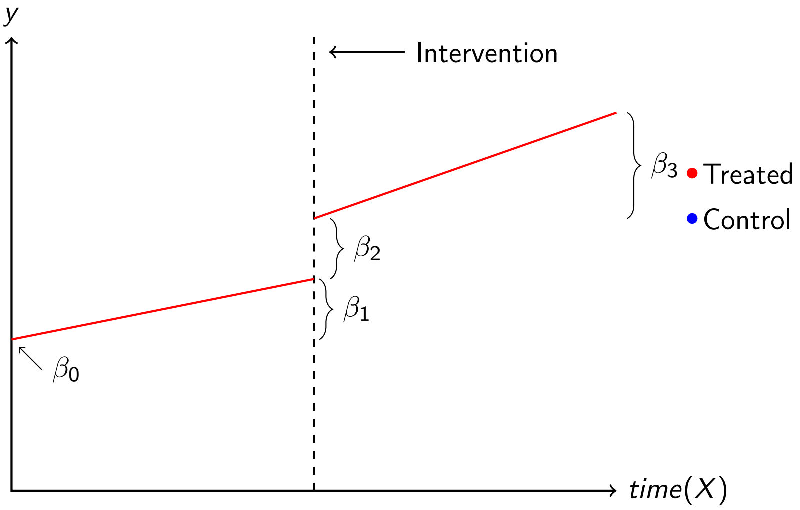


One group pretest-posttest design

- 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.

Visual interpretation of parameters from linear ITS

$$Y_t = \beta_0 + \beta_1 X + \beta_2 D_t + \beta_3 X D_t + \varepsilon_t$$



Estimating intervention effects

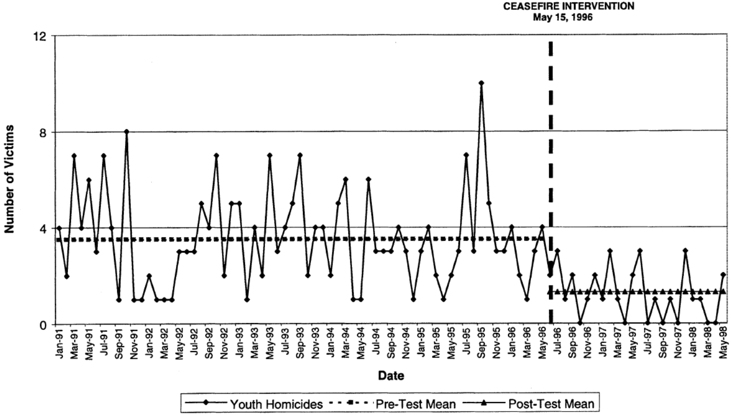
- Recall our model: $Y_t = \beta_0 + \beta_1 X + \beta_2 D_t + \beta_3 X D_t + \varepsilon_t$.
- We can estimate the impact of the intervention by comparing the predicted value of Y with and without the intervention at a given time.
- We do this by setting $\beta_2 = 0$ and $\beta_3 = 0$ and predicting $\hat{Y}$ at a specific time (e.g., end of follow-up):
- In the absence of intervention: $\hat{Y}_{D=0} = \hat{\beta}_0 + \hat{\beta}_1 X$.
- In the presence of intervention: $\hat{Y}_{D=1} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 D + \hat{\beta}_3 X D$.
- Difference: $\hat{\beta}_2 D + \hat{\beta}_3 X D$ is the estimate of the impact.

Analytic challenges with ITS and pre-post studies

- In the absence of a control group that is untreated, most ITS analyses must rely on extrapolating pre-existing trends to estimate counterfactual outcomes.
- As with any non-randomized study that uses models, model assumptions should be checked and sensitivity analyses for main assumptions should be done if possible.
- Key is modeling the pre-intervention trends correctly.
- Several specific challenges for modeling pre-existing trends:
    - Strong secular changes.
    - Autocorrelation.
    - Seasonality.

ITS example: Braga et al. (2001)

- Impact of "Operation Ceasefire" on Boston homicide rates.
- Basic graphical evidence (dashed lines are means):



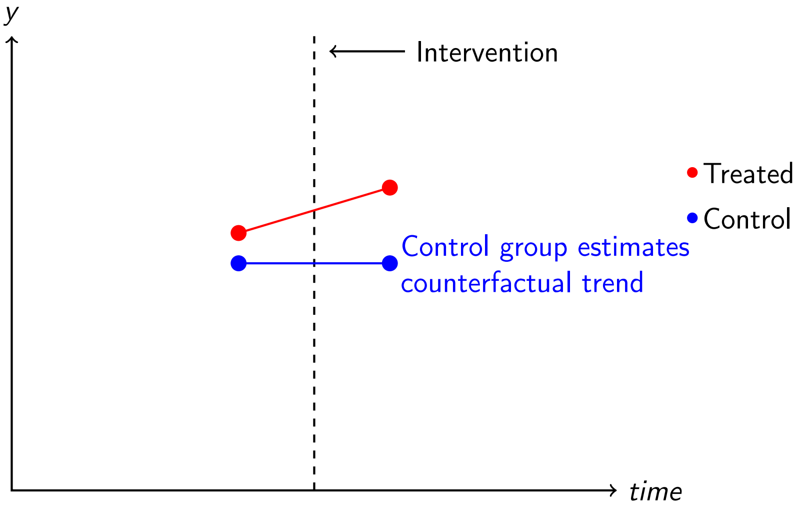
2: Monthly Counts of Youth Homicides in Boston

Braga et al. 2001 [35]

ITS example: Braga et al. (2001)

- Adjusted for seasonality and time trends and found reduction in homicide rates.
- Several sensitivity analyses:
  - Also looked # gun assaults and shots fired (should be affected).
  - Looked at subgroups that should have been more vs. less affected (EMM).
  - Could also have used a placebo outcome (e.g. outcomes that should not be affected by the intervention)
  - Included time-varying covariates (e.g., unemployment rates).
  - **Compared changes in Boston to changes in other cities.**
- Last point leads to comparing differences pre- and post- in one unit relative to an untreated unit.

Adding pretests for both groups



How does this help?

- Pre/post in a control group helps 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.

DD Regression: Two Groups, Two Periods

Basic setup for DD with a single treated and control group, two periods:

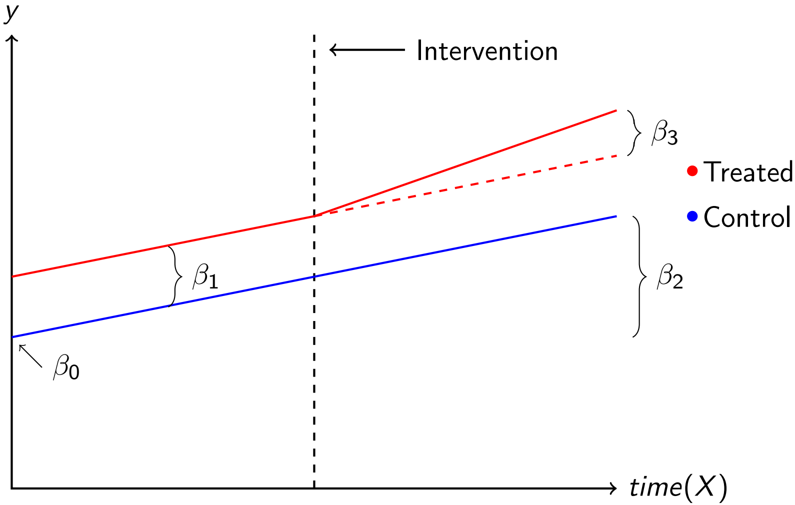
y	group	time	treated?	after?	treatXafter
$\vdots$	1	1	0	0	0
	1	2	0	1	0
$\vdots$	2	1	1	0	0
	2	2	1	1	1

$$Y = \boxed{\beta_0} + \boxed{\beta_1 * treat} + \boxed{\beta_2 * after} + \boxed{\beta_3 * treat * after}$$

- Treatment group
- Post-policy
- Interaction term

Visual interpretation of parameters from linear DD model

$$Y = \beta_0 + \beta_1 T + \beta_2 A + \beta_3 T * A + \varepsilon_t$$



Causal effects without regression?

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

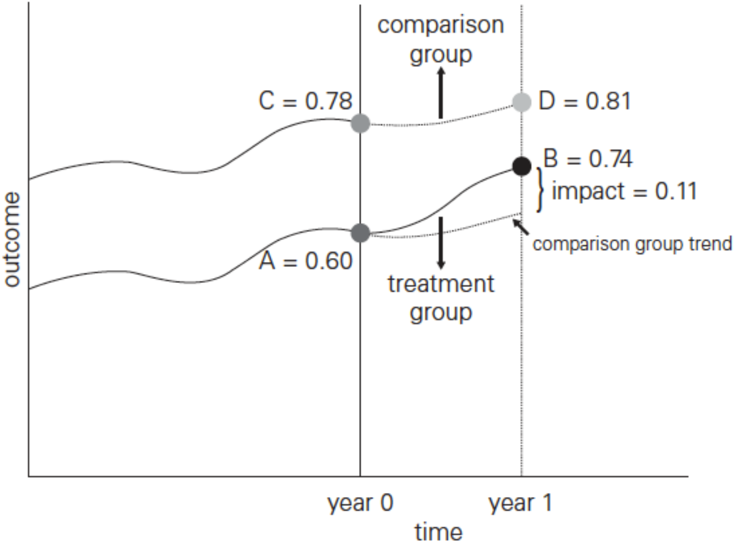
- $i = 0$ is control group, $i = 1$ is treatment.
- $t = 0$ is pre-period, $t = 1$ is post-period.
- One 'difference' estimate of causal effect is: $\mu_{11} — \mu_{10}$ (pre-post in treated)
- Differences-in-Differences estimate of causal effect is: $(\mu_{11} - \mu_{10}) - (\mu_{01} - \mu_{00})$

	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.

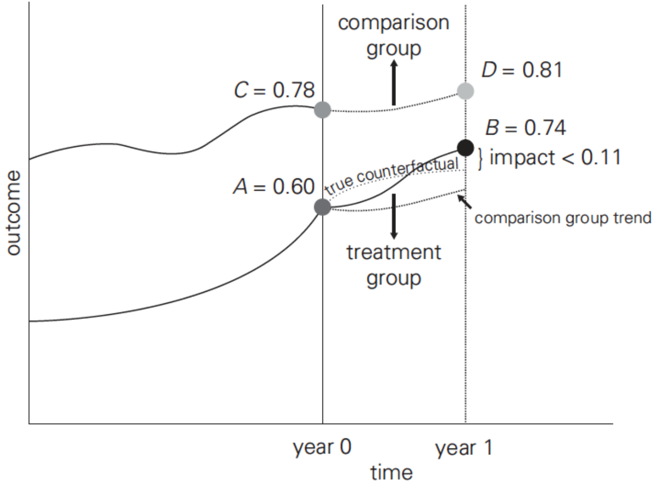
Visual Intuition of (good) DD



Gertler (2011)

Key assumption: parallel trends

- Non-parallel pre-intervention trends decrease study credibility.



Gertler (2011)

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

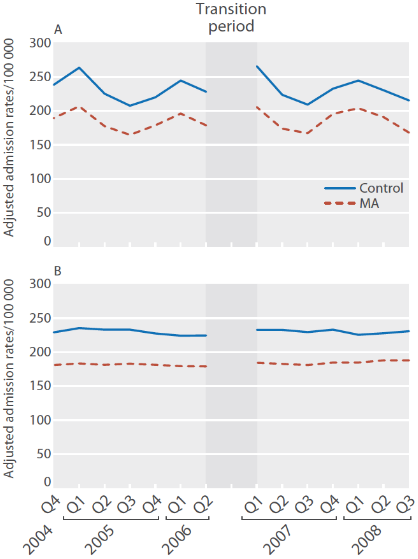
Danny McCormick,[1] Amresh D Hanchate,[2,3] Karen E Lasser,[3] Meredith G Manze,[3] Mengyun Lin,[3] Chieh Chu,[3] Nancy R Kressin[2,3]

- 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 [36]

Evaluating pre-intervention trends

- Strong visual evidence that pre-intervention trends similar in treated and control groups.
- Adds credibility to assumption that post-intervention trends would have been similar in the absence of the intervention.



- Comparison of pre-intervention covariates:

Table 1 | Characteristics of patients admitted to hospital for ambulatory care sensitive conditions in MA and control states (NY, NJ, PA) before (1 October 2004–30 June 2006) and after (1 January 2008–30 September 2009) healthcare reform. Figures are numbers (percentages) unless stated otherwise

	Massachusetts		Control states		
	Before reform (n=50 293)	After reform (n=52 248)	Before reform (n=393 900)	After reform (n=397 483)	P value*
Age (years):					
Mean (SD)	49.2 (11.9)	49.3 (11.9)	49.2 (11.5)	49.1 (11.6)	—
18–29	9.2	9.5	8.2	8.9	
30–39	10.8	10.2	11.4	10.7	
40–44	10.1	9.4	10.4	9.6	
45–49	13.1	12.9	13.7	13.8	0.64
50–54	15.6	16.7	16.1	17.1	
55–59	19.2	18.4	18.9	18.6	
60–64	22.1	23.0	21.4	21.3	

- Results:

Table 2 | Changes in rates of preventable hospital admissions* per 100 000 residents/year in Massachusetts and control states (NY, NJ, PA) before (1 October 2004–30 June 2006) and after (1 January 2008–30 September 2009) healthcare reform

	Massachusetts			Control states			Differences in differences estimates	
ACSC measure†	Before	After	% change	Before	After	% change	Unadjusted % change	Adjusted % change (95% CI)‡
Overall composite	745	730	−2.1	945	912	−3.5	1.4	1.2 (−1.6 to 4.1)
Acute composite	300	279	−7.0	308	292	−5.1	−1.9	−2.0 (−5.2 to 1.3)
Chronic composite	445	451	1.3	637	620	−2.7	4.0	3.7 (−0.04 to 7.6)

*Adjusted for age and sex with method of direct standardization.

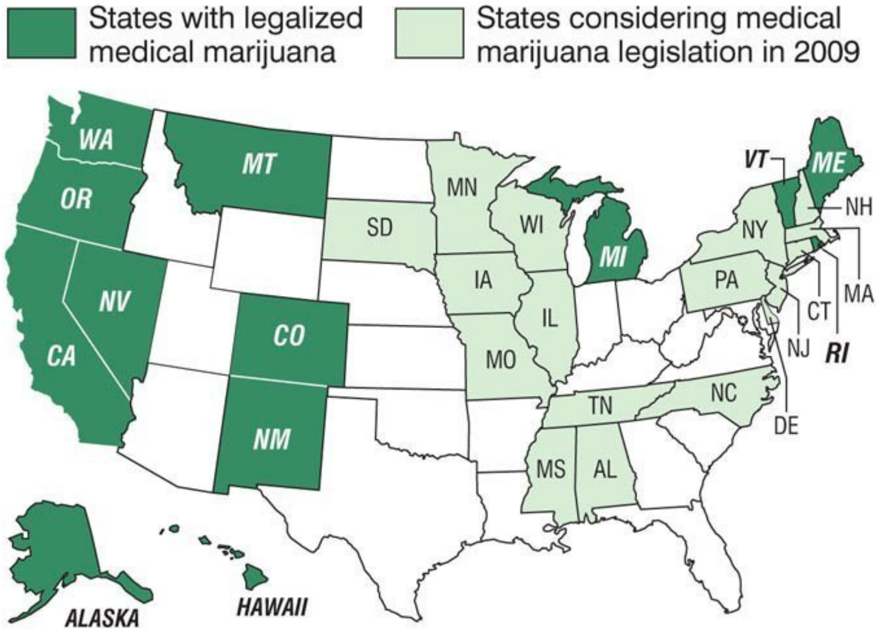
Clarity about the question really matters

Adolescent Marijuana Use from 2002 to 2008: Higher in States with Medical Marijuana Laws, Cause Still Unclear

Ann Epidemiol 2011;21:714–716.

MELANIE M. WALL, PhD, ERNEST POH, MS, MAGDALENA CERDÁ, DrPH,
KATHERINE M. KEYES, PhD, SANDRO GALEA, MD, DrPH, AND DEBORAH S. HASIN, PhD

- Wall et al. were interested in understanding the effect of legalizing medical marijuana on adolescent marijuana use.
  - Compared adolescent marijuana use in states with a law and without a law in each year from 2002-2008.
  - "States with MML had higher average adolescent marijuana use, 8.68% (95% CI: 7.95–9.42) compared to states without MML, 6.94% (95% CI: 6.60–7.28%)."


States with legalized medical marijuana
States considering medical marijuana legislation in 2009
WA
OR
CA
NV
MT
CO
NM
SD
MN
IA
MO
WI
IL
MI
TN
MS
AL
NC
PA
NY
VT
ME
NH
MA
CT
RI
NJ
DE
ALASKA
HAWAII

What was the question again?

Wall et al. also used a more sophisticated approach:

- "We used random-effects regression analysis that accounted for a common linear time trend and a random state intercept to compare "the prevalence of marijuana use in the years prior to MML passage (data available for 8 states prior to MML) to that of:
    - 1) post-MML years in states that passed MML and
    - 2) all years for states that did not pass MML by 2011
- But is that the question we want to answer? From a legislator's perspective, don't we want to know what would happen to adolescent marijuana use if we legalized medical marijuana?

Exchangeability is a key assumption

- Wall et al. assume there are no unmeasured factors that might make states with laws different from states without laws.
- Can you think of any other ways in which California might be different from Arkansas that could be related to marijuana use?

Exchangeability is a key assumption

- One solution: why not compare California to itself, before and after it changed its law? This leads to "difference-in-differences"* formulation
- Comparing each state to itself before and after a policy change and comparing to control states means we can control for:
  - Fixed characteristics of states that do not change over time (e.g., social norms)
  - Common secular trends that affect marijuana use in all states (kids today...)

*see Meyer, 1995[37]; Angrist & Pischke, 2009, 2015[38, 39]

Same Data, Different Question, Different Answer.

Replication and re-estimation using difference-in-differences:

	Past month marijuana use rate (%)			
	β	95% CI	β	95% CI
After law passed	1.87	(1.5, 2.2)	-0.59	(-1.1, -0.1)
Constant	8.23	(7.9, 8.6)	8.62	(8.4, 8.8)
Year fixed effects		Yes		Yes
State fixed effects		No		Yes

N=306 for all models.

- By using each state as its own control, we find that marijuana use decreases, whereas Wall et al. find an increase.
- Highlights the importance of framing the question in causal terms (i.e., what happens after policy implementation?)

*Harper, Strumpf, Kaufman, 2012[40]

Research Impact

WEST COAST LEAF

Current Issue: Spring 2012

inches)

ISSN 1945-221X

"The Cannabis Newspaper of Record"

« QuantaCann on-site testing system makes revolutionary change in industry

YouTube skips first-place query for Obama: Legalize? »

Medical cannabis laws may decrease adolescents' use

By Paul Armentano, **NORML Deputy Director**

State laws that allow for the limited legal use of cannabis by quali
according to data published online in February 2012 by the journ
Montreal, Canada, obtained state-level estimates of marijuana use

NORML

Working to reform marijuana laws

Toront
Promotions &

Home | About NORML | About Marijuana | State Info | Legal Issues | Library | News R

Main » NORML Blog » SCIENCE » Breaking News: Medicinal Cannabis Laws Have No Discernable Adverse Impact On Adolescents' Pot Use

Breaking News: Medicinal Cannabis Laws Have No Discernable Adverse Impact On Adolescents' Pot Use

:tor

The Daily Chronic

Medical Marijuana Laws Have No Discernible Adverse Impact On Adolescents' Use, Study Says

Tweet 2 | Like | 33 people like this. Be the first of your friends.

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.
  - But still check "Table 1": how close does it look to a RCT?
  - 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.

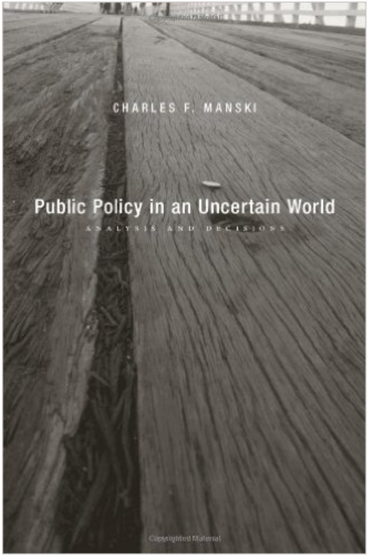
What are natural experiments good for?

1. To understand the effect of treatments *induced by policies* on outcomes, e.g., Policy $\rightarrow$ Treatment $\rightarrow$ 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 $\rightarrow$ Outcome:
    - Taxes, wages.
    - Environmental legislation.
    - Food policy.
    - Employment policy.
    - Civil rights legislation.

Glymour 2014 [41]

assumptions + data $\rightsquigarrow$ conclusions

"...the strength of the conclusions drawn in a study should be commensurate with the quality of the evidence. When researchers overreach, they not only give away their own credibility, they diminish public trust in science more generally." (Manski 2013[42])



Are natural experiments always more credible than regression adjustment?

- 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 treatment groups.
- Allows for simple, transparent analysis of average differences between groups.
- Allows us to rely on weaker assumptions.

Using qualitative information for stronger designs

- Inference in natural and quasi-experimental studies can be strengthened considerably in several ways.
- Additional design elements include:
  - Adding more pre-test outcome data.
  - Adding relevant pre-treatment covariates.
  - Replication in alternative populations.
  - Systematic removal of treatment units.
  - Multiple control groups.
  - Placebo outcomes.
- Qualitative information on why programs change is extremely valuable.
- Key idea is to push the data and rigorously test your assumptions.

Why is John Snow's work compelling?

- Good **qualitative** evidence of pre-treatment equivalence between groups:

"The mixing of the supply is of the most intimate kind... each Company supplies both rich and poor, both large houses and small; there is no difference either in the condition or occupation of the persons receiving the water of the different Companies (pp. 74–75)... [and this intermixing provided]... incontrovertible proof on one side or the other (p. 74)"

- Good **qualitative** evidence about the process of treatment assignment:

More than 300,000 individuals were "...divided into two groups without their choice, and, in most cases, without their knowledge; one group being supplied with water containing the sewage of London...the other group having water quite free from such impurity"

Snow [1855] (1965: 74-75), Freedman 1991 [43];Dunning, 2012[44]

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.

Back to basics: impact evaluation 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/stronger 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.

How can we capitalize on natural experiments?

- Take "as-if random" seriously in all study designs.
- Find 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.

Causal inference challenges for social epidemiology

- Bias: this should be reduced.
    - Propensity scores; studying exposure changes; finding instruments; bias modeling can all help.
- Specificity: this should be increased.
    - More precise descriptions of interventions and populations; transportability
- Imagination: this should be encouraged.
    - Qualitative work; diversify researchers; improve social theory; accept/model uncertainty.
- Impact evaluation incorporates all 3 of these elements!

Glymour and Rudolph, 2016[45]

Postscript: why this still matters.

Eliminating Racial Disparities in Colorectal Cancer in the Real World: It Took a Village

Stephen S. Grubbs, *Delaware Cancer Consortium, Dover; and Helen F. Graham Cancer Center, Newark, DE*
Blase N. Polite, *The University of Chicago, Chicago, IL*
John Carney Jr, *US House of Representatives, Washington, DC*
William Bowser, *Delaware Cancer Consortium, Dover, DE*
Jill Rogers, *Delaware Division of Public Health, Dover, DE*
Nora Katurakes, *Delaware Cancer Consortium, Dover; and Helen F. Graham Cancer Center, Newark, DE*
Paula Hess, *Delaware Cancer Consortium, Dover, DE*
Electra D. Paskett, *College of Medicine and Comprehensive Cancer Center, Ohio State University, Columbus, OH*

- "In this brief report, we demonstrate what can happen when the entire health care community of a state is mobilized toward a goal: eliminating health disparities in CRC."

J Clin Oncol 2013 [46]. Note the impact factor of JCO in 2015 was 18.5.

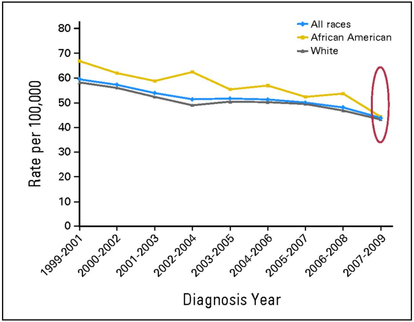
What happened?

- What is the intervention?
- What sort of design would you propose to answer the question?

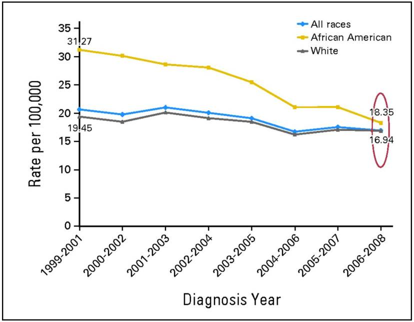
Did it take a village?

- Where is the evidence?
- Recall the "intervention" was implemented in 2003.

Trends in mortality rates by race



Trends in incidence rates by race



J Clin Oncol 2013 [46].

Discussion

"For all of the discussion about health care disparities, it sometimes seems that it has been so extensively documented that we have become numb to its implications or decided that it is too complex to fix. That there are complexities and nuances we do not deny, but the State of Delaware has shown us that if we have the will, there is a way."

"Delaware created a comprehensive statewide CRC screening program that included coverage for screening and treatment, patient navigation for screening and care coordination, and case management. By doing these common-sense things, we accomplished the following with respect to CRC health disparities from 2002 to 2009: elimination of screening disparities, equalization of incidence rates, reduction in the percentage of African Americans with regional and distant disease from 79% to 40%, and most importantly a near elimination of mortality differences."

More on this at http://samharper.org/new-blog/2017/3/6/did-it-really-take-a-village

Additional Resources (among many others)

- Published yesterday:

Accepted Manuscript

Quasi-experimental study designs series – Paper 7: assessing the assumptions

Till Bärnighausen, Catherine Oldenburg, Peter Tugwell, Christian Bommer, Cara Ebert, Mauricio Barreto, Eric Djimeu, Noah Haber, Hugh Waddington, Peter Rockers, Barbara Sianesi, Jacob Bor, Günther Fink, Jeffrey Valentine, Jeffrey Tanner, Tom Stanley, Eduardo Sierra, Eric Tchetgen Tchetgen, Rifat Atun, Sebastian Vollmer

PII: S0895-4356(17)30298-6

DOI: 10.1016/j.jclinepi.2017.02.017

Reference: JCE 9356

To appear in: *Journal of Clinical Epidemiology*

Quasi-experimental designs are gaining popularity in epidemiology and health systems research – in particular for the evaluation of healthcare practice, programs and policy – because they allow strong causal inferences without randomized controlled experiments. We describe the concepts underlying five important quasi-experimental designs: Instrumental Variables, Regression Discontinuity, Interrupted Time Series, Fixed Effects, and Difference-in-Differences designs. We illustrate each of the designs with an example from health research. We then describe the assumptions required for each of the designs to ensure valid causal inference and discuss the tests available to examine the assumptions.

Additional Resources (among many others)

- Expand your toolbox!

References I

[1] Miguel A Hernán, "Does water kill? A call for less casual causal inferences", *Ann Epidemiol*, 26(10), 2016, pp. 674–680.

[2] J S Kaufman and R S Cooper, "Seeking causal explanations in social epidemiology", *Am J Epidemiol*, 150(2), 1999, pp. 113–20.

[3] Tyler J VanderWeele and Whitney R Robinson, "On the causal interpretation of race in regressions adjusting for confounding and mediating variables", *Epidemiology*, 25(4), 2014, pp. 473–84.

[4] M Maria Glymour and Donna Spiegelman, "Evaluating Public Health Interventions: 5. Causal Inference in Public Health Research-Do Sex, Race, and Biological Factors Cause Health Outcomes?", *Am J Public Health*, 107(1), 2017, pp. 81–85.

[5] Jay S Kaufman, "There is no virtue in vagueness: Comment on: Causal Identification: A Charge of Epidemiology in Danger of Marginalization by Sharon Schwartz, Nicolle M. Gatto, and Ulka B. Campbell", *Ann Epidemiol*, 26(10), 2016, pp. 683–684.

[6] Commission on Social Determinants of Health, *Closing the gap in a generation: health equity through action on the social determinants of health: final report of the commission on social determinants of health*, Geneva: World Health Organization, 2008.

[7] Lisa F. Berkman, et al., *Social epidemiology*, Oxford University Press, second edition edn., 2016.

[8] Arijit Nandi and Sam Harper, "How consequential is social epidemiology? A review of recent evidence", *Current epidemiology reports*, 2(1), 2015, pp. 61–70.

[9] Mark Petticrew, et al., "Evidence for public health policy on inequalities: 1: the reality according to policymakers", *J Epidemiol Community Health*, 58(10), 2004, pp. 811–6.

[10] C Bambra, et al., "A labour of Sisyphus? Public policy and health inequalities research from the Black and Acheson Reports to the Marmot Review", *J Epidemiol Community Health*, 65(5), 2011, pp. 399–406.

[11] Paul J Gertler, et al., *Impact evaluation in practice*, World Bank Publications, 2016.

[12] William D Savedoff, et al., *When will we ever learn?: Improving lives through impact evaluation*, Center for Global Development, 2006.

References II

[13] Martin Ravallion, "Evaluating anti-poverty programs", *Handbook of development economics*, 4, 2007, pp. 3787–3846.

[14] Arijit Nandi, et al., "The effect of an affordable daycare program on health and economic well-being in Rajasthan, India: protocol for a cluster-randomized impact evaluation study", *BMC Public Health*, 16, 2016, p. 490.

[15] Malcolm R Macleod, et al., "Biomedical research: increasing value, reducing waste", *Lancet*, 383(9912), 2014, pp. 101–4.

[16] Rachel Glennerster and Kudzai Takavarasha, *Running randomized evaluations: a practical guide*, Princeton University Press, Princeton, NJ, 2013.

[17] Dean Karlan and Jonathan Zinman, "Microcredit in theory and practice: using randomized credit scoring for impact evaluation", *Science*, 332(6035), 2011, pp. 1278–84.

[18] Florencia Devoto, et al., "Happiness on tap: piped water adoption in urban Morocco", *American Economic Journal: Economic Policy*, 4(4), 2012, pp. 68–99.

[19] Michael Hallsworth, et al., *The behavioralist as tax collector: Using natural field experiments to enhance tax compliance*, Tech. rep., National Bureau of Economic Research, 2014, URL http://www.nber.org/papers/w20007.

[20] Adam Sacarny, et al., "Medicare Letters To Curb Overprescribing Of Controlled Substances Had No Detectable Effect On Providers", *Health Aff (Millwood)*, 35(3), 2016, pp. 471–9.

[21] Abhijit Banerjee, et al., "Development economics. A multifaceted program causes lasting progress for the very poor: evidence from six countries", *Science*, 348(6236), 2015, p. 1260799.

[22] Amm Quamruzzaman, et al., "Are tuition-free primary education policies associated with lower infant and neonatal mortality in low- and middle-income countries?", *Soc Sci Med*, 120, 2014, pp. 153–9.

[23] Jonathan Klick and Alexander Tabarrok, "Using terror alert levels to estimate the effect of police on crime", *The Journal of Law and Economics*, 48(1), 2005, pp. 267–279.

[24] 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.

References III

[25] 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.

[26] Joshua D Angrist, et al., “Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size”, *American Economic Review*, 88(3), 1998, pp. 450–77.

[27] Martin Ravallion and Quentin Wodon, “Does Child Labour Displace Schooling? Evidence on Behavioural Responses to an Enrollment Subsidy”, *The Economic Journal*, 110(462), 2000, pp. C158–C175.

[28] Sukumar Vellakkal, et al., “Food Price Spikes Are Associated with Increased Malnutrition among Children in Andhra Pradesh, India”, *J Nutr*, 145(8), 2015, pp. 1942–9.

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

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

[31] Ellen Moscoe, et al., “Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice”, *J Clin Epidemiol*, 68(2), 2015, pp. 122–33.

[32] 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.

[33] Leah M Smith, et al., “Strategies for evaluating the assumptions of the regression discontinuity design: a case study using a human papillomavirus vaccination programme”, *Int J Epidemiol*, 2016.

[34] 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.

[35] Anthony A Braga, et al., “Problem-oriented policing, deterrence, and youth violence: An evaluation of Boston's Operation Ceasefire”, *Journal of Research in Crime and Delinquency*, 38(3), 2001, pp. 195–225.

References IV

[36] 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.

[37] Bruce D Meyer, "Natural and Quasi-Experiments in Economics", *Journal of Business & Economic Statistics*, 13(2), 1995, pp. 151–161.

[38] Joshua David Angrist and Jörn-Steffen Pischke, *Mostly harmless econometrics: an empiricist's companion*, Princeton Univ Press, 2008.

[39] Joshua David Angrist and Jörn-Steffen Pischke, *Mastering 'metrics: the path from cause to effect*, Princeton University Press, 2015.

[40] Sam Harper, et al., "Do medical marijuana laws increase marijuana use? Replication study and extension", *Ann Epidemiol*, 22(3), 2012, pp. 207–12.

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

[42] Charles F Manski, *Public policy in an uncertain world: analysis and decisions*, Harvard University Press, 2013.

[43] David A Freedman, "Statistical models and shoe leather", *Sociological methodology*, 21(2), 1991, pp. 291–313.

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

[45] M Maria Glymour and Kara E Rudolph, "Causal inference challenges in social epidemiology: Bias, specificity, and imagination", *Soc Sci Med*, 166, 2016, pp. 258–65.

[46] Stephen S Grubbs, et al., "Eliminating racial disparities in colorectal cancer in the real world: it took a village", *J Clin Oncol*, 31(16), 2013, pp. 1928–30.

Ever tried.
Ever failed.
No matter.
Try again.
Fail again.
Fail better.

Samuel Beckett