**Title**:
Alternative Causal Inference Methods in Population Health Research: Evaluating Tradeoffs and Triangulating Evidence

**Authors**:
Ellicott C. Matthay
Erin Hagan
Laura Gottlieb
David Vlahov
Nancy Adler
M. Maria Glymour

## Introduction

Quantitative population health researchers are drawn from diverse disciplines, such as epidemiology, sociology, and economics. Investigators in these fields employ diverse terminologies and methodologies, even when addressing identical research questions. Lack of shared language and understanding has inhibited mutually beneficial interdisciplinary dialogue and collaboration. Moreover, preferences for the methodologies in one's home discipline have led to within-field reliance on only a subset of promising causal inference tools. These divides have motivated comments and efforts to bridge across disciplines (Krieger, 2000; Lynch, 2006; Abrams, 2006; Gunasekara et al., 2008; Kindig, 2007; Craig et al., 2012).

Many questions of central interest to population health researchers involve drawing causal inferences in the absence of conventional randomized controlled trials (RCTs). In this paper, we review and contrast common methodological strategies used by population health scientists to approach causal inference using nonexperimental designs. Drawing on examples from the literature on educational attainment, we contrast approaches that depend on accounting for factors that influence both the treatment and outcome (which we refer to as "covariate-control" methods) against approaches that leverage arguably random sources of variation in treatment, such as quasi-experiments or policy changes. We provide simplified summaries that aim to highlight strengths, weaknesses, and points of greatest contrast. Because inconsistent terminology is a persistent challenge for interdisciplinary research, we provide informal definitions for how we use important terms in this paper in **Boxes 1-3** (for formal definitions, see (Angrist and Pischke, 2008; Pearl, 2000; Rothman et al., 2008; Shadish et al., 2002)).

Choosing a method entails tradeoffs between statistical power, internal validity, measurement quality, and generalizability. Therefore, neither covariate-control nor quasi-experimental approaches will be preferable for all substantive questions. The assumptions for covariate-control methods differ markedly from those required in quasi-experimental designs, but both depend on unverifiable assumptions. Thus, convincing arguments need to triangulate evidence from different approaches, and the most useful research design is likely to evolve as evidence on a particular research question accumulates. Clarifying these tradeoffs between approaches will help population health researchers to strengthen the evidence in the population health sciences by selecting the most appropriate method for any given research question and better integrating evidence from heterogeneous designs.

## Defining the research question

Consider the question of whether college completion affects adult mortality. The research question of interest is causal and much more difficult to answer than a research question that is merely predictive or documenting an association. We define causal effects by contrasting potential outcomes associated with specific treatments (**Box 1**). For any individual, we would like to know whether her survival if she completes college is longer than her survival would be if she stopped her education at the end of high school. In practice, one of these survival outcomes is known and the other is unknown. The challenge of causal inference is to approximate this unknown potential outcome, using observed data on a sample of people. We

observe the actual survival outcomes for some individuals who completed college and others who did not. We would like to know what would have happened if we could roll back the clock and observe the same individuals, but under the scenario in which individuals with a high school education were instead college graduates and vice versa. Simply comparing survival of individuals with high school degrees to those with college degrees is unlikely to be successful in estimating the effect of education because those with differing levels of education also likely differ on other characteristics that will influence survival.

### Box 1. Terms Describing Variables in Causal Inference

**Causal model:** A tool, most often a system of equations or a diagram, used to describe background assumptions about hypothesized or known causal relationships among variables that are relevant to a particular causal question.
**Treatment** or **exposure** or **independent variable**: The explanatory variable of interest in a study. These terms are often used synonymously even for exposures that are not medical "treatments", such as social resources or environmental exposures.
**Outcome** or **dependent variable**: The causal effect of interest is the impact of the treatment or exposure on this variable.
**Potential outcome**: The outcome that an individual (or other unit of analysis, such as family or neighborhood) would experience if his/her treatment takes any particular value. Each individual is conceptualized as having a potential outcome for each possible treatment value. Potential outcomes are sometimes referred to as counterfactual outcomes.
**Exogenous versus endogenous variables**: These terms are common in economics, where a variable is described as exogenous if its values are not determined by other variables in the system under consideration. The variable is called endogenous if it is influenced by other variables in the system. If a confounder influences treatment variable and outcome, this implies the treatment is endogenous.
**Instrument** or **instrumental variable**: An external factor that induces treatment differences and has no other reason to be associated with the outcome. An instrument—for example, random assignment to treatment—can be used to estimate the effect of treatment on the outcome.

Randomized trials are typically conceptualized as an optimal approach to estimating causal effects, because random assignment helps to balance measured and unmeasured differences between treated and untreated groups that could otherwise lead to bias. However, many questions of interest to population health researchers involve situations where randomization is not ethical or feasible. Approaches to estimating causal effects in the absence of randomization can be broadly categorized into two groups. In this paper, we refer to these as observational and quasi-experimental (**Box 2**). In observational studies, researchers compare outcomes for people observed to have differing treatments and use covariate-control to account for imbalances in characteristics between treatment groups. In quasi-experimental studies, researchers leverage an external or "exogenous" (**Box 1**) source of variation in the treatment received—often a change in a program, policy, or other accident of time and space—that influences treatment but is not likely to be otherwise associated with outcomes. The assumptions required to estimate causal effects for each approach are distinct and we elaborate on these assumptions in the sections that follow.

### Box 2. Terminology for Study Designs and Causal Effects

**Observational study**: A study in which effects of a treatment are estimated by comparing outcomes of treated to untreated individuals. Treatment in these settings may be determined by the individual's own preferences, behaviors, or other naturally occurring influences.
**Quasi-experimental study**: A study in which effects of a treatment are estimated by leveraging the influence of an external factor that induces differences in treatments between individuals who are otherwise similar.
**Average treatment effect (ATE)**: The difference in average outcome if everyone in the population were treated compared to if nobody in the population were treated.
**Local average treatment effect (LATE)**: The average treatment effect among those whose treatment status is changed by the instrument (i.e., the effect among compliers).

## Observational approaches to estimating causal effects

Observational study designs are conducted in settings in which the treatment and outcome are each determined by a potentially large set of factors, and variation in the exposure is not due to the interventions of the researcher. Observational studies are particularly common in epidemiology. In cohort studies, the archetypical observational study design in epidemiology (Rothman et al., 2008), exposure is characterized in a group of individuals who are then followed to assess subsequent health outcomes. The National Longitudinal Study of Youth is an example of a typical cohort study that might be used to evaluate the effects of college completion on mortality.

The analytic strategy used to estimate the causal effect of interest is to ascertain, measure, and appropriately adjust for a "sufficient set" of variables to control confounding (**Box 3**). Modern frameworks define confounding as arising from shared causes of treatment and outcome; such factors can create non-causal influences linking treatment and outcome. Sufficient sets are often determined from substantive knowledge, prior research, or expert judgement. In the modern epidemiologic approach to causal inference, causal diagrams have emerged as popular tools to select sufficient sets of covariates (Pearl, 2000; van der Laan and Rose, 2011). Causal diagrams are causal models (**Box 1**) that visually represent background knowledge and assumptions about the causal structures linking variables. They are similar to the conceptual models used in many disciplines but are drawn and interpreted with formal mathematics-based rules that provide a rigorous method for determining sufficient sets. Usually, there is some uncertainty about the correct diagram, and several diagrams are considered plausible. Ideally a set of covariates is available that would be sufficient to control confounding under any of the causal diagrams. If the variables in a sufficient set are correct, have been measured in the available data, and can be appropriately controlled—the key assumption of observational approaches—then valid effect estimates can be delivered.

Once a sufficient set of covariates has been selected, several options can be used to account for these covariates. Researchers typically adopt a modeling approach. Because confounding arises from variables that influence both exposure and outcome, strategies to reduce confounding may focus on breaking the association of the confounders with the outcome (e.g., regression

adjustment), the association of the confounders with the exposure (e.g., matching, adjustment, or weighting based on propensity scores), or both (i.e., doubly robust methods). These methods all effectively reduce confounding bias by making comparisons within subgroups or pseudo populations that have balanced covariates, such that the covariates cannot bias the treatment-outcome association.

Both covariate-control approaches can be incorporated into numerous statistical models, such as generalized linear regressions or time-to-event (survival) models. The choice of a particular statistical model is driven by concerns about the parameter of interest, bias-efficiency tradeoffs, and convenience. For example, the investigator might use a regression to model the risk of mortality by age 60 as a function of whether the individual completed college, as well as baseline individual, psychosocial, interpersonal, and community covariates such as gender, conscientiousness, marital status, and access to care. The parameter most commonly estimated by covariate-control analyses is the average treatment effect (ATE; **Box 2**). That is, the average outcome if everyone in the population were exposure versus if no one were exposed—for example, the difference average survival times if everyone in the population completed college versus if nobody completed college. The ATE is commonly of interest in population health.

Panel fixed effects can also be considered a type of covariate-control approach, particularly when a program or policy is itself the treatment. In this approach, treatments and outcomes are typically measured on the same participants or places over time. Binary variables (also known as indicator variables) representing each place are used to control for features of participants/places that do not change over the study period (e.g. genes). Indicator variables representing each time period are used to control for features of time that are common across places (e.g. a nationwide recession). For example, the investigator might leverage variation in the timing and location of compulsory schooling law (CSL) implementation across states, modeling mortality rates across states and years as a function of state indicators, time indicators, and a variable representing CSL implementation (Fletcher, 2015). This approach relies on the same assumption of adequately identifying, measuring, and adjusting for all confounders, but indicator variables serve to control for time-invariant aspects of units and unit-invariant aspects of time, so the remaining confounders of concern are those that vary in time and are specific to each place.

## Quasi-experimental approaches to estimating causal effects

Quasi-experiments yield variation in the treatment other than the processes typically involved in determining treatment. They therefore create differences in treatments between individuals who are presumably otherwise similar that can be leveraged to estimate the causal effect of the treatment on the outcome. The key assumption for this category of approaches is that the external source of variation in treatment is unrelated to the potential health outcomes of the individuals in the study.

Sources of quasi-experimentation include lotteries (sometimes used to assign housing vouchers (Sanbonmatsu et al., 2011) or other resources (Eisenberg and Rowe, 2009; Pallais, 2009) when

there is not enough for all eligible individuals), arbitrarily assigned judges (who have different propensities for leniency (Roach and Schanzenbach, 2015)) or clinicians (who have different preferences for treatment modalities (Brookhart and Schneeweiss, 2007)), month or quarter of birth (which influences years of schooling (Acemoglu and Angrist, 1999)), or biological chance, such as the sex of a child (which influences chances parents will wish to conceive another child (Angrist and Evans, 1998)). Quasi-experiments also take the form of arbitrary discontinuities or determinants of treatment that are not associated with other determinants treatment—for example, an arbitrary cutoff for social program eligibility or arbitrary variation across states and time in the implementation of a policy.

Study designs such as instrumental variables (IV), regression discontinuity (RD), and differences-in-differences (DiD) (each described below) are often discussed separately, but all rely quasi-experiments. This collection of techniques is common in economics and other social and behavioral sciences (Angrist and Pischke, 2008; Shadish et al., 2002). For these techniques, it is useful to distinguish between research questions about the health effects of a specific policy and research questions about the health effects of an exposure, treatment, or resource delivered by a policy. Both are usually of interest. IV analyses deliver estimates of the effects of exposure, which may be of interest for two reasons. First, once the effect of exposure is known, policy alternatives that influence treatment can be compared to one another and some may be preferable for other reasons (e.g., political feasibility). Second, the overall effect of the policy is partly dependent on how many people were influenced by the policy, i.e., how many people became eligible because of the policy change, or how the policy was enforced. These factors may change as evidence accrues, and knowing the effects of the exposure is more likely to be useful to predict health impacts of future policy changes. Quasi-experimental designs can be deployed to evaluate effects of policies themselves, but they are also powerful designs for evaluating the effects of treatments determined by those policies. In the latter case, each design (IV, RD, DiD) can be conceptualized and executed as an IV approach, in which case the quasi-experiment is considered an IV that can be used to evaluate the effect of the treatment they influence on health outcomes (Angrist and Pischke, 2008).

IV analyses control confounding by leveraging a source of variation in the treatment (the instrument) that is distinct from other determinants of treatment. A typical IV analysis requires the assumptions of relevance (the instrument must affect the treatment), exclusion (the instrument only affects the outcome through the treatment), and exchangeability (instrument does not share unmeasured common causes with the outcome) (**Box 3**). Some assumptions (e.g. relevance) can be tested, but others (e.g. exclusion) cannot be tested and must be judged substantively. The treatment itself may have numerous determinants, but when these assumptions are met, the variation in treatment that is predicted by the instrument will be independent from these other determinants. IV analyses quantify the effect of this instrument-predictable variation in exposure on the outcome. If the instrument is randomization, as in an RCT, this is the effect of the treatment among compliers. This is the core of IV analysis.

In RD methods, there is an arbitrary discontinuity in the probability of being treated depending on the value of a third variable (such as class size, date or hour of birth, age, or income). It is

assumed that individuals immediately above and below that discontinuity have equivalent potential health outcomes, but stark differences in treatment probability, and that the shape of the relationship between the third variable and treatment probability is known. Under these assumptions, this third variable can be considered an IV to evaluate the effect of treatment on health outcomes. Such an analysis is termed "fuzzy" RD and analyzed in the same manner as a traditional IV. For example, Goodman et al (Goodman et al., 2015) noted that admission to Georgia's State University System was granted only to applicants with math SAT scores above 400, creating a discontinuity at this score threshold in probability of beginning college at a 4-year institution. They took advantage of this discontinuity to estimate the effect of starting college at a 4-year institution (the treatment) on chances of college completion (the outcome). If the research question were instead about the effect of Georgia State's SAT score admission policy itself as the treatment, the investigator could use a regression approach to directly compare college completion for those just meeting and just missing the SAT score threshold. Termed "sharp" RD, this approach effectively reduces to covariate-control and assumes that no unmeasured factors affecting college completion coincide with the SAT score threshold and that functional form of the model relating SAT scores to mortality is known.

DiD methods combine an RD with one or more comparison groups, which can account for other sources of variation at the discontinuity. This approach is especially valuable when the date of change in a policy affecting treatment (e.g., mandatory schooling law changes) is used as a discontinuity. For example, in 1918, Mississippi implemented a law requiring children to attend a minimum number of years of schooling. We might hope to use this discontinuity to estimate the health effects of extra schooling. However, World War I or the influenza pandemic of 1918 might have altered long-term outcomes for those cohorts in ways completely unrelated to the additional schooling. In a DiD design, we might include a state that did not change its schooling law in those years to control for these historical events. The key assumption of DiD is that, conditional on measured covariates, if the state that changed its mandatory schooling policy had not done so, the trends in outcomes would be parallel for individuals in that state as in states that did not change their policy. This amounts to an assumption of no confounders that vary at the same time as the mandatory schooling policy. DiD methods are commonly analyzed as traditional IVs where the interaction of treatment and time variables serves as an instrument. However, if the research question is about the policy as a treatment itself, DiD reduces to an observational covariate-control approach analogous to panel fixed effects.

The strength of an analysis drawing on a valid quasi-experiment is that it may deliver accurate effect estimates even if there are unmeasured confounders of the treatment-outcome association (Duncan, 2008; Moffitt, 2005). However, in many cases, the assumptions for quasi-experimental approaches are assumed to hold only after conditioning on a set of covariates. When this is needed is determined by the assumed causal model (i.e., background assumptions about hypothesized or known causal relationships between variables; **Box 1**), which can be expressed as equations or with graphical models.

Interpreting IV estimates—whether from a discontinuity, a difference-in-difference, or another exogenous source of variation in treatment—requires some additional assumptions. If one

assumes that the effect of treatment on outcome is identical for everyone in the population, then the IV estimates the ATE. This rarely seems likely however. Most IV analyses instead adopt the assumption of monotonicity: that the IV does not have opposite effects on the treatment for any two people in the population, i.e., if the policy increases treatment for some people, it must not decrease treatment for anyone (Pearl, 2000). Under this assumption, the parameter estimated by an IV approach is the local average treatment effect (LATE; **Box 2**). That is, the effect among those whose treatment is affected by the instrument—for example the effect of attending a four-year college on people who would attend a four-year college if and only if they scored above the SAT threshold. The LATE is generally estimated using two-stage least squares (2SLS) regression. The choice of parameter has important implications for generalizability which we discuss in the next section.

## Box 3. Types of Bias and Assumptions for Causal Inference

**Confounding** or **omitted variable bias** or **bias from selection into treatment:** A bias that occurs when the association between treatment and outcome is partially attributable to the influence of a third factor that affects both the treatment and the outcome. This bias is the key problem posed by lack of randomization. It is often referred to as omitted variables bias because it is a problem when the common cause is omitted from a regression model. Selection bias in this context specifically refers to selection into treatment and is distinct from biases due to selection into the study sample (which is the phenomenon typically referred to as selection bias in epidemiology).

**Information bias** or **measurement error**: A bias arising from a flaw in measuring the treatment, outcome, or covariates. This error may result in differential or non-differential accuracy of information between comparison groups.

**Reverse causation** or **simultaneity:** When the outcome causes the treatment, rather than the treatment causing the outcome.

**Exchangeability**, **ignorability, no confounding,** or **randomization assumptions:** The assumption that which treatment an individual receives is unrelated to her potential outcomes if given any particular treatment. This assumption is violated for example if people who are likely to have good outcomes regardless of treatment are more likely to actually be treated. In the context of instrumental variables analysis, exchangeability is the assumption that the instrument does not have shared causes with the outcome.

**Conditional exchangeability**, **conditional ignorability**, or **conditional randomization**: The assumption that exchangeability, ignorability, or randomization is fulfilled after controlling for a set of measured covariates. When this assumption is met, we say that the set of covariates—known as a **sufficient set**—fulfills the **backdoor criterion** with respect to the treatment and outcome. **Relevance**: In the context of instrumental variables, the assumption that the instrument affects the treatment.

**Exclusion**: In the context of instrumental variables, the assumption that, conditional on measured covariates, the instrument only affects the outcome through the treatment.

**Monotonicity**: In the context of instrumental variables, the assumption that all those affected by the treatment are affected in the same direction

**Stable unit treatment value assumption (SUTVA)**: The assumption that there is only one unique version of the treatment, and each unit's outcomes are unaffected by the treatment values of other units.

## Considerations and tradeoffs for all population health studies

Choosing among observational and quasi-experimental approaches entails tradeoffs (**Table 1**). Shadish, Cook, and Campbell's (SCC) causal inference framework (Shadish et al., 2002), which has been widely influential an a range of population health disciplines (Cook, 2018), is useful to consider which study design is preferable. SCC categorizes the types of validity necessary for studies to provide convincing causal inferences into four types:

1. Internal validity: the extent to which the estimated association in the study sample corresponds to a causal effect from treatment to outcome;
2. Statistical conclusion validity: appropriate use of statistical methods to assess the relationships between study variables;
3. Construct validity: the extent to which measured variables capture the concepts the investigator intends to assess with those measures; and
4. External validity: the extent to which study results can be generalized to other units, treatments, observations made on units, and settings of study conduct.

Under this framework, the study design and analysis are critical to internal validity. Correct statistical inference, measurement, and external validity, which are arguably insufficiently addressed in training and practice in many population health sciences, are also critical for accurate and precise interpretation and relevance of causal inferences.

### Internal validity

Internal validity requires some type of conditional exchangeability or randomization assumption (**Box 3**). The choice between observational and quasi-experimental approaches is often driven by which untestable assumptions to achieve exchangeability seem most plausible. Adequately accounting for confounders is particularly challenging for social determinants of health where causal pathways are complex, cyclical, and difficult to identify. It is clear that those who pursue college differ from those who do not on a wide variety of factors that can impact health outcomes. Thus, the primary limitation of standard observational approaches is the reliance on identifying, measuring, and correctly adjusting for a sufficient set of confounders. Observational study designs are particularly appealing when achieving this task seems feasible, or when observational approaches can make improvements in covariate-control over previous studies. Quasi-experimental strategies that can address unmeasured confounding factors are powerful tools for preventing threats to internal validity that may be particularly persistent in research on social determinants of health. The major weakness of quasi-experimental approaches is that finding an adequate quasi-experiment or valid instrument to answer the study question of interest can be challenging.

Confounding constitutes the core threat, but internal validity may also be threatened by imperfectly measured variables, regression model misspecification, reverse causation or simultaneity (**Box 3**), inadvertently controlling for factors that are influenced by exposure, or differential loss-to-follow-up, among others. For example, in a covariate-control regression model, if a continuous confounder with a linear relationship to the outcome is modeled as a binary variable with a threshold effect, the model will not fully account for that variable. In both observational and quasi-experimental approaches, design tools such as falsification tests or negative control exposures or outcomes can help to rule out alternative explanations and contribute to internal validity.

## Statistical conclusion validity

All causal inference approaches rely on appropriate statistical inference. This includes ruling out random error, having sufficient support in the data for the statistical estimate of the target causal quantity to be defined, meeting necessary assumptions of the statistical test or model (e.g. independent and identically distributed observations on units; no interference or spillover), accounting for multiple testing (e.g. through a Bonferroni correction), and correctly specifying the statistical model (e.g. the association between age and mortality is linear). The stable unit treatment value assumption (SUTVA; **Box 3**)—that each unit's outcomes are unaffected by the treatment values of other units—is assumed for the statistical validity of many analyses.

All approaches can be threatened by low statistical power, but power may be a particular challenge in quasi-experimental studies, because inferences are constrained to the fraction of the study population whose exposure is affected by the quasi-experiment. For example, if compulsory schooling laws only impact educational attainment for a fraction of the study population (a common occurrence in quasi-experiments), the sample size is effectively limited to that fraction. The result can be wide confidence intervals or under-powered studies. An observational approach to the same question could potentially leverage the entire study population.

## Construct validity

Construct validity concerns relate to whether study measurements capture the constructs they are intended to capture. Causal inferences will be invalid if observed effects are interpreted or attributed incorrectly. Many threats to construct validity could be described as information bias or measurement error (**Box 3**). Misunderstanding the "active" component of a program (e.g., college completion may improve health outcomes because of the college credential, the knowledge and skills gained through coursework, or the social network established) threatens construct validity. Failing to recognize that program participation had multiple consequences besides the intentionally delivered services (e.g. if college attendance is accompanied by job search support that substantially enhances subsequent earnings) is another threat to construct validity of particular relevance for population health research. Similar concerns relate to measurement error (e.g. if self-reports of educational attainment are affected by investigator expectations). When threats to construct validity are recognized, they can be addressed in

design or measurement innovations (e.g., incorporating multiple or objective measures of the outcome) or simply by tempering interpretation of the study's findings.

Greater construct validity can come at the expense of statistical power, because the highest quality measurements are often expensive and time-consuming to collect, and thus performed on smaller samples. This tradeoff is one that population health researchers must often grapple with as they seek to make valid causal inferences. Studies grounded in large administrative datasets benefit from greater statistical power but tend to have less detailed measurements while smaller studies can afford more and higher quality measurements. To the extent that observational or quasi-experimental approaches more commonly use one data type over another, they may bring different strengths. Important approaches to solving measurement quality problems for both designs include detailed measurements on subsamples of large data sets (Langa et al., 2005), large data initiatives (Sudlow et al., 2015) and, in the case of quasi-experiments, targeted data collection enrolling a smaller sample of individuals most affected by the quasi-experiment (Schneider and Harknett, 2018).

## External validity

External validity concerns relate to the populations and places to which study results can be generalized. Causal inferences about populations external to the one under study will be invalid if the causal relationship of interest is modified by participant characteristics, settings, the types of outcomes measured, or treatment variations. Researchers address generalizability based on a priori knowledge or theory guiding interpretation of results, delineating the target population to whom the results refer (e.g. with respect to sociodemographics or geography), and judging the extent to which the findings are relevant to settings beyond the ones studied. External validity concerns can also be addressed with design or analytic features such as oversampling of underrepresented groups, modeling causal interactions, or applying analytic methods of generalization such as transportability estimators (Pearl and Bareinboim, 2011).

Population representative, or at least diverse, data sources are necessary to understand how treatments influence both population average health outcomes and inequalities in health outcomes, the central issues in population health research. Many observational studies are well-suited to these goals, because they are frequently based on large, population-representative samples and estimate population average treatment effects. The diversity of participants these studies also supports the evaluation of differential effects across population subgroups, facilitating generalization of effect estimates to new populations with different compositions. For example, observational studies of education and health typically include both White and Black participants, and differential effects can be directly evaluated (Assari and Mistry, 2018; Cohen et al., 2013; Kaplan et al., 2008; Liu et al., 2015; Vable et al., 2018).

Generalization can be more challenging in quasi-experimental studies, because they typically deliver local average treatment effects (LATE), which apply only to the subset of participants whose treatment is affected by the quasi-experiment. Additionally, it can be challenging to find instruments that affect treatment for diverse population subgroups such that treatment effects can be estimated for each subgroup. For example, Lleras-Muney found evidence that

compulsory schooling laws were not historically enforced for Black children, and thus cannot be used to tell us about the effects of education on black populations, unless we are willing to assume that effects in White students can be generalized to Black students (Lleras-Muney, 2002).

However, the LATE can be an important population health parameter in some situations, such as when there is no possibility that everyone in a population would be treated. For example, when estimating the health effects of incarceration, it is most relevant to consider cases for which either incarceration or release is a reasonable sentence. Convicted murderers will always be incarcerated. Jaywalkers will not be incarcerated. Of interest are health effects for individuals with intermediate crimes, for whom reasonable people might disagree about a "just" sentence. In this case, the LATE delivered by quasi-experiments leveraging arbitrary differences in judicial leniency can be extremely informative in population health research.

## Table 1. Comparison of common approaches to nonexperimental causal inference for population health scientists studying the effects of treatments

| Feature | Observational | Quasi-experimental |
|---|---|---|
| Main strategies for estimating causal effects | Identify, measure, and control for a sufficient set of confounders through regression adjustment, propensity score methods, or some combination. | Identify and leverage a random or conditionally random source of variation in treatment through instrumental variables, regression discontinuity, differences-in-differences, or related approaches. |
| Key assumptions | Conditional exchangeability, including no unmeasured common causes of treatment and outcome. | Relevance, exclusion, exchangeability, and monotonicity or constant treatment effects. |
| Assessment of assumptions | Assumptions cannot be proven and are primarily evaluated based on background knowledge, negative controls, or testable implications of the hypothesized causal mechanisms. Measured covariates are often assumed to proxy for unmeasured covariates and used to inform sensitivity analyses. | The "relevance" assumption can be proven. Other assumptions cannot be proven and are primarily evaluated using background knowledge, falsification tests drawing on multiple IVs, mathematical constraints implied by the assumptions (bounding approaches), or testable implications of the hypothesized causal mechanisms. |
| Typical statistical analysis | Regression with covariate-control<br>Propensity score matching, adjustment, or weighting<br>Doubly robust analyses | Two-stage least-squares |
| Key methodological advantages | Analysis can incorporate entire populations, resulting in greater statistical power relative to studies of sub-populations affected by quasi-experiments.<br>Often based on diverse and representative samples that facilitate assessment of treatment effects across and within populations. | Study design and analytic approaches can circumvent bias from unmeasured confounders. |
| Key methodological challenges | Reliance on identifying, measuring, and adjusting for all confounders. | Valid instruments can be difficult to identify.<br>Analysis limited to the sub-population affected by quasi-experiment, resulting in reduced statistical power relative to total-population studies.<br>Treatment effects (LATE) only generalize to the subset of participants whose treatment is affected by the quasi-experiment. |

See Boxes 1-3 for definitions. For simplicity, the characterizations in this table generally refer to analytic methods for questions about the effect of receiving a treatment or resource, such as additional education, not the effect of a policy or program itself (e.g. providing vouchers to subsidize college completion). We present a simplified characterization of each approach to highlight key points of potential divergence.

## Discussion

All population health researchers strive to generate compelling evidence on causal effects in situations when randomization is not possible or not ethical. Investigators use a variety of approaches to address this challenge, broadly falling into the categories of observational and quasi-experimental designs. The approaches presented in this paper are distinct, but rarely in conflict. Each approach entails tradeoffs, and within each approach are further analytic decisions with their own tradeoffs. Untestable assumptions must be made to derive useful inferences with any approach. Which set of untestable assumptions is more appealing depends on the disciplinary traditions of the investigator and the problem and data at hand. The preferred approach also depends on the prior research: if all prior research depends on the same untestable assumptions, additional work that does not depend on those assumptions will be more valuable than work invoking identical assumptions as prior studies. In other words, alternative methods allow triangulation (Lawlor et al., 2016). Limitations from one study can be addressed by inferences from another; a variety of studies with diverse strengths and weaknesses will provide stronger evidence than any single study alone (Cordray, 1986, 1986; Duncan, 2008).

To our knowledge, there has been little systematic attention to categorizing the types of problems amenable to observational approaches, problems where quasi-experimental approaches are preferable, or problems for which neither will deliver particularly informative answers. Existing research comparing the performance of different analytic approaches relies primarily on "within-study comparisons". Such comparisons align randomized trial effect estimates against estimates obtained using observational or quasi-experimental methods applied to the trial's treatment group and an externally derived untreated population such as a population-representative survey (Wong and Steiner, 2018). These studies demonstrate that the performance of different approaches is highly context- and application-dependent. In some settings, no approach succeeds in replicating the experimental result, while in others, numerous quasi-experimental and observational approaches perform well (Pirog et al., 2009; Shadish, 2011). Reasons for this variability are not fully understood (Oliver et al., 2010). Although it has been suggested that regression discontinuity more reliably replicates experimental results than other observational and quasi-experimental approaches (Pirog et al., 2009; Shadish, 2011), it is unclear whether this approach actually performs better, or whether the situations in which such methods can be applied are more promising for validity regardless of the analytic approach. Moreover, such studies rarely consider applications to social determinants of health (examples of exceptions are (Gennetian et al., 2010; Handa and Maluccio, 2010)), have not utilized modern methods that provide rigorous procedures for covariate selection or that make fewer assumptions about shapes of relationships between variables, rarely consider external validity (exceptions exist, e.g. (Jaciw, 2016)), and by definition cannot address the types of questions that are not amenable to randomization. At this point, there are simply too few truly parallel comparisons of effect estimates for social determinants of health relying on divergent research designs to draw general conclusions.

Fortunately, the techniques and concepts of observational and quasi-experimental approaches described in this paper are not mutually exclusive. Tools from one approach may complement or strengthen the tools in another. For example, in quasi-experimental studies where some covariate-control is needed, the causal diagrams popular with epidemiologists could provide a closed form procedure for covariate selection (Pearl, 2000; van der Laan and Rose, 2011). It is often useful to have evidence from both approaches, and when the results align, findings are especially convincing. For example, good correspondence has been seen for observational results and randomized trials in clinical epidemiology (Anglemyer, 2014), as well as for a subset of observational and quasi-experimental studies of educational attainment and mortality (cites). New evidence arising from diverse approaches is more convincing than evidence relying on the same, unverifiable assumptions. Preference for research design approaches often aligns with disciplinary tradition. Cross-disciplinary exchange can enhance technical toolkits and avoid research duplication.

## Limitations

We present simplified characterizations of approaches to causal inference for non-randomized empirical studies on social determinants of health with the goal of fostering cross-disciplinary communication and enhanced use of the full spectrum of causal inference tools available to population health scientists. These are vast generalizations of rich methodologies that have substantial heterogeneity in implementation. We present the most common approaches, and highlight their main strengths and weaknesses. Other methods exist (e.g. synthetic control (Abadie et al., 2010; Doudchenko and Imbens, 2016) or comparative regression discontinuity (Tang et al., 2017)), and often build upon the core approaches we described.

## Conclusions

For research on social determinants of health, the challenge of causal inference in the absence of randomization is a so-called “wicked” problem. For these wicked problems, no single approach will likely provide conclusive evidence. Diversity of data sources and methodological approach is a boon, not a problem. However, understanding how these diverse approaches fit together is critical for forming a full picture of the current state of the evidence. This broader lens will help population health researchers across disciplines to select the approaches that will most effectively answer their research questions and to strengthen the evidence needed to improve population health and health equity outcomes.

# References


Abadie, A., Diamond, A., Hainmueller, J., 2010. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. Journal of the American Statistical Association 105, 493–505.

Abrams, D.B., 2006. Applying Transdisciplinary Research Strategies to Understanding and Eliminating Health Disparities. Health Education & Behavior 33, 515–531. https://doi.org/10.1177/1090198106287732

Acemoglu, D., Angrist, J., 1999. How Large are the Social Returns to Education? Evidence from Compulsory Schooling Laws (Working Paper No. 7444). National Bureau of Economic Research. https://doi.org/10.3386/w7444

Anglemyer, A., 2014. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials [WWW Document]. URL http://publichealthwell.ie/search-results/healthcare-outcomes-assessed-observational-study-designs-compared-those-assessed-rand?&content=resource&member=572160&catalogue=none&collection=none&tokens_complete=true (accessed 7.3.18).

Angrist, J.D., Evans, W.N., 1998. Children and their parents' labor supply: Evidence from exogenous variation in family size. American Economic Review 88, 450–477.

Angrist, J.D., Pischke, J.-S., 2008. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

Assari, S., Mistry, R., 2018. Educational Attainment and Smoking Status in a National Sample of American Adults; Evidence for the Blacks' Diminished Return. International Journal of Environmental Research and Public Health 15, 763. https://doi.org/10.3390/ijerph15040763

Brookhart, M.A., Schneeweiss, S., 2007. Preference-Based Instrumental Variable Methods for the Estimation of Treatment Effects: Assessing Validity and Interpreting Results. The International Journal of Biostatistics 3. https://doi.org/10.2202/1557-4679.1072

Cohen, A.K., Rehkopf, D.H., Deardorff, J., Abrams, B., 2013. Education and obesity at age 40 among American adults. Social Science & Medicine 78, 34–41. https://doi.org/10.1016/j.socscimed.2012.11.025

Cook, T.D., 2018. Twenty-six assumptions that have to be met if single random assignment experiments are to warrant "gold standard" status: A commentary on Deaton and Cartwright. Social Science & Medicine, Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue 210, 37–40. https://doi.org/10.1016/j.socscimed.2018.04.031

Cordray, D.S., 1986. Quasi-experimental analysis: A mixture of methods and judgment. New Directions for Program Evaluation 1986, 9–27. https://doi.org/10.1002/ev.1431

Craig, P., Cooper, C., Gunnell, D., Haw, S., Lawson, K., Macintyre, S., Ogilvie, D., Petticrew, M., Reeves, B., Sutton, M., Thompson, S., 2012. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. J Epidemiol Community Health 66, 1182–1186. https://doi.org/10.1136/jech-2011-200375

Doudchenko, N., Imbens, G., 2016. Balancing, regression, difference-in-differences, and synthetic control methods: a synthesis. NBER Working Paper No. 22791. arXiv. https://doi.org/1610.07748

Duncan, G. j., 2008. When to promote, and when to avoid, a population perspective. Demography 45, 763–784. https://doi.org/10.1353/dem.0.0031

Eisenberg, D., Rowe, B., 2009. The Effect of Smoking in Young Adulthood on Smoking Later in Life: Evidence based on the Vietnam Era Draft Lottery. Forum for Health Economics & Policy 12. https://doi.org/10.2202/1558-9544.1155

Fletcher, J.M., 2015. New evidence of the effects of education on health in the US: Compulsory schooling laws revisited. Social Science & Medicine, Special Issue: Educational Attainment and Adult Health: Contextualizing Causality 127, 101–107. https://doi.org/10.1016/j.socscimed.2014.09.052

Gennetian, L.A., Hill, H.D., London, A.S., Lopoo, L.M., 2010. Maternal employment and the health of low-income young children. Journal of health economics 29, 353–363. https://doi.org/10.1016/j.jhealeco.2010.02.007

Goodman, J., Hurwitz, M., Smith, J., 2015. College access, initial college choice and degree completion. National Bureau of Economic Research, Cambridge, MA.

Gunasekara, F.I., Carter, K., Blakely, T., 2008. Glossary for econometrics and epidemiology. Journal of Epidemiology & Community Health 62, 858–861. https://doi.org/10.1136/jech.2008.077461

Handa, S., Maluccio, J.A., 2010. Matching the Gold Standard: Comparing Experimental and Nonexperimental Evaluation Techniques for a Geographically Targeted Program. Economic Development and Cultural Change 58, 415–447. https://doi.org/10.1086/650421

Jaciw, A.P., 2016. Assessing the Accuracy of Generalized Inferences From Comparison Group Studies Using a Within-Study Comparison Approach: The Methodology. Eval Rev 40, 199–240. https://doi.org/10.1177/0193841X16664456

Kaplan, G., Ranjit, N., Burgard, S., 2008. Lifting Gates - Lengthening Lives: Did Civil Rights Policies Improve the Health of African-American Women in the 1960's and 1970's. Russell Sage.

Kindig, D.A., 2007. Understanding Population Health Terminology. The Milbank Quarterly 85, 139–161. https://doi.org/10.1111/j.1468-0009.2007.00479.x

Krieger, N., 2000. Epidemiology and Social Sciences: Towards a Critical Reengagement in the 21st Century. Epidemiol Rev 22, 155–163. https://doi.org/10.1093/oxfordjournals.epirev.a018014

Langa, K.M., Plassman, B.L., Wallace, R.B., Herzog, A.R., Heeringa, S.G., Ofstedal, M.B., Burke, J.R., Fisher, G.G., Fultz, N.H., Hurd, M.D., Potter, G.G., Rodgers, W.L., Steffens, D.C., Weir, D.R., Willis, R.J., 2005. The Aging, Demographics, and Memory Study: Study Design and Methods. NED 25, 181–191. https://doi.org/10.1159/000087448

Lawlor, D.A., Tilling, K., Davey Smith, G., 2016. Triangulation in aetiological epidemiology. Int J Epidemiol 45, 1866–1886. https://doi.org/10.1093/ije/dyw314

Liu, S.Y., Manly, J.J., Capistrant, B.D., Glymour, M.M., 2015. Historical Differences in School Term Length and Measured Blood Pressure: Contributions to Persistent Racial Disparities among US-Born Adults. PLOS ONE 10, e0129673. https://doi.org/10.1371/journal.pone.0129673

Lleras-Muney, A., 2002. Were Compulsory Attendance and Child Labor Laws Effective? An Analysis from 1915 to 1939. The Journal of Law and Economics 45, 401–435. https://doi.org/10.1086/340393

Lynch, J., 2006. It's not easy being interdisciplinary. Int J Epidemiol 35, 1119–1122. https://doi.org/10.1093/ije/dyl200

Moffitt, R., 2005. Remarks on the analysis of causal relationships in population research. Demography 42, 91–108. https://doi.org/10.1353/dem.2005.0006

Oliver, S., Bagnall, A., Thomas, J., Shepherd, J., Sowden, A., White, I., Dinnes, J., Rees, R., Colquitt, J., Oliver, K., Garrett, Z., 2010. Randomised controlled trials for policy interventions: a review of reviews and meta-regression. Health technology assessment (Winchester, England) 14, 1–iii. http://eprints.leedsbeckett.ac.uk/510/1/mon1416.pdf

Pallais, A., 2009. Taking a Chance on College Is the Tennessee Education Lottery Scholarship Program a Winner? J. Human Resources 44, 199–222. https://doi.org/10.3368/jhr.44.1.199

Pearl, J., 2000. Causality: Models, Reasoning and Inference Applications. Cambridge University press, New York.

Pearl, J., Bareinboim, E., 2011. Transportability of Causal and Statistical Relations: A Formal Approach, in: 2011 IEEE 11th International Conference on Data Mining Workshops. Presented at the 2011 IEEE 11th International Conference on Data Mining Workshops, pp. 540–547. https://doi.org/10.1109/ICDMW.2011.169

Pirog, M.A., Buffardi, A.L., Chrisinger, C.K., Singh, P., Briney, J., 2009. Are the alternatives to randomized assignment nearly as good? Statistical corrections to nonrandomized evaluations. Journal of Policy Analysis and Management 28, 169–172. https://doi.org/10.1002/pam.20411

Roach, M.A., Schanzenbach, M.M., 2015. The Effect of Prison Sentence Length on Recidivism: Evidence from Random Judicial Assignment (SSRN Scholarly Paper No. ID 2701549). Social Science Research Network, Rochester, NY.

Rothman, K.J., Greenland, S., Lash, T.L., 2008. Modern Epidemiology. Lippincott Williams & Wilkins, Philadelphia, PA.

Sanbonmatsu, L., Katz, L.F., Ludwig, J., Gennetian, L.A., Duncan, G.J., Kessler, R.C., Adam, E.K., McDade, T., Lindau, S.T., 2011. Moving to Opportunity for Fair Housing Demonstration Program: Final Impacts Evaluation.

Schneider, D., Harknett, K., 2018. What's not to like? Facebook as a tool for survey data collection, in: Proceedings of the Population Association of America Annual Meeting.

Shadish, W.R., 2011. Randomized Controlled Studies and Alternative Designs in Outcome Studies: Challenges and Opportunities. Research on Social Work Practice 21, 636–643. https://doi.org/10.1177/1049731511403324

Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton-Mifflin, Boston.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R., 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine 12, e1001779. https://doi.org/10.1371/journal.pmed.1001779

Tang, Y., Cook, T.D., Kisbu-Sakarya, Y., Hock, H., Chiang, H., 2017. The Comparative Regression Discontinuity (CRD) Design: An Overview and Demonstration of its Performance Relative to Basic RD and the Randomized Experiment, in: Regression Discontinuity Designs, Advances in Econometrics. Emerald Publishing Limited, pp. 237–279. https://doi.org/10.1108/S0731-905320170000038011

Vable, A.M., Cohen, A.K., Leonard, S.A., Glymour, M.M., Duarte, C. d. P., Yen, I.H., 2018. Do the health benefits of education vary by sociodemographic subgroup? Differential returns to education and implications for health inequities. Annals of Epidemiology 28, 759-766.e5. https://doi.org/10.1016/j.annepidem.2018.08.014

van der Laan, M.J., Rose, S., 2011. Targeted Learning: Causal Inference for Observational and Experimental Data. Springer Science & Business Media.

Wong, V.C., Steiner, P.M., 2018. Designs of Empirical Evaluations of Nonexperimental Methods in Field Settings. Eval Rev 0193841X18778918. https://doi.org/10.1177/0193841X18778918