Quasi-experimental causality in neuroscience and behavioral research

Ioana E. Marinescu[1*], Patrick N. Lawlor[2], Konrad P. Kording[3]

[1]School of Social Policy and Practice, University of Pennsylvania, Caster Building D6, 3701 Locust Walk, Philadelphia, PA 19104

[2]Division of Neurology, Children's Hospital of Philadelphia, Colket Translational Research Building, 3501 Civic Center Blvd Office 10200-11, Philadelphia, PA 19104

[3]Departments of Neuroscience and Bioengineering, Leonard Davis Institute, Warren center for network science, Wharton Neuroscience Initiative, University of Pennsylvania, 106 Hayden Hall, 240 S 33rd St., Philadelphia, PA 19104, Canadian Institute For Advanced Research

*Corresponding author. Email: ioma@upenn.edu

Abstract

In many scientific domains, causality is the key question. For example, in neuroscience, we might ask whether a medication affects perception, cognition, or action. Randomized controlled trials (RCTs) are the gold standard to establish causality, but they are not always practical. The field of empirical economics developed rigorous methods to establish causality even when RCTs are not available. Here we review these quasi-experimental methods and highlight how neuroscience and behavioral researchers can use them to do research that can credibly demonstrate causal effects.

Introduction

Behavioral research asks a broad range of questions, and most of them are of a causal nature [1]. When we ask how a drug affects a patient, we want to know its causal effect: does it make the patient better? We do not want to ask the correlational question: does taking the drug correlate with well-being? Characteristics of the patient such as socioeconomic status may affect both the probability of being prescribed a drug and the patient's well-being. Similarly, in neuroscience we have many causal questions. For example, we are interested in how one brain area affects another brain area, as opposed to how the two brain areas are correlated. In psychology, we ask which interventions improve people's thriving, again not to be confused with correlation (people with big yachts are happier, but see [2]). The primary goal of the bulk of scientific research is to ask how elements of a system causally affect other elements. Causality is at the heart of many questions in behavior and neuroscience.

Ignoring the difference between correlation and causality frequently leads scientists to incorrect conclusions. In one notorious example from medicine, a correlational study suggested that hormone replacement therapy (HRT) may decrease the risk of cardiovascular disease in post-menopausal women [3]. A later randomized controlled trial, however, showed the opposite [4] – that HRT actually led to worse cardiovascular outcomes. The discrepancy likely resulted from influences of socioeconomic status [5]; women with higher socioeconomic status were both more likely to receive

HRT and to have better outcomes. In neuroscience, we have the same problem; many studies are inherently correlational due to the difficulty of controlling neurons and neural states. Reliably identifying causality without randomized experiments is difficult.

The crucial problem in causal inference is confounding. We would like to estimate the influence between two variables X and Y, but there may be other variables that affect both X and Y. If we *set* the value of X, as we do in a randomized experiment, there is no issue because the other variables can only affect Y. Running such experiments is the basis of the experimental method [6], and allows for a direct reading out of causal effects. This has been done, often at great cost, for education (e.g., The Perry Preschool Project and The Carolina Abecedarian Project), neuroscience (e.g., optogenetics, slice stimulation), clinical psychology (e.g., therapy comparison), and is frequently done in online user interactions (e.g., AB tests). In medicine, randomization is the gold standard and is called a randomized controlled trial (RCT). When we can set the relevant variables and randomize, answering causal questions meaningfully is far more straightforward.

But if we can only *observe* a system, then confounding is a serious problem; we can never know if an apparent interaction between X and Y is real or is *confounded* by the other variables. This is often the case for a number of reasons. First, there are many variables that we cannot easily set, e.g., the activity of neurons somewhere in the brain. Second, setting variables is often expensive, e.g. in the case of large clinical trials [7-9]. Third, randomized experiments can be unethical, since they can force us to withhold the intervention that we believe to be best. When we cannot set all of the variables of interest, confounding is a serious issue which makes it hard to learn about the effect of one variable on another. Yet, most of the world's ever-growing data do not come from randomized experiments and we should not waste this data.

In response to the confounding problem in observational data, there are two important schools of thought. One school attempts to builds large, complex models that observe and model all confounders. Such models thus assume that confounding is unlikely or impossible. There are many widely-used methods that make this *unconfoundedness* assumption in neural and behavioral research e.g., Granger causality [10] (see Box 1 for others). However, unconfoundedness is rarely plausible as virtually all systems that we study have more variables of importance than we can realistically measure or model. A second school of thought that has arisen in response to the confounding problem focuses on quasi-experiments [11]. Although we may not assume unconfoundedness in general, we may still be able to find variables in our data that are assigned in a way that is as good as random. This paper focuses on discussing this second way of thinking about causal inference.

This second school of thought mainly comes from econometrics, and over the last few decades has developed a number of ways in which meaningful causal estimates can be obtained without randomization. Economists were obtaining unreliable results based on correlational methods, so they decided to "take the con out of econometrics" [12] by developing better tools for causal inference. Some of these methods include the Regression Discontinuity Design [13, 14], the Difference-in-Differences approach [11] and Instrumental Variables [11]. These techniques are standard in economics yet are rarely used in many branches of behavioral and neuroscience research (although see Box 1 and Discussion for causal inference techniques already used in neuroscience).

Here we review these alternatives to randomization. We take published examples, and explain the methods. For each method, we then sketch how it could be used more widely across behavioral and neuroscience research using existing and emerging data. By systemically replacing correlational techniques with causal techniques, economics went through what they call a credibility revolution. Perhaps as a result, empirical work in economics has progressively overtaken theoretical work both in terms of citations within economics [15], and in terms of citations to economics papers made by articles in other fields [16].

There are subfields of neuroscience that aspire to causal inference from observational data. Network neuroscience, for example, seeks to identify connectivity (often termed *functional* or *effective* connectivity) between brain regions using recorded time series from a variety of modalities (e.g., fMRI, EEG, spike trains) [17-20]. This connectivity is sometimes interpreted as causal, but the validity of this interpretation depends on context, and *a priori* plausibility [21]. Importantly, the techniques listed below generally assume a lack of confounding by unobserved variables. In our view, this seems unlikely given the small number of observed signals and the high dimensionality of the brain. In this box, we review the most prominent techniques used in these fields. Note that these models can overlap, and that we have only presented the essence of each.

Granger causal models: The core intuition of Granger causality is that causes temporally precede effects [10]. A variable X (a time series) is said to Granger-cause Y (also a time series) if earlier values of X and Y predict Y better than earlier values of Y alone. I.e., if the history of X improves predictions of Y, this is evidence that X causes Y. Granger causality has been used extensively in network neuroscience [22, 23] and macroeconomics [24], but not without criticism [25] .

State-space models: This is a broad family of models [26, 27] in which one represents a system with one or more "state" variables to characterize "the way the system is". State variables may or may not be observed, typically evolve over time, and can be related to system inputs and output. For example, hippocampal neural activity could be a state variable that is affected/caused by experimental conditions and gives rise to (causes) an fMRI BOLD signal which is measured. States can be modeled as causally affecting one another as well [28, 29]. This family of models includes Dynamic Causal Modeling [28, 30], some types of point-process models [31], and others.

Structural equation models: This is a type of regression model with multiple equations [32]. There can be multiple dependent variables, and multiple independent variables. Dependent variables can also depend on *other* dependent variables. The dependencies between variables can, in some contexts, be given a causal interpretation. The parameters of these models are often found by regression approaches. Such models are also used in economics [33].

Bayesian networks: This is a type of model that includes variables and their statistical dependencies [1, 34]. In this framework, causality can be viewed as the probabilistic influence variable X has on variable Y after taking into account other variables in the network. It is said to be Bayesian because variables have prior and likelihood distributions, and other tools of Bayesian statistics can be used. This is also a broad family of models, and has been used in both network neuroscience [35, 36] as well as human causal learning [37-40].

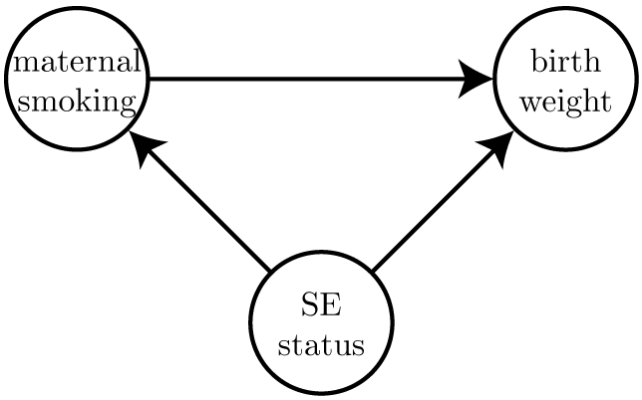
Box 1: Overview of causal inferences techniques already used in neuroscience.

When trying to infer causal effects, it is helpful to visually represent the variables under consideration and the relationships between them. Graphical models, widely used in computational

fields, provide a way to do this by representing each relevant variable as a circle (with a label), and each putative statistical relationship between two variables as a line connecting them. Causal relationships are represented by arrows which can be unidirectional or bidirectional. The variables included in a graphical model should be all relevant independent and dependent variables, as well as confounding variables which may affect the independent and dependent variables. Framed this way, the goal of causal inference becomes clearer: to estimate the strength and direction of the statistical relationships (lines/arrows) between variables after maximally accounting for the important components in the system. Actually estimating the statistical relationships depends on the specifics of the proposed model.

Importantly, we should not assume that the variables we use in a graphical model can be experimentally controlled. We should, therefore, distinguish between *observed* variables and *controlled (*or *set)* variables. Observed variables always have the possibility of being affected by unknown and unmodeled confounding variables, whereas controlled variables are immune to this problem. Following Pearl [1], we use the notation of *do(X)* to indicate an experimentally controlled variable, and the unqualified *X* to indicate a variable that is simply observed.

Consider an example addressed later in the text, in which we seek to find out whether maternal smoking affects a child's birth weight. To form a graphical model of this scenario, we would specifically model maternal smoking and birth weight as variables in the system. Because we believe that smoking may influence birth weight, we would draw an arrow that points from maternal smoking to birth weight. We would also want to account for confounding factors, like socioeconomic status; these factors may influence both a mother's smoking as well as birth weight. Socioeconomic status should also be specifically modeled in the graphical model, and we should draw arrows from socioeconomic status to both maternal smoking and birth weight. Furthermore, because ethically we cannot randomize maternal smoking, we cannot use the *do()* notation. The graphical model of this simplified system is shown below.



Box 2: Introduction to graphical models in causal reasoning

Regression Discontinuity Design

It is possible to approximate causality from a common property of decisions: thresholds for treatment. Treatment effects can often be estimated near thresholds, even without randomized experiments, because subjects near the threshold are similar. This approach was originally developed by Thistlethwaite, who was interested in the effect of academic recognition of student outcomes [13]. In that study, students with test scores above a certain threshold were given Certificates of Merit and public recognition. The students who received the certificate were clearly different from those that did not, e.g. in intelligence and socioeconomic status, so it was not possible to simply ask how Certificates of Merit affected outcomes. However, as we approach the threshold score from either side, the students will become arbitrarily similar. This is because there is randomness in the exact score a student received due to e.g., question selection, or their last night's sleep. Thistlethwaite found that Certificates of Merit led to more future scholarships, but not to differences in long-term career plans [13]. This strategy, which is based on the idea that samples just above and just below the threshold are nearly indistinguishable, is the basis of the regression discontinuity design.

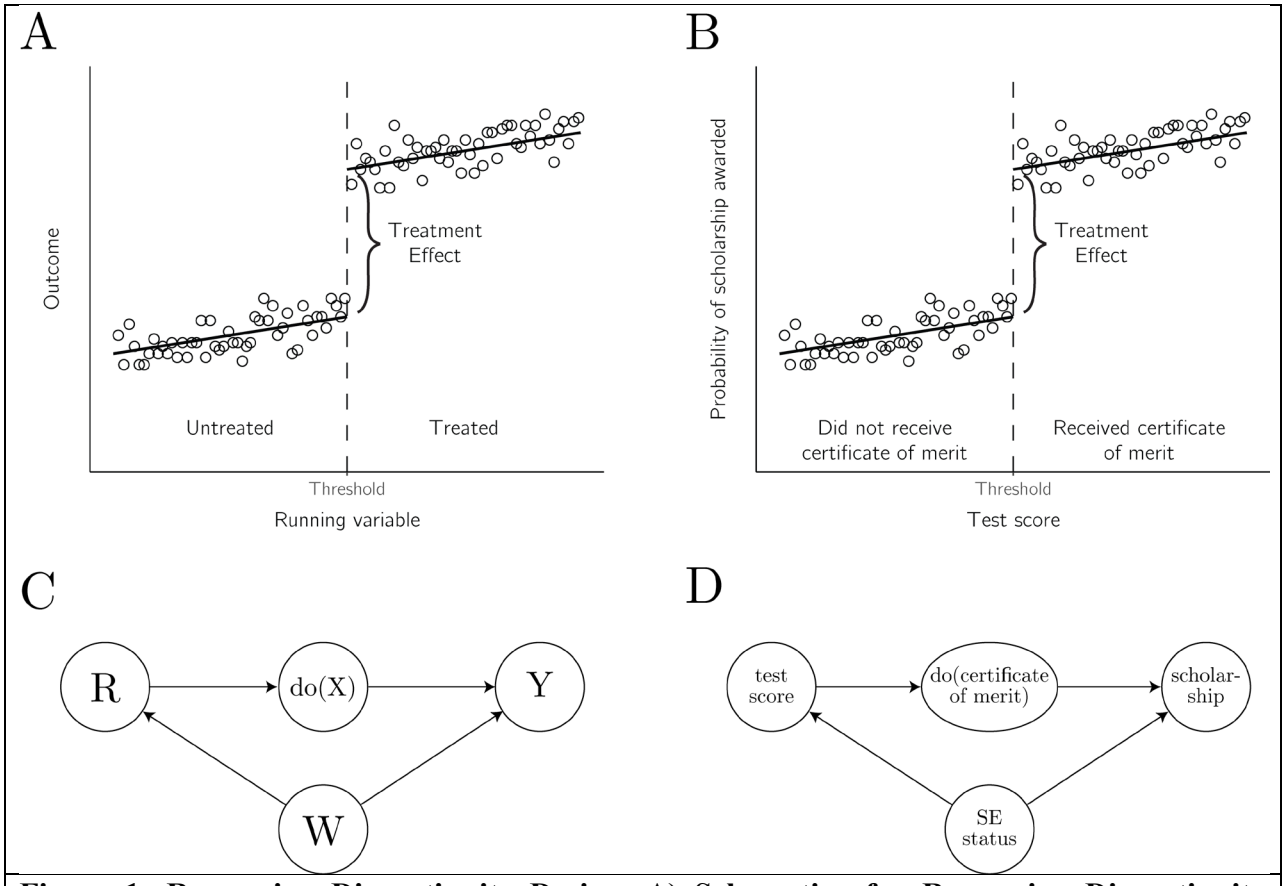


Figure 1: Regression Discontinuity Design. A) Schematic of a Regression Discontinuity Design analysis. The treatment is only administered if the running variable is above the threshold. The outcome (y-axis) is plotted as a function of a running variable (x-axis). The magnitude of the treatment effect, the difference in outcome at the threshold, is estimated using regression. **B) Schematic figure representing the analysis performed** in [13]. Academic outcome (probability

of scholarship) is plotted as a function of test score, and a discontinuity is seen at the cutoff for receiving a certificate of merit. Note that this figure is stylized and does not use the data used in the original analysis; it is intended only to demonstrate the approach. **C) Graphical model of Regression Discontinuity Design.** W are confounding variables; R is the running variable which determines the treatment along with the threshold; X is the treatment (independent variable) which is either administered (do(X)) or not administered (do(not X)) depending on R; and Y is the outcome (dependent variable) of interest. **D) Graphical model representing this analysis**. Socioeconomic status (for example) is likely to affect both test score and the probability of receiving a scholarship. Test score determines whether a certificate of merit is awarded, which in turn affects the probability of receiving a scholarship.

To perform the analysis, a regression of outcome as a function of the running variable (e.g. test scores in the study by Thistlethwaite) is fit on both sides of the threshold. A causal effect would be manifested by a discontinuity between the regression line on the left and on the right of the threshold (see Fig. 1). This discontinuity in the outcome can only be from the treatment because no confounder is likely to have a discontinuity at exactly the same threshold (although it is standard to check this assumption). The RDD gives a meaningful and often unbiased estimate of the causal effect of the treatment in the vicinity of the threshold [14].

Area	Question	Running variable	Threshold	Outcome variable
Education	How much does enrichment help?	Test scores used for enrichment program	Minimum test score	Education outcome, income
Medicine	How much does blood pressure medication help?	Blood pressure	Treatment guideline	Death by cardiovascular disease
Counseling	How many people should receive depression treatment?	Risk score	Enrollment threshold	Mental health
Advertising	How much does an advertisement affect consumer behavior?	Affinity score	Money limit	Sale of product
Neural data science	What are the neural requirements for movement?	Neural drive	Firing threshold	Activity of a downstream neuron or muscle
Neural theory	How much would a larger synaptic weight increase reward-seeking behavior?	Neural drive	Firing Threshold	Behavioral change

Table 1: Possible applications of RDD in neuroscience and behavioral research

It is important to be aware of the conditions needed for causal validity in RDD. Subjects must not be able to precisely control their score – and thus their treatment – e.g. by working long enough hours to achieve exactly the score that will put them over the threshold. A test for this assumption was developed by [41], which looks for a discontinuity in the number (i.e., density) of subjects on either side of the threshold. Moreover, subjects must not be able to override the threshold mechanism for selection. Sometimes so-called fuzzy RDD approaches can deal with the problem of treatments not being perfectly administered [42]. Best practices for implementing RDD can be found in [14, 43]. Importantly, these methods allow checking whether the assumptions underlying RDD are valid. For example, potential confounders should not have a discontinuity at the threshold. While there are many statistical issues to consider for the RDD, there is an active community of practitioners furthering our already-strong understanding.

RDD should allow us to discover causal effects in many domains (see table 1). Thresholds exist widely in human behavior and neuroscience, and there are very few variables that subjects can noiselessly control. For example, in the field of neural theory, we might ask how neurons can estimate their causal effect on animal performance which would allow asking if a larger weight would be better. The translation of neural drive to spiking has a firing threshold, which could allow neurons to estimate their causal effect [44]. There are countless possibilities to expand on the small set of current applications, e.g., [13, 45]. RDD thus promises to be useful for nearly every sub-field of behavioral research and neuroscience, ranging from education, medicine all the way to neural theory.

Difference-in-Differences

Another approach for approximating causality is to look for temporal trends in treated versus untreated subjects, even if they were not randomly assigned. The core idea of this approach is to use longitudinal data for two groups where only one is treated, but where the two groups are similarly affected by extraneous factors. For example, [46] investigated the effect of academic year length on student outcomes. It exploited a transient reduction in school year length which occurred in some but not all German states. Thus, it was possible to compare outcomes between short-school-year states (the treatment group) and a regular-school-year state (the control group) to measure the effect of the shortened school year. Grade repetition increased in the short-school-year states relative to the regular-school-year state after the short school year was introduced. The length of the school year was thus found to have a causal effect on repeating grades.

To perform the analysis, the temporal evolution of the outcome is measured for both the treated and the untreated group. This, in a way, generalizes the idea of baseline-controlled or two-factorial designs sometimes used in clinical trials. A quantification of the temporal difference between the treated and the untreated group then allows estimating the treatment effect (see Fig. 2).

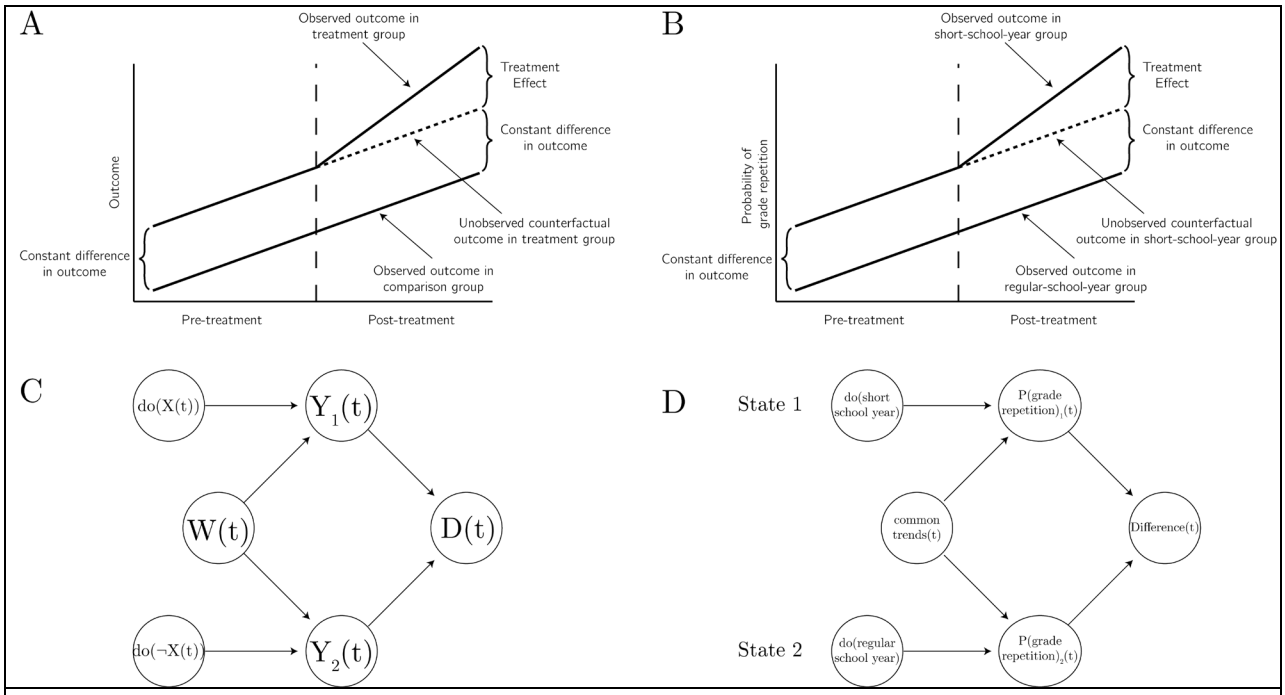


Figure 2: A) Schematic of a Difference-in-Differences analysis. The trend of two groups, treated and untreated, is plotted as a function of time. Before the treatment, the trends of the two groups should be parallel (a constant difference-in-differences). The treatment effect is estimated by the degree to which the trends diverge after the treatment is administered. **B) Schematic figure representing the analysis performed** in [46]. Outcome (probability of grade repetition) is plotted as a function of time, before and after the implementation of the short school year in some states. The difference between state outcomes changes after the change in school year (i.e., there is an increase in difference in differences). Note that this figure is stylized and does not use the data used in the original analysis; it is intended only to demonstrate the approach. **C) Graphical model for Difference-In-Differences.** All variables are considered as a function of time, t. W are confounding variables; X is the treatment (independent variable) which is administered (do(X)) to population 1, and not administered (do(not X)) to population 2; Y1 and Y2 are the outcomes (dependent variables) for populations 1 and 2, respectively; D is the difference between Y1 and Y2 and is tracked over time**. D) Graphical model representing the analysis performed**. Common trends such as federal taxes and economic conditions are likely to affect the two states similarly. The short school year is implemented only in one state. The difference in outcome is calculated from the two states' outcomes.

The Difference-in-Differences (DiD) approach naturally comes with its own assumptions and caveats, many of which we can explicitly test. Most importantly, it assumes that the two groups are chosen such that they are similarly affected by relevant and perhaps unmeasured factors: this is the common trends assumption. In the above example, the two groups of German states should be similarly affected by the economic context, other policy changes, etc. One way to provide support for this assumption is to check that trends in the outcome prior to the new treatment are parallel. The groups should also be stable in composition (e.g. percentage of women in each group) over the period of

comparison. Extensions such as nonlinear DiD have also been developed [47]. Best practices for DID can be found in [11].

Field	Question	Comparison	Outcome measured over time
Educational policy	How do smartphones affect middle school students?	Two nearby school districts before and after a new smartphone policy change	Standardized test scores, disciplinary action
Rehabilitation	How well does rehab work?	Affected limb and unaffected limb before and after rehab	Strength, coordination scores
Neurology	What are the effects of new brain lesions in MS?	before and after unilateral lesion	Strength, coordination scores
Public relations	Do ads have negative side effects?	In versus outside of target area, before and after the start an ad campaign	Attitudes towards ads

Table 2: Possible applications of Difference-In-Difference approaches

DiD approaches should also be broadly applicable in behavioral science and neuroscience (see table 2). Many, if not most, variables in neuroscience and behavior are measured over time. And many interventions affect some people or neurons (the treatment group) but not others (the control group). DiD thus promises to be useful across most sub-disciplines that deal with behavior.

Instrumental Variables

A third common approach for quasi-experimental causal inference is Instrumental Variables (IV) [48]. With this approach, we seek to identify variables Z ("instruments") that causally affect the independent variable of interest X, but only causally affect the dependent variable Y through X (Fig. 3). For example, [49] sought to ask how maternal smoking affects birth weight. We should expect heavy confounding as e.g., low socioeconomic status may affect both smoking and health. Instead, the authors leveraged tobacco taxes as an instrument, which arguably affects smoking but does not directly affect birth weight. Differences in tobacco taxes across years and across states could then be exploited to estimate the causal effect of smoking on birth weight. They found that maternal smoking decreased birth weight by between 300 and 600 grams.

To perform an IV analysis, we first identify the independent and dependent variables. Next we find, through an understanding of the system, another variable that can serve as an instrument (taxes in the above example) that only affects the independent variable. Next, we build a predictive model of the treatment X based on the instrument Z (first stage regression). And then we use this prediction in a second stage regression to quantify the causal effect of treatment (e.g. smoking) on the dependent variable (birth weight in the above example). The essence of this approach is that it identifies changes in the dependent variable that occur as a result of varying the independent variable; the instrument can be viewed as a rudimentary experimental manipulation of the independent variable. Simply regressing the dependent variable on the independent variable would be confounded by e.g.,

socioeconomic status in the above example. Such IV approaches allow yet one more way to meaningfully deal with unobserved confounders.

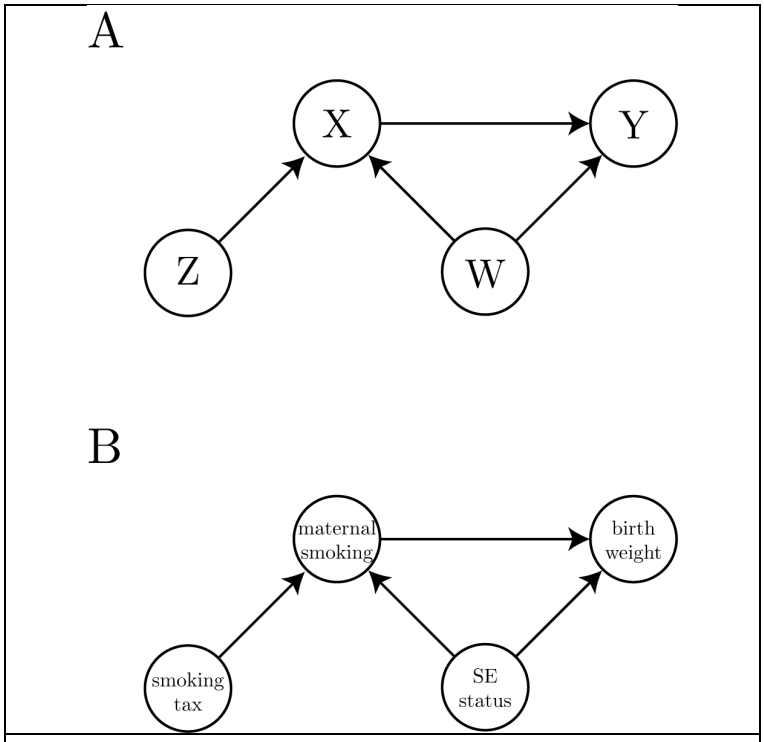


Figure 3: A) Graphical model for Instrumental Variables. W are confounding variables; X is the independent variable; Y is the outcome (dependent variable); Z is the instrument which only affects Y through its effect on X. **B) Graphical model representing the analysis performed.** Graphical model representing this analysis performed in [49]. Maternal smoking is thought to affect birth weight. But socioeconomic status (for example) likely affects both a mother's decision to smoke as well as the child's birth weight. A tax on cigarette smoking could affect maternal smoking but is unlikely to directly influence the birth weight, except through an effect on maternal smoking. Such a tax is therefore a good instrument to examine the effect of smoking on birth weight without being confounded by socioeconomic status.

The IV approach also has its assumptions and caveats. The most important assumption is the exclusion restriction: the instrument (e.g., tobacco taxes) should affect the dependent variable (e.g., birth weight) only through its effect on smoking. I.e., we should be able to exclude that the instrument affects the outcome other than through the independent variable. The exclusion restriction is not directly testable and must therefore be assessed on plausibility grounds given what we know about

the phenomenon at hand. Furthermore, the instrument should not be too weakly correlated with the independent variable of interest, in order to produce useful estimates. The F test for the first stage can measure how strong the instrument is [50]. Best practices and further discussion of IV can be found in [11].

Medicine	Does a medication affect patient outcomes?	Medication use	Patient outcome	Hospital rules about pharmaceutical reps
Neuroscience	How does brain region A affect region B?	Brain region A activity	Brain region B activity	Diffuse optogenetic stimulation of brain region A
Behavioral health	Does alcohol consumption make you a bad parent	Alcohol consumption	Parenting license exam	Alcohol taxes
Education	Does having more disposable income improve educational outcomes?	Income	Educational outcome	Income tax cut

Table 3: Possible applications of Instrumental Variable approach

Instrumental variable approaches should also be broadly applicable in behavioral science (see table 3). Typical experiments have many variables that are not experimentally randomized, yet affect the system. For example, metabolism affects neural activity which affects behavior; markers of metabolism could therefore be viewed as instruments to ask how neural activity gives rise to behavior. Many such variables are random with respect to behavior, and could just as easily be viewed as instruments. Even standard techniques like optogenetics may be better viewed as instruments [51]. Optogenetics does not precisely set neural activity; it only affects it. Hence, it may be useful to view optogenetic stimulation of brain region X as an IV, and then use that model to ask how brain region X affects region Y. More generally, it may be possible to create biological constructs with IV-like properties; e.g., a molecular construct that inactivates individual neurons at random times. In summary, Instrumental Variables are an approach to get good results by using existing, non-experimental randomization.

Discussion

Here we have argued that understanding causal effects is the goal of the bulk of both behavioral science and neuroscience, and that these fields need to adopt better techniques for making causal inferences. We have reviewed three prominent quasi-experimental approaches developed in economics, explained their application, and suggested ways that they might be applied in a number of examples using already-existing data. These techniques promise to move our data analysis towards a causal understanding. We chose three particular techniques – Regression Discontinuity Design, Difference-In-Differences, and Instrumental Variables – but many other techniques for estimating

causal influence exist. For example, Bayesian networks [34] and Structural Equations [32] can be used to model networks of relevant variables and to estimate causal relationships between them. Propensity Score Matching estimates causal effects between treated and untreated subjects by adjusting for observed confounders that predict treatment [52-54]. Other techniques use noise distributions to estimate the direction of causal influence [55]. Synthetic controls may be a valuable way to construct better control/comparison groups for case control and DiD studies [56]. In this Perspective, however, we focused on quasi-experimental approaches as these approaches readily allow dealing with unobserved confounders that make causal inference difficult in behavioral science as well as neuroscience.

Neuroscience and psychology do have a history of using techniques that attempt to recover causal influences from data. Network neuroscience has used a large suite of approaches (see Box 1) with the goal of deciphering the complex networks of the brain. This is an important line of work, but whether these techniques actually recover true causal effects in this context remains an open question [21, 25]. Many challenges exist such as possible omitted variables, and the difficulty of modeling information transformation between brain regions [18, 21]. Although some of the techniques we present could be applied towards this purpose, we view RD, DiD, and IV as more general approaches that exploit different aspects of data than network approaches. RD exploits thresholds, which is clearly different from the network approaches. DiD exploits common trends even when confounders are not always identifiable, whereas network approaches generally are sensitive to omitted confounders [35]. IV identifies non-experimental randomization, and although it could be incorporated into network approaches, we believe that its use is far more general. We therefore believe that the techniques discussed in this paper widen the scope of data available for causal analysis in neuroscience and behavioral science.

Techniques for quasi-experimental causal inference are ripe for application in behavioral science and neuroscience. They could fruitfully be applied to existing laboratory data, such as neuroimaging, virtual reality behavior, or neural spike recordings. This may allow us to extract more valuable information from these data. But these techniques also make it possible to perform credible analyses of the kind of observational data offered by the information age [57] that is much cheaper and much more common than laboratory data. Thresholds exist everywhere, in online systems [58], economic activity (e.g. tax notches, see [59]) but especially in medicine [45], making RDD, with its clean treatment of confounders, an invaluable tool. Wherever parallel trends exist, DiD promises to give our analyses better controls. And identifying valid instruments on independent variables of interest will help us to tease apart causal relationships with IV. Every neuroscientist and behavioral scientist should become familiar with these techniques.

In the 1980s it became obvious in economics that the typical correlational findings were not overly indicative of real causal effects. This led the field of econometrics to decide to work towards methods that allow the quantification of causality [60]. In the following decades, causal inference improved massively and today the bulk of top economics papers uses standard causal inference strategies [60].

Neuroscience and behavioral science have the same problem: we write stories about causality in behavior and brains based on correlational data, whereas we need techniques that can reliably demonstrate causal effects. Many techniques currently used in neuroscience may actually be misleading us [25] because we misunderstand whether they measure causal effects. Furthermore, a

focus on causal effects should help us to focus on which effects are worth caring about, such as behavior [61]. This understanding of the deeply problematic basis of this kind of inference is slowly taking hold in the community. To lead to a deeper understanding of minds and brain, we need to take the causal questions seriously. We can only do so by applying techniques that allow answering those questions.

References

1. Pearl J. Causality: Cambridge university press; 2009.
2. Notorious B, Combs S, Mase. Mo Money Mo Problems: Bad Boy Records; 1997.
3. Grodstein F, Manson JE, Colditz GA, Willett WC, Speizer FE, Stampfer MJ. A prospective, observational study of postmenopausal hormone therapy and primary prevention of cardiovascular disease. Annals of Internal Medicine. 2000;133:933-41. doi: 10.7326/0003-4819-133-12-200012190-00008. PubMed PMID: 11119394.
4. Manson JE, Hsia J, Johnson KC, Rossouw JE, Assaf AR, Lasser NL, et al. Estrogen plus progestin and the risk of coronary heart disease. The New England Journal of Medicine. 2003;349:523-34. doi: 10.1056/NEJMoa030808. PubMed PMID: 12904517.
5. Humphrey LL, Chan BK, Sox HC. Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. Annals of Internal Medicine. 2002;137:273-84. PubMed PMID: 12186518.
6. Greenland S. Randomization, statistics, and causal inference. Epidemiology. 1990:421-9.
7. Ismail-Beigi F, Craven T, Banerji MA, Basile J, Calles J, Cohen RM, et al. Effect of intensive treatment of hyperglycaemia on microvascular outcomes in type 2 diabetes: an analysis of the ACCORD randomised trial. The Lancet. 2010;376(9739):419-30.
8. Officers TA. Major Outcomes in High-Risk Hypertensive Patients Randomized to or Calcium Channel Blocker vs Diuretic. Journal of the American Medical Association. 2002;288:2981-97. doi: 10.1001/jama.288.23.2981. PubMed PMID: 12479763.
9. Group SR. A randomized trial of intensive versus standard blood-pressure control. New England Journal of Medicine. 2015;373(22):2103-16.
10. Granger CW. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society. 1969:424-38.
11. Angrist JD, Pischke J-S. Mostly harmless econometrics: An empiricist's companion: Princeton University Press; 2008.
12. Leamer EE. Let's Take the Con Out of Econometrics. The American Economic Review. 1983;73:31-43.
13. Thistlethwaite DL, Campbell DT. Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment. The Journal of Educational Psychology. 1960;51:309-17. doi: 10.1037/h0044319. PubMed PMID: 13149141.
14. Imbens GW, Lemieux T. Regression discontinuity designs: A guide to practice. Journal of Econometrics. 2008;142:615-35. doi: 10.1016/j.jeconom.2007.05.001. PubMed PMID: 25246403.
15. Angrist J, Azoulay P, Ellison G, Hill R, Lu SF. Economic Research Evolves: Fields and Styles. American Economic Review. 2017;107(5):293-97.
16. Angrist J, Azoulay P, Ellison G, Hill R, Lu SF. Inside job or deep impact? Using extramural citations to assess economic scholarship. National Bureau of Economic Research, 2017.
17. Pillow JW, Shlens J, Paninski L, Sher A, Litke AM, Chichilnisky EJ, et al. Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature. 2008;454(7207):995-9. doi: 10.1038/nature07140.
18. Valdes-Sosa PA, Roebroeck A, Daunizeau J, Friston K. Effective connectivity: Influence, causality and biophysical modeling. NeuroImage. 2011;58:339-61. doi: 10.1016/j.neuroimage.2011.03.058. PubMed PMID: 21477655.
19. Stevenson IH, Körding KP. On the similarity of functional connectivity between neurons estimated across timescales. PloS one. 2010;5(2):e9206-e. doi: 10.1371/journal.pone.0009206.
20. Sakkalis V. Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG. Computers in biology and medicine. 2011;41(12):1110-7.
21. Ramsey JD, Hanson SJ, Hanson C, Halchenko YO, Poldrack RA, Glymour C. Six problems for causal inference from fMRI. NeuroImage. 2010;49:1545-58. doi: 10.1016/j.neuroimage.2009.08.065. PubMed PMID: 19747552.

22. Bressler SL, Seth AK. Wiener–Granger causality: a well established methodology. Neuroimage. 2011;58(2):323-9.
23. Ding M, Chen Y, Bressler SL. 17 Granger causality: basic theory and application to neuroscience. Handbook of time series analysis: recent theoretical developments and applications. 2006;437.
24. Hiemstra C, Jones JD. Testing for linear and nonlinear Granger causality in the stock price-volume relation. The Journal of Finance. 1994;49(5):1639-64.
25. Jonas E, Kording KP. Could a neuroscientist understand a microprocessor ? 2016;XXI:1-5.
26. Chen Z. Advanced State Space Methods for Neural and Clinical Data: Cambridge University Press; 2015.
27. Shumway RH, Stoffer DS. State-Space Models. Time series analysis and its applications: Springer; 2011. p. 319-404.
28. Friston KJ, Harrison L, Penny W. Dynamic causal modelling. NeuroImage. 2003;19(4):1273-302. doi: 10.1016/S1053-8119(03)00202-7.
29. Semedo J, Zandvakili A, Kohn A, Machens CK, Byron MY, editors. Extracting latent structure from multiple interacting neural populations. Advances in neural information processing systems; 2014.
30. Daunizeau J, David O, Stephan KE. Dynamic causal modelling: A critical review of the biophysical and statistical foundations. NeuroImage. 2009;58(2):312-22. doi: 10.1016/j.neuroimage.2009.11.062.
31. Latimer KW, Yates JL, Meister MLR, Huk AC, Pillow JW. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. Science. 2015;349(6244):184-7. doi: 10.1126/science.aaa4056.
32. Ullman JB, Bentler PM. Structural equation modeling. Handbook of Psychology, Second Edition. 2012;2.
33. Nevo A, Whinston MD. Taking the dogma out of econometrics: Structural modeling and credible inference. Journal of Economic Perspectives. 2010;24(2):69-82.
34. Koller D, Friedman N. Probabilistic graphical models: principles and techniques: MIT press; 2009.
35. Smith SM, Miller KL, Salimi-Khorshidi G, Webster M, Beckmann CF, Nichols TE, et al. Network modelling methods for FMRI. NeuroImage. 2011;54(2):875-91. doi: 10.1016/j.neuroimage.2010.08.063.
36. Song L, Kolar M, Xing EP, editors. Time-varying dynamic Bayesian networks. Advances in neural information processing systems; 2009.
37. Goodman ND, Ullman TD, Tenenbaum JB. Learning a theory of causality. Psychological review. 2011;118:110-9. doi: 10.1037/a0021336. PubMed PMID: 21244189.
38. Gopnik A, Sobel DM, Danks D, Glymour C, Schulz LE, Kushnir T. A Theory of Causal Learning in Children: Causal Maps and Bayes Nets. Psychological Review. 2004;111:3-32. doi: 10.1037/0033-295X.111.1.3. PubMed PMID: 14756583.
39. Gopnik A, Tenenbaum JB. Bayesian networks, Bayesian learning and cognitive development. Developmental science. 2007;10(3):281-7.
40. Körding KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L. Causal inference in multisensory perception. PLoS one. 2007;2(9):e943.
41. McCrary J. Manipulation of the running variable in the regression discontinuity design: A density test. Journal of Econometrics. 2008;142:698-714. doi: 10.1016/j.jeconom.2007.05.005.
42. Trochim WM. Research design for program evaluation: The regression-discontinuity approach: Sage Publications, Inc; 1984.
43. Jacob RT, Zhu P, Somers M-A, Bloom H. A Practical Guide to Regression Discontinuity. Mdrc. 2012:1-100.
44. Lansdell B, Kording K. Spiking allows neurons to estimate their causal effect. bioRxiv. 2018:253351.
45. Moscoe E, Bor J, Bärnighausen T. Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: A review of current and best practice. Journal of Clinical Epidemiology. 2015;68:122-33. doi: 10.1016/j.jclinepi.2014.06.021. PubMed PMID: 1643436087.
46. Pischke JS. The impact of length of the school year on student performance and earnings: Evidence from the German short school years. Economic Journal. 2007;117:1216-42. doi: 10.1111/j.1468-0297.2007.02080.x.
47. Athey S, Imbens GW. Identification and inference in nonlinear difference-in-differences models. Econometrica. 2006;74:431-97. doi: 10.1111/j.1468-0262.2006.00668.x. PubMed PMID: 18563856.
48. Angrist JD, Imbens GW, Rubin DB, Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association. 2016;91:444-55.
49. Evans WN, Ringel JS. Can higher cigarette taxes improve birth outcomes? Journal of Public Economics. 1999;72:135-54. doi: 10.1016/S0047-2727(98)00090-5. PubMed PMID: 491319.
50. Stock JH, Yogo M. Testing for weak instruments in linear IV regression. National Bureau of Economic Research Cambridge, Mass., USA; 2002.

51. Li X, Yamawaki N, Barrett JM, Kording KP, Shepherd GM. Scaling of optogenetically evoked signaling in a higher-order corticocortical pathway in the anesthetized mouse. bioRxiv. 2018:154914.
52. Dehejia RH, Wahba S. Propensity score-matching methods for nonexperimental causal studies. Review of Economics and statistics. 2002;84(1):151-61.
53. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.
54. Imbens GW, Rubin DB. Causal inference in statistics, social, and biomedical sciences: Cambridge University Press; 2015.
55. Hoyer PO, Janzing D, Mooij JM, Peters J, Schölkopf B, editors. Nonlinear causal discovery with additive noise models. Advances in Neural Information Processing Systems; 2009.
56. Abadie A, Diamond A, Hainmueller J. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. Journal of the American Statistical Association. 2010;105:493-505. doi: 10.1198/jasa.2009.ap08746. PubMed PMID: 741578133.
57. Kwak H, Lee C, Park H, Moon S, editors. What is Twitter, a social network or a news media? Proceedings of the 19th International Conference on World Wide Web; 2010: ACM.
58. Bem J. Using match confidence to adjust a performance threshold. Google Patents; 2008.
59. Slemrod J. Buenas notches: lines and notches in tax system design. eJournal of Tax Research. 2013;11(3):259.
60. Angrist JD, Pischke J-S. The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. Journal of Economic Perspectives. 2010;24:3-30. doi: 10.1257/jep.24.2.3.
61. Neuroscience Needs Behavior: Correcting a Reductionist Bias, (2017).