Title

Alternative causal inference methods in population health research: Evaluating tradeoffs and triangulating evidence.

Authors

Matthay, Ellicott C
Hagan, Erin
Gottlieb, Laura M
et al.

Publication Date

2020-04-01

DOI

10.1016/j.ssmph.2019.100526

Peer reviewed

SSM - Population Health 10 (2020) 100526

Contents lists available at ScienceDirect

SSM - Population Health

journal homepage: http://www.elsevier.com/locate/ssmph

Article

Alternative causal inference methods in population health research: Evaluating tradeoffs and triangulating evidence

Ellicott C. Matthay [a,b,*], Erin Hagan [a], Laura M. Gottlieb [a], May Lynn Tan [a], David Vlahov [c], Nancy E. Adler [a], M. Maria Glymour [a,b]

[a] Center for Health and Community, University of California, San Francisco, 3333, California St, Suite, 465, Campus Box 0844, San Francisco, CA, 94143-0844, USA
[b] Department of Epidemiology and Biostatistics, University of California, San Francisco, 550 16th Street, 2nd Floor, Campus Box 0560, San Francisco, CA, 94143, USA
[c] Yale School of Nursing at Yale University, 400 West Campus Drive, Room 32306, Orange, CT, 06477, USA

ARTICLE INFO

Keywords:
Causal inference
Quasi-experiment
Instrumental variable
Econometrics
Epidemiologic methods
Threats to validity

ABSTRACT:

Population health researchers from different fields often address similar substantive questions but rely on different study designs, reflecting their home disciplines. This is especially true in studies involving causal inference, for which semantic and substantive differences inhibit interdisciplinary dialogue and collaboration. In this paper, we group nonrandomized study designs into two categories: those that use confounder-control (such as regression adjustment or propensity score matching) and those that rely on an instrument (such as instrumental variables, regression discontinuity, or differences-in-differences approaches). Using the Shadish, Cook, and Campbell framework for evaluating threats to validity, we contrast the assumptions, strengths, and limitations of these two approaches and illustrate differences with examples from the literature on education and health. Across disciplines, all methods to test a hypothesized causal relationship involve unverifiable assumptions, and rarely is there clear justification for exclusive reliance on one method. Each method entails trade-offs between statistical power, internal validity, measurement quality, and generalizability. The choice between confounder-control and instrument-based methods should be guided by these tradeoffs and consideration of the most important limitations of previous work in the area. Our goals are to foster common understanding of the methods available for causal inference in population health research and the tradeoffs between them; to encourage researchers to objectively evaluate what can be learned from methods outside one's home discipline; and to facilitate the selection of methods that best answer the investigator's scientific questions.

1. Introduction

Quantitative population health researchers come from diverse disciplines, including epidemiology, sociology, demography, psychology, and economics. Investigators in these fields use different terminologies and methodologies, even when addressing identical research questions. Many researchers rely almost exclusively on methods common in their home discipline, which may comprise only a subset of informative research designs for any given area. Moreover, lack of shared language and understanding inhibits mutually beneficial interdisciplinary dialogue and collaboration.

Disciplinary preferences are especially pronounced in research intended to support causal inferences. Randomized trials can provide compelling evidence of causation, but many questions of interest to population health researchers involve situations where randomization is not ethical or feasible. In these instances, researchers employ alternative designs, most of which can be categorized informatively as either "confounder-control" or "instrument-based". In confounder-control studies, researchers compare outcomes for people observed to have differing treatments (Box 1, definition 2) and use various statistical adjustment methods to account for imbalances in characteristics between treatment groups. In instrument-based studies, researchers leverage an apparently arbitrary or "exogenous" (Box 1, definition 5) 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. Many instrument-based study designs can be described as "quasi-experimental", although this term has been used inconsistently in prior research. Confounder-control and instrument-based approaches tend to be specific to disciplines. For example, in the first 6 months of 2018, 22

* Corresponding author. UCSF Department of Epidemiology and Biostatistics, Box 0560 550 16th Street, 2nd Floor, San Francisco, CA, 94143, USA.
E-mail address: Ellicott.matthay@ucsf.edu (E.C. Matthay).

https://doi.org/10.1016/j.ssmph.2019.100526
Received 14 May 2019; Received in revised form 4 November 2019; Accepted 1 December 2019
Available online 9 December 2019

of 25 nonrandomized studies of the causal effects of social exposures on health outcomes in the *American Journal of Epidemiology* employed confounder-control designs while all 6 such studies in the *American Economic Journal* used instrument-based approaches. The divide has motivated comments and efforts to bridge across disciplines (Abrams, 2006; Craig et al., 2012; Gunasekara, Carter, & Blakely, 2008; Kindig, 2007; Krieger, 2000; Lynch, 2006).

In our role evaluating investigator-initiated submissions for a grant-making program focused on improving population health and addressing health inequities (the Evidence for Action program of the Robert Wood Johnson Foundation), these disciplinary methodological divides are evident and have compelled us to take into consideration the pros and cons of different designs. Drawing on examples from the literature on educational attainment and health (Galama, Lleras-Muney, & van Kippersluis, 2018), in this paper we compare confounder-control and instrument-based approaches. Specifically, we apply Shadish, Cook, and Campbell's threats to validity framework to consider the tradeoffs in confounder-control versus instrument-based studies. We also provide simplified summaries of these two approaches, highlighting important distinctions, strengths, and limitations. Because inconsistent terminology is a persistent challenge for interdisciplinary research, we include informal definitions for how we use key terms in this paper in Boxes 1-3 (also see (Angrist & Pischke, 2008; Pearl, 2000; Rothman, Greenland, & Lash, 2008; Shadish, Cook, & Campbell, 2002)).

In this paper, we aim to enhance appreciation of the methodological landscape in which design choices are made by emphasizing an essential distinction between research methods, namely covariate-control versus instrument-based designs. We emphasize that the required assumptions and means of achieving valid causal inference differ markedly between approaches but both depend on unverifiable assumptions. Choosing a method entails tradeoffs between statistical power, internal validity, measurement quality, and generalizability. Therefore, neither confounder-control nor instrument-based approaches will always be preferable for all questions. Depicting the tradeoffs implicit in these different approaches should encourage researchers to evaluate inferences that can be culled from alternative approaches on a case-by-case basis, recognizing that complementary evidence from diverse designs will probably provide the best path to robust causal inferences. This paper is also intended to provide a framework for new population health researchers, who are currently being trained in numerous disciplines, to recognize the potential value of other methodological traditions and to select the research approach that will best address their scientific questions.

2. Defining the research question

Consider the question of whether college completion affects adult mortality. This research question is causal and considerably 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, definition 4). For any individual, we want to know whether survival is longer if she completes college than if she stops 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 specific sample. 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 correctly estimate the effect of education because those with differing levels of education likely differ on other characteristics that influence survival. Confounder-control and instrument-based study designs address this fundamental problem in distinct ways.

The distinction between confounder-control and instrument-based studies is grounded in Pearl's transdisciplinary causal inference framework (Pearl, 2000), which defines three approaches to achieve causal inferences: approaches based on the backdoor criterion (which we refer to as confounder-control), approaches based on an instrumental variable (which we refer to as instrument-based), and approaches based on the front door criterion (which are rarely used and not addressed in this paper). This distinction usefully frames myriad popular methodological approaches within a contemporary causal framework. We describe these approaches in the next two sections.

3. Estimating causal effects by controlling for confounders

In confounder-control studies, researchers address imbalances in characteristics between treatment groups (e.g. age, education, income, etc.) by statistically adjusting for these characteristics. This study design is particularly common in epidemiology, in which variation in an exposure of interest is commonly characterized in a group of individuals who are then followed to assess subsequent health outcomes (a "cohort study") (Rothman et al., 2008). For all confounder-control studies, the analytic strategy used to estimate the causal effect of interest is to ascertain, measure, and appropriately adjust for a "sufficient set" of variables (or proxies for those variables) to control confounding (Box 3, definition 1). Modern frameworks define confounding as arising from shared causes of treatment and outcome; such factors can create associations between treatment and outcome even if the treatment has no causal effect on the outcome. Sufficient sets of confounders are selected

Box 1
Terminology for Describing Causal Questions

1. **Causal model:** A description, most often expressed as a system of equations or a diagram, of a researcher's assumptions about hypothesized or known causal relationships among variables relevant to a particular research question.
2. **Treatment**, **exposure,** or **independent variable**: The explanatory variable of interest in a study. In this paper, we use these terms synonymously even for exposures that are not medical "treatments", such as social resources or environmental exposures. Some writers also describe this as the "right-hand-side variable".
3. **Outcome**, **dependent variable**, or **left-hand-side variable**: The causal effect of interest in a research study is the impact of an exposure(s) on an outcome(s).
4. **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.
5. **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 causal model. The variable is called endogenous if it is influenced by other variables in the causal model. If a third variable influences both the exposure and outcome, this implies the exposure is endogenous.

Box 2
Terminology for Study Designs and Causal Effects

1. **Confounder-control study**: A study in which effects of a treatment are estimated by comparing outcomes of treated to untreated individuals and potential imbalances in confounding variables between treated and untreated groups are addressed with adjustment, stratification, weighting, or similar methods. Treatment in these settings may be determined by the individual's own preferences, behaviors, or other naturally occurring influences. This study type corresponds to causal inference by fulfilling the backdoor criterion (Box 3, definition 5) under Pearl's framework (Pearl, 2000).
2. **Instrument-based study**: A study in which effects of a treatment are estimated by leveraging apparently random or arbitrary factors that alter the chances an individual will receive a treatment, e.g., due to external factors such as the timing of policy changes. This is analogous to randomization in a randomized controlled trial, in which random assignment affects the chances an individual will be treated but is otherwise unrelated to the outcome. The source of variation is often called an instrumental variable (Box 2, definition 3). This study type corresponds to causal inference by leveraging an instrumental variable under Pearl's framework (Pearl, 2000).
3. **Instrument** or **instrumental variable**: An external factor that induces differences in the chance an individual will be exposed to the treatment of interest but 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.
4. **Forcing variable:** A variable with a threshold such that people just above the threshold are much more likely to be treated than people just below the threshold (or vice-versa). The threshold provides the discontinuity in regression discontinuity designs. The forcing variable, sometimes called the running variable, may also have a continuous, dose-response association with the outcome.
5. **Population average treatment effect (PATE)**: The difference in the average outcome if everyone in the population were treated compared to the average outcome if nobody in the population were treated. Because the effect of treatment might not be the same for everybody in the population, the PATE is distinguished from treatment effects in various subgroups.
6. **Average treatment effect among the treated** or **effect of treatment on the treated (ATT** or **ETT):** The average treatment effect among those people who actually received treatment. This might differ from the PATE, for example, if the people most likely to benefit from treatment were also the most likely to be treated.
7. **Local average treatment effect (LATE)**: The average treatment effect among those whose treatment status was changed by the instrumental variable. This might differ from the PATE, for example, if the instrumental variable was a policy change that increased the chances of treatment for the people who were most likely to benefit from treatment.

Box 3
Types of Bias and Assumptions for Causal Inference

1. **Confounding** or **omitted variable bias** or **bias from selection into treatment:** The key bias introduced by lack of randomization. This bias 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 (e.g., parental education may influence both a child's own education and that child's later health; if not accounted for, parental education confounds the association between the child's education and subsequent health). This bias 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.
2. **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.
3. **Reverse causation:** When the outcome causes the treatment, rather than the treatment causing the outcome.
4. **Exchangeability**, **ignorability**, **no confounding**, or **randomization assumption:** 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.
5. **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.
6. **Relevance**: In the context of instrumental variables, the assumption that the instrument affects the treatment.
7. **Exclusion restriction**: In the context of instrumental variables, the assumption that, conditional on measured covariates, the instrument only affects the outcome through the treatment.
8. **Monotonicity**: In the context of instrumental variables, the assumption that the instrument does not have the opposite direction of effect on chances of treatment for different people in the population.
9. **Positivity** or **common support**: All subgroups of individuals defined by covariate stratum (e.g., every combination of possible covariate values) must have a nonzero chance of experiencing every possible exposure level. Put another way, within every covariate subgroup, all exposure values of interest must be possible.
10. **Consistency**: The assumption that an individual's potential outcome setting treatment to a particular value is that person's actual outcome if s/he actually has that particular value of treatment. This could be violated if the outcome might depend on how treatment was delivered or some other variation in the meaning or content of the treatment. Some researchers consider consistency a truism rather than an assumption.
11. **Stable unit treatment value assumption (SUTVA)**: The assumption that all versions of the treatment has the same effect (i.e., versions of the treatment with differences substantial enough to have different health effects are referred to as some other type of treatment), and each unit's outcomes are unaffected by the treatment values of other units.

based on substantive knowledge, prior research, or expert judgement. In epidemiology and other disciplines, causal diagrams have emerged as popular tools to visually represent the content area assumptions which guide selection of sufficient sets of covariates (Pearl, 2000; van der Laan & Rose, 2011).

Causal diagrams, including directed acyclic graphs (DAGs), are causal models (Box 1, definition 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, researchers acknowledge 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.

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 cause both exposure and outcome, strategies to reduce confounding focus on breaking the association of the confounders with the outcome (e.g., regression adjustment); breaking the association of the confounders with the exposure (e.g., matching, adjustment, or weighting based on propensity scores); or breaking both the association with the exposure and the outcome (e.g., doubly robust methods). These methods all eliminate confounding by making comparisons within subgroups or pseudo populations that have balanced covariates, such that the covariates cannot bias the treatment-outcome association. Under Pearl (Pearl, 2000), this is called fulfilling the backdoor criterion (Box 3, definition 5). Some investigators also conceptualize matching as fulfilling the backdoor criterion because of a perfect offsetting balance between the influence of confounding variables and the spurious association within matched pairs (Kim, Steiner, Hall, & Su, 2016; Kim & Steiner, 2019). These concepts are relevant across all approaches to matching based on observable covariates, regardless of the specific algorithms used to create matches (Ho, Imai, King, & Stuart, 2007; Stuart, 2010).

Confounder-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 healthcare. The parameter most commonly estimated is the sample average treatment effect, which is commonly interpreted as an estimate of the population average treatment effect (PATE; Box 2, definition 5), although some methods by default deliver the effect of treatment on the treated (ETT).

Panel fixed effects can also be considered confounder-control. In this approach, treatments and outcomes are measured repeatedly on the same participants over time. Binary indicator variables representing each participant are used to control for characteristics of participants that do not change over the study period (e.g., genes). This approach treats the unchanging characteristics of the individual as the confounders requiring adjustment and estimates the effect of treatment by comparing how differences in the treatment received for the same individual at different times relate to differences in that individual's outcomes. Similar fixed effects approaches have also been applied to studies of twins (Lundborg, Lyttkens, & Nystedt, 2016).

A parallel approach is applicable for questions about the health effects of policies in which places are the unit of analysis. Time-constant features of places (e.g., altitude) are controlled by including indicator variables representing each place. Analyses also commonly include indicator variables representing each time period to account for events 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). Thus indicator variables control for both time-invariant aspects of units and unit-invariant aspects of time. The only remaining confounders of concern are those that change over time in different ways in different places, and standard principles for confounder-control can be applied to these.

Confounder-control study designs deliver valid effect estimates if and only if a sufficient set of confounders is correctly identified, the available data include a high-quality measure of each confounder (or proxy for each confounder, in the case of panel fixed effects), and each confounding variable is modeled correctly in its relation with either the exposure or the outcome.

4. Instrument-based approaches to estimating causal effects

Despite growing recognition of opportunities to implement randomization to evaluate complex social risk factors (Duflo, Glennerster, & Kremer, 2007), studies without formal randomization remain essential in population health research. Instruments create differences in the treatment received between individuals who are otherwise similar. Conceptually, these designs share characteristics with experimental randomized studies. To make causal inferences, both designs require two key assumptions to be true: (1) the exogenous factor—whether randomization or some other program, policy, or arbitrary variation—must create differences in the chance of receiving treatment between groups of individuals with otherwise similar potential outcomes, and (2) the exogenous factor must have no other mechanism to influence the outcome except via the treatment under consideration (Angrist & Pischke, 2008). For all instrument-based and experimental randomized studies, when these assumptions are met, the statistical associations between the exogenous factor, the treatment, and the outcome can then be leveraged to estimate the causal effect of the treatment on the outcome. In the rare situation when the instrument and treatment are equal, as in a randomized trial with perfect adherence, confounder-control and instrument-based approaches are equivalent. This is the only situation in which the methods converge (Pearl, 2000). Sources of instruments include lotteries (sometimes used to assign wartime drafts (Buckles, Hagemann, Malamud, Morrill, & Wozniak, 2016), housing vouchers (Sanbonmatsu et al., 2011) or other resources (Eisenberg & Rowe, 2009; Pallais, 2009) when there is not enough for all eligible individuals), arbitrary assignment of judges (who have different propensities for leniency (Roach & Schanzenbach, 2015)) or clinicians (who have different preferences for treatment modalities (Brookhart & Schneeweiss, 2007)), changes in policies at unprecedented times (Andriano & Monden, 2019; Clark & Royer, 2013; Malamud, Mitrut, & Pop-Eleches, 2018), month or quarter of birth (which influences age of school enrollment (Acemoglu & Angrist, 1999)), or biological chance, such as the sex of a child (which influences chances parents will wish to conceive another child (Angrist & Evans, 1998)). Instruments 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 on instruments. This collection of techniques is common in economics and other social and behavioral sciences (Angrist & 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 that policy. Both types of questions are usually of interest. IV analyses deliver estimates of the effects of the exposure, treatment, or resource.

These estimates are useful, because once the effect of exposure is known, alternative policies to influence treatment can be considered and some may be preferable for reasons such as political feasibility. Furthermore, 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. Instrument-based designs can be deployed to evaluate effects of policies themselves, but in this paper we focus on research about the effects of the treatments determined by those policies. Designs variously referred to as RD and DiD can all be conceptualized and statistically analyzed as IVs (Angrist & Pischke, 2008).

IV analyses control confounding by leveraging a source of variation in the treatment (the instrument) that is unrelated to other determinants of the outcome. A typical IV analysis requires the assumptions of relevance (the instrument must influence the treatment), exclusion (if the instrument affects the outcome, it is only via the treatment), and exchangeability (the instrument does not share unmeasured causes with the outcome) (Box 3, definitions 4, 6, and 7). Relevance can be tested. The other assumptions (i.e., exclusion and exchangeability) cannot be proven to be true and must be judged substantively. The treatment itself may be influenced by numerous confounders (i.e., variables that also affect the outcome), but when these assumptions are met, the variation in treatment that is predicted by the instrument is independent of the confounders. IV analyses quantify how this variation in treatment induced by the instrument affects the outcome. In an RCT, thinking of random assignment as an instrument, this corresponds to the effect of treatment received among those who would have adhered to their random assignment, regardless of whether assigned to treatment or placebo. This is the conceptual core of IV analysis.

RD methods are applicable when there is an arbitrary discontinuity in the probability of being treated depending on the value of a third "forcing" variable (Box 2). Examples of forcing variables include: age, when eligibility for a resource such as Medicare begins at a certain age; household income, when there is a sharp eligibility cutoff at a certain income level; or class size, when policy requires that classes with more than a certain number of students must be broken into two classrooms. Individuals immediately above and below that cut point are expected to have equivalent potential health outcomes, except for the effects of the sharp differences in treatment probability. If the discontinuity does not perfectly determine who is treated or untreated, this is termed a "fuzzy" RD and can be analyzed in the same manner as a traditional IV to evaluate the effect of treatment on health outcomes.

Goodman and colleagues (Goodman, Hurwitz, & Smith, 2015) described a fuzzy RD created by Georgia's State University System admission rules, which require SAT scores above 400 in math and 430 in reading. This created a discontinuity in the probability of beginning college at a 4-year institution for students just above or just below these SAT scores. Of course, some students with scores below these thresholds may attend college elsewhere, and some students with scores above them may enroll in 2-year colleges or forego attendance, so the RD is "fuzzy": students who scored a 400 in math and 430 in reading had about a 10 percentage point higher probability of beginning college at a 4-year institution than students who scored a 399 or 429 in math or reading, respectively. Goodman et al. 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—a distinct but related question—the investigator could use a regression approach to directly compare college completion for those just meeting and just missing the SAT score threshold. This approach reduces to confounder-control; for estimating the causal effect, no instrument is used.

DiD methods combine an RD with one or more external 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 of 6 years of schooling, whereas previously there was no required minimum. As a result of the policy change, children who began schooling right before policy implementation completed slightly less school than children who began in the years after the policy was adopted. We might hope to use this discontinuity to estimate the health effects of extra schooling. However, World War I or the great influenza pandemic might have altered long-term outcomes for those cohorts in ways completely unrelated to the additional schooling. In a DiD design, we would include comparison states that did not change their schooling laws in those years (say, Alabama) to control for these historical events. The key assumption of DiD is that, conditional on measured covariates, if Mississippi had not changed its policy in 1918, the trends in outcomes across birth cohorts would be parallel for Mississippi and Alabama. This amounts to an assumption of no confounding factors that changed at the same time as the mandatory schooling policy. DiD scenarios can be analyzed as traditional IVs where the interaction of the policy and timing of implementation serves as an instrument (e.g., Mississippi state and years after 1918 as an IV for educational attainment). We focus here on how to evaluate the health effects of receiving extra schooling, but if the research question is instead about the compulsory schooling policy itself, the causal inference approach is DiD via confounder-control, with state as the unit of analysis, analogous to panel fixed effects; no instrument is used. The latter approach can also be considered a special case of the comparative interrupted time series design (Shadish et al., 2002).

The strength of an analysis drawing on a valid instrument 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 instrument-based approaches are judged to be plausible only after conditioning on a set of covariates. We regard these as instrument-based studies, not confounder-control, because these studies ultimately leverage an instrument to estimate causal effects and adjustment for confounders is used to meet the instrument-based study assumptions.

Several statistical approaches can be applied to evaluate causal effects of a treatment on an outcome using an IV but we do not detail them here. Interpreting IV estimates—whether from a discontinuity, a difference-in-difference, or another exogenous source of variation in treatment—requires additional assumptions. If the effect of treatment on outcome is identical for everyone in the population, then the standard statistical IV estimate can be interpreted as the PATE (Box 2). However, most IV analyses instead adopt the monotonicity assumption (Box 3): that the IV does not have opposite effects on the chances of 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. Under this assumption, the parameter estimated by an IV approach is the LATE (Box 2). For example, the LATE corresponding with using the Georgia State University SAT score threshold as an IV would describe the effect of beginning college at a four-year institution on people who would have begun at a four-year college if and only if they scored above the SAT threshold. The choice to estimate the PATE, the LATE, or some other subgroup effect has important implications for generalizing findings to new settings.

5. Considerations and tradeoffs for all population health studies

Choosing among confounder-control and instrument-based approaches entails tradeoffs (Table 1). Shadish, Cook, and Campbell's causal inference framework (Shadish et al., 2002), which has been widely influential in a range of population health disciplines (Cook, 2018), is useful to consider which study design is preferable in any given setting. Shadish, Cook, and Campbell categorize study validity for causal

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

Feature	Confounder-control	Instrument-based
Main strategies for estimating causal effects	Identify, measure, and control for a sufficient set of confounders through matching, regression adjustment, propensity score methods, or related methods.	Identify and leverage a random or conditionally random source of variation in chances of treatment which would be otherwise unrelated to the outcome through instrumental variables, regression discontinuity, differences-in-differences, or related approaches.
Key assumptions	Conditional exchangeability between treated and untreated individuals, including no uncontrolled common causes of treatment and outcome.	Variation in the instrumental variable alters chances of treatment, is unrelated to potential outcomes, and influences the outcome via no other mechanism except the treatment at hand. The instrument's variation cannot have opposite effects on probability of treatment for different people in the study.
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 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 instrumental variables, or testable implications of the hypothesized causal mechanisms.
Typical analyses	Regression with confounder-control. Propensity score matching, adjustment, or weighting. Doubly robust analyses.	Two-stage least-squares regression. Method of moments. Residual control methods.
Key methodological advantages	Analyses leverage treatment variation in the entire populations, improving statistical power relative to instrument-based approaches with the same data source. Often based on diverse and representative samples that facilitate assessment of differential treatment effects across and within populations.	Study design and analytic approaches can circumvent bias from unmeasured confounders of the treatment-outcome association. Can deliver a treatment effect specific to the individuals most affected by the instrument.
Key methodological challenges	Reliance on identifying, measuring, and appropriately adjusting for all confounders.	Valid sources of instruments can be difficult to identify. Reduced statistical power relative to total-population studies. Treatment effects (LATE) only generalize to the subset of participants whose treatment is affected by the instrument.

See Boxes 1-3 for definitions. We present a simplified characterization of each approach to highlight key distinctions.

inference 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 people, similar treatments, alternative measures, or other settings.

Below we discuss how confounder-control and instrument-based approaches typically perform with respect to each of these categories of validity.

5.1. *Internal validity*

Internal validity requires some type of conditional exchangeability or randomization assumption (Box 3). The choice between confounder-control and instrument-based 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. For example, those who pursue college likely differ from those who do not on a wide variety of factors that may impact health. Thus, the primary limitation of confounder-control approaches is the reliance on identifying, measuring, and correctly adjusting for a sufficient set of confounders. Confounder-control study designs are particularly appealing when achieving this task seems feasible, or when previous efforts can be improved upon.

Instrument-based strategies are appealing for internal validity because of the possibility that they can address unmeasured confounders of the treatment-outcome association. Instruments are hardly a panacea, however, because it is often difficult to identify a valid instrument to answer the study question of interest. Many examples of failed instruments exist; for example, Barua and Lang found that legal school entry age, a commonly used instrument for the effects of schooling, fails to meet the monotonicity assumption (Barua & Lang, 2016). The exchangeability and exclusion assumptions are controversial in many applied examples (French & Popovici, 2011).

In addition to exchangeability, causal inference requires the assumptions of "positivity" and "consistency" (Box 3, definitions 9–10). In covariate-control methods, a positivity violation implies the exposure and a confounder cannot be disentangled—for example, if all treated individuals were also veterans and many untreated individuals were not, it would be impossible to disentangle the effects of treatment from veteran status. Consistency violations (including violations of the stable unit treatment value assumption [SUTVA]—Box 3, definition 11) occur when there are many different "flavors" of exposure with the same measured value (e.g., "college degree completion") that may have very different consequences for the outcome (e.g., the health benefits of the degree may depend on the institution and major). Consistency violations undermine covariate-control methods. In contrast, for instrument-based methods, violations of positivity and consistency for the exposure are not necessarily problematic because, by definition, these methods estimate effects of variation in exposure induced by the instrument (Glymour, Tchetgen Tchetgen, & Robins, 2012).

Internal validity may also be threatened by imperfectly measured variables, regression model misspecification, reverse causation (Box 3), inadvertently controlling for factors that are influenced by exposure, or differential loss-to-follow-up, among others. For example, in a confounder-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 confounder-control and instrument-based 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 (Lipsitch, Tchetgen, & Cohen, 2010). For example, if we found

that Georgia students who scored just above an irrelevant math SAT threshold such as 600 were substantially more likely to complete college than students who scored just below the threshold, this would call into question the validity of that particular threshold for estimating causal effects of the child's own college experience.

5.2. 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). SUTVA—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 is a particular challenge in instrument-based studies, because inferences are constrained to the fraction of the study population whose exposure is affected by the instrument. For example, compulsory schooling laws only impact educational attainment for a fraction of the study population (most students do not determine when to complete their education by referring to state law), and the sample size is effectively limited to that fraction. This can result in wide confidence intervals or underpowered studies. Exclusive reliance on instrument-based approaches thus may risk failing to identify important interventions with only small to moderate effect sizes. A confounder-control approach to the same question in the same population may deliver more precise effect estimates and be powered to identify smaller effects sizes.

5.3. 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. Program participation may have multiple consequences besides the intentionally delivered services (e.g. if college attendance is accompanied by job search support which substantially enhances subsequent earnings). Such multi-faceted causal links between an intervention and health threaten construct validity. 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 with limited budgets, researchers may opt for smaller sample sizes. Studies grounded in large administrative datasets benefit from greater statistical power but tend to have less detailed measurements; smaller studies can often afford higher quality measurements. Because instrument-based approaches intrinsically sacrifice statistical power, they may have to rely on large, frequently administrative data sets. 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 targeted enrollment of participants most affected by a given instrument (Schneider & Harknett, 2018).

5.4. External validity

For population health, we nearly always hope to extrapolate study results to a larger group of people than just those directly involved in the study (Westreich, Edwards, Lesko, Cole, & Stuart, 2019). Causal inferences in one population cannot be generalized to a new setting if the causal relationship of interest is modified by participant characteristics, settings, or treatment variations which differ in the new setting. Researchers address generalizability based on a priori theory guiding interpretation of results or empirical evidence on the characteristics of the study population compared to the target population (e.g. with respect to sociodemographics or geography). 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 & 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. Many confounder-control 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 in these studies also supports the evaluation of differential effects across population subgroups. Once we understand differential effects, we can anticipate how a treatment would play out in a new population. For example, if studies of education and health include both White and Black participants, differential effects can be directly evaluated (Assari & Mistry, 2018; Cohen, Rehkopf, Deardorff, & Abrams, 2013; Kaplan, Ranjit, & Burgard, 2008; Liu, Manly, Capistrant, & Glymour, 2015; Vable et al., 2018) and then applied to anticipate population average treatment effects in predominantly White or predominantly Black populations.

Generalizing can be more challenging in instrument-based studies, because they typically deliver the LATE, not the PATE. It is technically impossible to determine if any individual study participant would have adhered to their assigned treatment whether assigned to treatment or control, and thus impossible to identify the individuals whose effects are described by the LATE. 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 historically enforced for White but not 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 an instrument-based approach leveraging arbitrary differences in judicial leniency can be extremely informative in population health research.

6. Discussion

Population health researchers use a variety of approaches to derive causal inferences in the absence of ideal randomization. We present simplified characterizations these approaches with the goal of fostering cross-disciplinary communication and enhancing use of the full

spectrum of causal inference tools available to population health scientists. We find the distinction between confounder-control approaches and instrument-based approaches valuable for highlighting the complementary strengths and limitations of alternative designs. The approaches presented in this paper are distinct, but rarely in conflict. Both rely on unverifiable assumptions and each approach has tradeoffs. Which set of untestable assumptions is more appealing depends on the problem and data at hand as well as 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, Tilling, & Davey Smith, 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 or any set of studies using the same design (Cordray, 1986, 1986; Duncan, 2008).

To our knowledge, there has been little systematic attention to categorizing the types of problems amenable to confounder-control approaches, problems where instrument-based approaches are preferable, or problems for which neither will deliver 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 from confounder-control or instrument-based methods applied to the trial's treatment group and an externally derived untreated population such as a population-representative survey (Wong & 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 instrument-based and confounder-control approaches perform well (Pirog, Buffardi, Chrisinger, Singh, & Briney, 2009; Shadish, 2011).

Reasons for the inconsistent performance of instrument-based or confounder-control methods in recapitulating experimental results are not fully understood (Oliver et al., 2010) and whether results from randomized trials are the appropriate benchmark continues to be debated (Deaton & Cartwright, 2018; Gelman, 2018; Hanin, 2017; Ioannidis, 2018; Trentino, Farmer, Gross, Shander, & Isbister, 2016). Although it has been suggested that regression discontinuity more reliably replicates experimental results than other confounder-control or instrument-based approaches (Pirog et al., 2009; Shadish, 2011), this may be because the situations in which regression discontinuity can be applied are more likely to succeed in controlling confounding regardless of the analytic approach because treatment is mainly determined by known, measured variables. Prior comparison studies have several limitations: they rarely consider applications to social determinants of health (examples of exceptions are (Gennetian, Hill, London, & Lopoo, 2010; Handa & Maluccio, 2010)); they have not utilized modern methods for analysis in confounder-control studies, methods which provide rigorous procedures for covariate selection and make fewer assumptions about shapes of relationships between variables; the comparison studies rarely consider external validity (exceptions exist, e.g. (Jaciw, 2016)); and by definition they 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 about the performance of instrument-based and confounder-control methods in anticipating RCT results. Good correspondence has been seen for observational results and randomized trials in clinical epidemiology (Anglemyer, 2014), but causal inferences about exposures with the greatest relevance to population health may be more challenging. Further research is needed.

For population health research, the challenge of causal inference in the absence of randomization is a so-called "wicked" problem. No single approach is likely to provide conclusive evidence, but may be complementary to other methods. Researchers must recognize and appreciate how different approaches fit together. Understanding the tradeoffs entailed by methodologic choices is critical for either interpreting or contributing to the current evidence on the drivers of population health.

Role of the funding source

This work was supported by the Evidence for Action program of the Robert Wood Johnson Foundation (RWJF). RWJF had no role in the study design; collection, analysis, or interpretation of data; writing of the article; or the decision to submit it for publication.

Ethics

No human subjects were involved in the research presented in our manuscript.

Declaration of competing interest

The authors have no competing interests to declare.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ssmph.2019.100526.

References

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*. National Bureau of Economic Research. https://doi.org/10.3386/w7444 (Working Paper No. 7444).

Andriano, L., & Monden, C. W. S. (2019). The causal effect of maternal education on child mortality: Evidence from a quasi-experiment in Malawi and Uganda. *Demography*. https://doi.org/10.1007/s13524-019-00812-3.

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. *The 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.

Barua, R., & Lang, K. (2016). School entry, educational attainment, and quarter of birth: A cautionary tale of a local average treatment effect. *Journal of Human Capital, 10*, 347–376. https://doi.org/10.1086/687599.

Brookhart, M. A., & Schneeweiss, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects: Assessing validity and interpreting results. *International Journal of Biostatistics, 3*. https://doi.org/10.2202/1557-4679.1072.

Buckles, K., Hagemann, A., Malamud, O., Morrill, M., & Wozniak, A. (2016). The effect of college education on mortality. *Journal of Health Economics, 50*, 99–114. https://doi.org/10.1016/j.jhealeco.2016.08.002.

Clark, D., & Royer, H. (2013). The effect of education on adult mortality and health: Evidence from Britain. *The American Economic Review, 103*, 2087–2120. https://doi.org/10.1257/aer.103.6.2087.

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., et al. (2012). Using natural experiments to evaluate population health interventions: New medical research council guidance. *Journal of Epidemiology & Community Health, 66*, 1182–1186. https://doi.org/10.1136/jech-2011-200375.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. Social science & medicine, randomized controlled trials and evidence-based policy. *A Multidisciplinary Dialogue, 210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Chapter 61 using randomization in development economics research: A toolkit. In *Handbook of development economics* (pp. 3895–3962).

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.

French, M. T., & Popovici, I. (2011). That instrument IS lousy! IN search OF agreement when using instrumental variables estimation IN substance use research. *Health Economics, 20*, 127–146. https://doi.org/10.1002/hec.1572.

Galama, T. J., Lleras-Muney, A., & van Kippersluis, H. (2018). *The effect of education on health and mortality: A review of experimental and quasi-experimental evidence*. National Bureau of Economic Research. https://doi.org/10.3386/w24225 (Working Paper No. 24225).

Gelman, A. (2018). Benefits and limitations of randomized controlled trials: A commentary on deaton and Cartwright. Social science & medicine, randomized controlled trials and evidence-based policy. *A Multidisciplinary Dialogue, 210*, 48–49. https://doi.org/10.1016/j.socscimed.2018.04.034.

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.

Glymour, M. M., Tchetgen Tchetgen, E. J., & Robins, J. M. (2012). Credible mendelian randomization studies: Approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology, 175*, 332–339. https://doi.org/10.1093/aje/kwr323.

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

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.

Hanin, L. (2017). Why statistical inference from clinical trials is likely to generate false and irreproducible results. *BMC Medical Research Methodology, 17*, 127. https://doi.org/10.1186/s12874-017-0399-0.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*, 199–236. https://doi.org/10.1093/pan/mpl013.

Ioannidis, J. P. A. (2018). Randomized controlled trials: Often flawed, mostly useless, clearly indispensable: A commentary on deaton and Cartwright. Social science & medicine, randomized controlled trials and evidence-based policy. *A Multidisciplinary Dialogue, 210*, 53–56. https://doi.org/10.1016/j.socscimed.2018.04.029.

Jaciw, A. P. (2016). Assessing the accuracy of generalized inferences from comparison group studies using a within-study comparison approach: The methodology. *Evaluation Review, 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.

Kim, Y., & Steiner, P. M. (2019). Gain scores revisited: A graphical models perspective. *Sociological Methods & Research*. https://doi.org/10.1177/0049124119826155, 0049124119826155.

Kim, Y., Steiner, P. M., Hall, C. E., & Su, D. (2016). Graphical models for quasi-experimental designs. *Society for Research on Educational Effectiveness*.

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. *Epidemiological Review, 22*, 155–163. https://doi.org/10.1093/oxfordjournals.epirev.a018014.

van der Laan, M. J., & Rose, S. (2011). *Targeted learning: Causal inference for observational and experimental data*. Springer Science & Business Media.

Langa, K. M., Plassman, B. L., Wallace, R. B., Herzog, A. R., Heeringa, S. G., Ofstedal, M. B., et al. (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. *International Journal of Epidemiology, 45*, 1866–1886. https://doi.org/10.1093/ije/dyw314.

Lipsitch, M., Tchetgen, E. T., & Cohen, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology, 21*, 383–388. https://doi.org/10.1097/EDE.0b013e3181d61eeb.

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.

Lundborg, P., Lyttkens, C. H., & Nystedt, P. (2016). The effect of schooling on mortality: New evidence from 50,000 Swedish twins. *Demography, 53*, 1135–1168. https://doi.org/10.1007/s13524-016-0489-3.

Lynch, J. (2006). It's not easy being interdisciplinary. *International Journal of Epidemiology, 35*, 1119–1122. https://doi.org/10.1093/ije/dyl200.

Malamud, O., Mitrut, A., & Pop-Eleches, C. (2018). *The effect of education on mortality and health: Evidence from a schooling expansion in Romania*. National Bureau of Economic Research. https://doi.org/10.3386/w24341 (Working Paper No. 24341).

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., et al. (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? *Journal of Human Resources, 44*, 199–222. https://doi.org/10.3368/jhr.44.1.199.

Pearl, J. (2000). *Causality: Models, reasoning and inference applications*. New York: Cambridge University press.

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*. Rochester, NY: Social Science Research Network (SSRN Scholarly Paper No. ID 2701549).

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology*. Philadelphia, PA: Lippincott Williams & Wilkins.

Sanbonmatsu, L., Katz, L. F., Ludwig, J., Gennetian, L. A., Duncan, G. J., Kessler, R. C., et al. (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*. Boston: Houghton-Mifflin.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*, 1–21. https://doi.org/10.1214/09-STS313.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., et al. (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.

Trentino, K., Farmer, S., Gross, I., Shander, A., & Isbister, J. (2016). Observational studies - should we simply ignore them in assessing transfusion outcomes? *BMC Anesthesiology, 16*. https://doi.org/10.1186/s12871-016-0264-4.

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. https://doi.org/10.1016/j.annepidem.2018.08.014. e5.

Westreich, D., Edwards, J. K., Lesko, C. R., Cole, S. R., & Stuart, E. A. (2019). Target validity and the hierarchy of study designs. *American Journal of Epidemiology, 188*, 438–443. https://doi.org/10.1093/aje/kwy228.

Wong, V. C., & Steiner, P. M. (2018). Designs of empirical evaluations of nonexperimental methods in field settings. *Evaluation Review*. https://doi.org/10.1177/0193841X18778918, 0193841X18778918.