HHS Public Access

Author manuscript

Nat Comput Sci. Author manuscript; available in PMC 2022 June 04.

Published in final edited form as:

Nat Comput Sci. 2021 January ; 1(1): 24–32. doi:10.1038/s43588-020-00005-8.

Quantifying causality in data science with quasi-experiments

Tony Liu[1], **Lyle Ungar**[1], **Konrad Kording**[2,3,✉]

[1]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA.

[2]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA.

[3]Department of Neuroscience, University of Pennsylvania, Philadelphia, PA, USA.

Abstract

Estimating causality from observational data is essential in many data science questions but can be a challenging task. Here we review approaches to causality that are popular in econometrics and that exploit (quasi) random variation in existing data, called quasi-experiments, and show how they can be combined with machine learning to answer causal questions within typical data science settings. We also highlight how data scientists can help advance these methods to bring causal estimation to high-dimensional data from medicine, industry and society.

Although much of data science has centred around using tools from statistics and machine learning (ML) to make predictions and 'extract insight' from data[1], many questions are ultimately causal in nature as the predictions made and insights gained are used to inform decision-making. For example, data scientists may build a ML model to make predictions of which patients are likely to be re-admitted to a hospital within 30 days, but stake-holders ultimately want to know what interventions or policy decisions can be made to reduce the risk of a patient being re-admitted. Data scientists need to add causal methods to their toolkit to answer the causal questions embedded in many of the problems they aim to solve.

As correlation is not sufficient for causation, there is danger in misinterpreting the correlational results common in data science analysis as causal. Any typical supervised ML method (for example, predictions from a set of features) is inadequate for inferring causation, as training ML models for prediction is an inherently correlational task: the ML model only 'observes' the relationship between the features and outcome in the data when making predictions, instead of 'changing' the features to determine their impact on the outcome[2,3]. An example of these shortcomings was a model developed to predict mortality

✉**Correspondence** should be addressed to K.K. koerding@gmail.com.

Author contributions

T.L. helped write and prepare the manuscript. L.U. and K.K. jointly supervised this work and helped write the manuscript. All authors discussed the structure and direction of the manuscript throughout its development.

Competing interests

The authors declare no competing interests.

Code availability

We provide interactive widgets of Figs. 2–4 in a Jupyter Notebook hosted in a public GitHub repository (https://github.com/tliu526/causal-data-science-perspective) and served through Binder (see link in the GitHub repository).

in patients with pneumonia that paradoxically concluded that having asthma reduced pneumonia mortality risk[4]. The model did nothing wrong: patients in the training data with asthma were more likely to be admitted to the hospital and were administered more aggressive care, ultimately resulting in lower mortality. However, this model is potentially harmful from a decision-making perspective as following the model may deny patients with asthma the extra care they actually need. With standard ML there is always a risk that unobserved features may be driving the model predictions.

Numerous techniques have been developed to estimate causality more appropriately from observational data (where there are no randomized perturbations). Here we review standard quasi-experimental methods, which have been largely used in economics but underutilized outside of this domain. We discuss their assumptions, why we believe that they should be more widely adopted and motivate our explanations with real-world examples along with potential applications in common data science domains. We close by advocating for the cross-pollination of quasi-experimental methods and data science: quasi-experiments can make causal inference possible in typical data science settings, while innovations in ML can in turn improve these methods for wider application in complex data domains.

Confounders and colliders in observational studies

Observational data can be problematic for causal inference because of the core problem of unobserved confounding. We wish to determine the effect of a treatment X (hospitalization) on outcome Y (pneumonia risk), but there is a hidden confounder Z (asthma) that simultaneously influences both X and Y. If we can conduct an experiment, that is, set X ourselves, then we can remove confounding, allowing identification of the causal effect of X on Y (Fig. 1a). Experiments are used across medicine (clinical trials), public health (Oregon Medicaid lottery experiment[5]), software development and advertising (A/B testing), and ML (bandits and reinforcement learning[6–8]). However, experiments can be expensive, unethical or otherwise impractical. Furthermore, most data in data science settings are observational, yet we still would like to estimate causal effects. To infer causality from observational data, untestable assumptions about the data-generating process are usually needed.

Many techniques that estimate causality from observational data rely on unconfoundedness, which is the assumption that all confounders (variables that influence both the treatment X and outcome Y) are observed and accounted for in the data[9]. In this case, the confounders can be ‘controlled for’ to obtain valid causal estimates, for example, by adding them to a regression. However, we cannot generally know whether we capture all confounders[10], and so strong beliefs about the system are needed (Fig. 1b). Practitioners must argue that their data do in fact contain all relevant confounders (and none of the colliders, see below), and critiques of analyses relying on unconfoundedness often centre on variables missing from analysis. Because there is no way of testing whether the missing variables are confounders, belief of the causal estimates hinges on belief that the unconfoundedness assumption is credible.

For example, suppose that in our pneumonia risk problem, we measure patients’ asthma and control for it in our analysis. A reviewer of the study could raise that there are numerous

other potential confounders that were not accounted for, such as race, socioeconomic status, smoking status, other medical history and so on. And even if more of these potential confounders were measured as part of the study and controlled for, there is always the possibility of another missing confounder that could threaten the validity of the results. Any study making causal claims using observational data needs to either use methods that do not rely on unconfoundedness or provide a convincing argument that unconfoundedness is at least approximately true.

In addition to unconfoundedness, data scientists must also be able to identify variables that should not be controlled for, which are called colliders. Unlike confounders, which affect both X and Y, colliders are variables that are affected by X and Y (Fig. 1b). Controlling for colliders makes our causal estimates worse[11,12]. For example, we can imagine a possibility in our pneumonia question where asthma itself is a collider of hospitalization (X) and pneumonia (Y): pneumonia may make it more likely for a hospitalized patient to experience asthmatic symptoms, perhaps due to allergies encountered during hospitalization. In addition to measuring relevant variables to control for confounding, data scientists seeking causal estimates need to ensure that the variables are not colliders.

One framework of causality that is prominent in computer science is the structural causal model (SCM) popularized by Judea Pearl[11]. SCMs combine directed acyclic graphs (DAGs) to express causal relationships with a mathematical framework for computing causal quantities from the DAGs called do-calculus. The beauty of SCMs is that the causal structure of a problem is represented graphically, making both assumptions and identifiability of causal effects easy to understand[13]. Expert knowledge is needed to obtain these graphs and to argue that there are no confounders missing in the assumed form[3]. Though we can test the correctness of connections between variables in the graph to some degree by checking whether the implied conditional independencies between variables is reflected in the data[2], unconfoundedness is generally untestable. Unless there is a strong belief in the accuracy of the assumed graphical structure[14], there will always be the threat of unobserved confounders biasing causal estimates.

There are numerous other techniques that aim to estimate causal effects from observational data, though they often rely on unconfoundedness or some variant of unconfoundedness. Matching and inverse probability weighting, which are popular in social science settings, can be used to control for confounding bias in the causal estimates[12,15]. Double ML methods convert treatment effect estimation into two ML prediction tasks, allowing for causal inference in high-dimensional data[16]. There is also a rich field on causal discovery[17] with recent active ML research in this area[18–20], where the goal is to infer the causal graph from the data itself, often using unconfoundedness as an assumption. Latent variable modeling has also been applied to observational causal inference, which does not require unconfoundedness but does require specific assumptions about the independence and number of causes affecting the outcome[21]. Because unconfoundedness is often difficult to justify in real-world problems, we choose to highlight quasi-experimental strategies here.

Quasi-experimental strategies

A framework of observational causality that has largely been developed in the field of economics, called quasi-experiments, leverages randomness occurring naturally in observed data to estimate causal effects. After undergoing a 'credibility revolution' over the past few decades[22,23], economists have increasingly employed quasi-experimental techniques to estimate causal effects in real-world problems. These methods also rely on assumptions about the causal structure of the data, but these assumptions can be more plausible than unconfoundedness, as we will discuss. The following techniques we review leverage naturally occurring randomness to estimate causal effects.

Instrumental variables.

The first technique we will examine that uses this idea is instrumental variable (IV) estimation[24]. The goal is to identify observable variables, instruments, that affect our system only through their influence on the treatment of interest X (Fig. 2a). In other words, we are not asking for the system to be unconfounded but for a component to be unconfounded, which we then use to estimate causality.

A famous example of IV analysis in economics considers the effect of the amount of required schooling (X) on an individual's future wages (Y), using birth season (Z) as an instrument[25]. In many US states, children are required to enter school the calendar year they are age six, so individuals born later in the year are young for their school grade. Because state laws require school attendance until a particular age, for example, 16 years old, individuals are required to be in school for different amounts of time because of their birth season. As long as we can assume that birth season is effectively random and only affects an individual's future wages through the amount of required schooling they receive, birth season can be used as an instrument to estimate causal effects.

To obtain causal estimates using IV analysis, the treatment (required schooling), outcome (future wages) and instrument (birth season) must be identified, with the assumption that the instrument only affects the outcome through its effect on the treatment. We then commonly perform what is known as two-stage least squares[26] (alternative IV estimation frameworks have also been proposed[27]), regressing the treatment on the instrument in the first stage ($\hat{X} = \alpha \mathrm{IV}$, where α is the fitted coefficient) and then regressing the outcome on the estimate of the treatment from the first stage ($\hat{Y} = \beta\hat{X}$, where β is the fitted coefficient). The first stage extracts the 'unconfounded component' of the treatment due to the effect of the instrument and uses that component in our second stage regression to estimate the causal effect on the outcome, controlling for confounding in the process: Fig. 2b depicts simulated data where the true treatment effect of X on Y is confounded, and we see that IV gives unbiased causal estimates even when confounding is present. For our motivating example from the introduction, we could use the distance that patients with pneumonia live from a hospital as an IV[28] (as hospital distance is plausibly random).

There are assumptions and considerations to be made when performing IV analysis to ensure valid estimation. We need to assume that there is no confounding variable influencing both the instrument itself and the outcome Y (Fig. 2a), which is similar to the unconfoundedness

assumption but is often more justifiable because the instrument is chosen to be random: it is more plausible to argue that birth season is unconfounded with future wages than it is to argue that our treatment X schooling is unconfounded with future wages. We also need to ensure the instrument affects the system only through its effect on the treatment, known as exclusion restriction (Fig. 2a): Fig. 2c depicts simulated data where the true treatment effect of X on Y is confounded and where the exclusion restriction is violated, and we see that neither IV nor regression give correct causal estimates. Another key assumption that IV analysis requires is the monotonicity or no defiers assumption: no individual units in our data sample can respond in the opposite direction of treatment uptake when affected by the instrument[29]. In our above example, a violation of monotonicity would be if some states defied national trends and held back individuals born late in the year from entering school, which would produce the opposite effect of the instrument compared with other states. Furthermore, successful instruments must correlate strongly with the treatment, as weakly correlated instruments lack the precision to produce practically useful estimates. It is important to ensure that these assumptions are met, as violations of any of these assumptions could threaten the validity of the results.

For example, though weather is a popular instrument considered for IV analysis[30,31] because changes in weather are plausibly random, it may not be appropriate depending on the specific causal question being asked. Consider a behavioral scientist who wishes to study the effect of exercise on mental well-being and plans to use temperature as an instrument on the amount of exercise individuals get. Though it tells an intuitive story where temperature ‘randomizes’ the amount of exercise individuals get, temperature may in fact be only weakly correlated with exercise (for example, people tend to exercise indoors, so the outside temperature does not matter much), or more concerningly, may violate the exclusion restriction (temperature may also influence mental health outside of exercise, such as through seasonal affective disorder). In this situation, though it seems that weather is a good candidate for IV analysis, a closer consideration of all the assumptions reveal flaws in the study design.

Provided that we find an appropriate instrument such that these assumptions are plausible, IV analysis can be used broadly across many disciplines to estimate causal effects in non-experimental data by leveraging observable sources of randomness (Table 1).

Regression discontinuity designs.

Another method of leveraging naturally occurring randomness to estimate causality is the regression discontinuity design (RDD). In RDDs, our treatment of interest X is assigned according to a sharp cutoff of a continuous running variable R, such as age, a standardized test score or a blood pressure reading. Because the cutoff is sharp (for example, if age 50 or above, patients get screened for cancer, otherwise they do not), the treatment assignment X is ‘as good as random’ (quasi-random) for individuals near the cutoff, allowing for the estimation of the causal effect of treatment X on the outcome Y (Fig. 3a).

The classic example of RDD concerns high school academic recognition (X) and their effect on subsequent academic achievement such as receiving scholarships (Y)[32]. US high school students take a standardized exam called the National Merit Scholarship Qualifying Test

(NMSQT), with students who meet a minimum score cutoff being nationally recognized with a Certificate of Merit. Here the continuous running variable is the NMSQT test score (R). Students who just meet the cutoff are not materially different from students who just miss the cutoff, essentially randomizing assignment of the Certificate of Merit near the cutoff score. This quasi-randomization design allows estimation of the causal relationship between the Certificate of Merit and receiving scholarships for students around the threshold. For our motivating example in the introduction, an RDD using pneumonia risk score cutoffs for hospitalization[33] could be used to obtain causal estimates.

A common method to obtain causal estimates using RDD is to fit two models of the running variable R (test score) on the outcome Y (scholarship amount) both to the left and right of the threshold, called local linear regression in the linear case (Fig. 3b depicts a local linear regression on simulated data)[34,35]. The causal effect of the treatment X (Certificate of Merit) at the threshold is the difference in the two fitted regressions' predicted value at the threshold. A practical consideration of implementing RDDs concerns bandwidth size[34]: how far from the threshold can an individual be and still have effectively random treatment? Smaller bandwidths make it more plausible for treatment to be quasi-random but also reduce the sample size of the regressions fitted. However, larger bandwidths may introduce bias into our estimates, particularly if the relationship between the running variable R and the outcome Y is nonlinear, as shown in the simulated data in Fig. 3c: note how the difference between the left and right local regressions at the cutoff value do not match the actual discontinuous change in the outcome Y. Fortunately, there has been work in econometrics exploring optimal and data-driven bandwidth selection methods[36,37]. Fitting local models on either side of the threshold leverages the quasi-randomness in treatment at the cutoff, allowing us to quantify the causal effect.

Similar to IV analysis, there are assumptions needed when performing RDD analysis to ensure valid estimation. We assume that only the running variable R has a discontinuous jump that drives the causal relationship between the treatment and the outcome (Fig. 3a). This assumption can be falsified in practice by verifying the continuity of other measured covariates at the cutoff value of R. Another critical assumption requires that individuals cannot perfectly manipulate the running variable, as this could make the groups above and below the threshold incomparable. In our above example, if some students can precisely control their test score such that they do just enough studying to achieve the Certificate of Merit cutoff, that would violate the quasi-randomization of treatment. Economists have developed tests for this form of running variable manipulation enabling falsification of the assumption[38]. Notably, the running variable does not need to be unconfounded to make valid causal estimates. Because of its relatively weak and often falsifiable assumptions, RDD is known as one of the most credible quasi-experimental methods for estimating causality from observational data[39].

However, although it is tempting to try and apply RDD analysis whenever there is a threshold determining treatment assignment, there are plausible situations where these assumptions do not hold. The assumption that units just above and below the threshold are comparable needs to be carefully considered in practice, as violations often arise when individuals in the study know the cutoff and the score. For example, one may study the effect

of grant awards on future academic success of young scientists, utilizing National Institutes of Health payline cutoffs for fellowship and grant funding for an RDD analysis[40,41]. However, as payline cutoffs are made public, it is likely that scientists who know that they just missed the cutoff are further motivated to work harder than scientists that just barely made the cutoff, potentially biasing estimates of future success because the scientists just above and below the cutoff differ in motivation.

Provided that the assumptions can be justified, RDD analysis has the potential to be widely applicable in estimating causal effects as thresholds exist in many real-world data settings (Table 1), even in unconventional settings such as geographic boundaries[42].

Difference-in-differences.

A third standard econometric approach called difference-in-differences (DiD) addresses confounding in observational data by looking at a treated and a control group, and comparing their outcome trends over time. We have measurements of outcome Y in time periods before (Y_1) and after (Y_2) treatment X for both a treated group that receives the treatment and a control group that does not receive treatment. The simple difference $Y_{2,\text{treat}} - Y_{1,\text{treat}}$ of the treated group could serve as an estimate of the causal effect of X, but it may be confounded by unobserved factors or time effects. The idea of the DiD is to use the difference $Y_{2,\text{control}} - Y_{1,\text{control}}$ of the control group as an estimate of confounding affecting the treated group, correcting for this effect by subtracting this second difference (Fig. 4a).

The classic example of DiD in economics concerns the effect of raising the minimum wage (X) on employment (Y)[9,43]. New Jersey raised the minimum wage in 1992, while the bordering state of Pennsylvania did not. The study compared fast food restaurant employment numbers from both states before and after the policy change. Changes in employment when looking only at New Jersey could have been confounded by other factors, such as a national recession. However, by subtracting the difference in employment observed in Pennsylvania, which did not see a minimum wage increase but is plausibly equally affected by confounders due to geographic and demographic similarities, the authors could control for potential confounding.

To perform DiD analysis, we use longitudinal data to make estimates of Y_1 and Y_2 (for example, pre- and post-minimum wage increase) for both the control and treated groups. The estimates can be single expectations computed over the pre and post time periods or, if the practitioner wishes to include multiple time points and control for other covariates, the estimates can be made with a time series regression[44,45]. We then take the difference between the estimated differences of the outcome values between the treated and control groups $((Y_{2,\text{treat}} - Y_{1,\text{treat}}) - (Y_{2,\text{control}} - Y_{1,\text{control}}))$, which yields a valid estimate of the effect treatment X has on outcome Y by subtracting out confounding factors that influence both the control and treated group equally: Fig. 4b depicts this visually, where the trend lines pre-treatment for both the treatment and control group are parallel. We note here that DiD analyses are a particular regression method for estimating causality from time series, with variations of DiD available for non-parametric and nonlinear settings[46–48]. The synthetic control, related to DiD, is another technique developed in economics to estimate causal effects from time series data[41].

Like the other quasi-experimental designs, DiD analysis requires assumptions to ensure valid causal estimates. The most important assumption is the presence of parallel trends: we require that the treated and control groups are not differently affected by confounding factors over time. In our minimum wage example, the parallel trends assumption would be violated if the Pennsylvania labour market reacted differently from the one in New Jersey to the ongoing recession. If the longitudinal data contain multiple time points, the parallel trends assumption can be falsified by examining the outcome Y in the two groups before the treatment X is applied, with Fig. 4c visually depicting such a violation. Another assumption required is the absence of spillover effects, where intervention itself results in a change in the composition of the control and treated groups[49]. This would be violated if the new minimum wage in New Jersey resulted in different individuals re-entering the job market or individuals from Pennsylvania moving to New Jersey. Another example of DiD assumptions failing to hold was a study examining the impact of US food aid delivered to countries on subsequent civil conflict, with random variations in US wheat production impacting the amount of aid provided to countries[50]. The problem was that the association between wheat production and conflict varied over time across the 'control' and 'treated' countries, violating parallel trends and showing that the effect of food aid on conflict may have been driven by spurious correlations[51]. For DiD analyses to be valid, we need evidence to support both parallel trends and the absence of spillover effects.

Nevertheless, the transparency and flexibility of the DiD analysis make it an attractive technique for causal inference provided that appropriate controls are selected and that the DiD assumptions are plausible. As many data science problems involve time series, DiD has the potential to be used widely for obtaining causal estimates from observational data (Table 1).

Limitations of quasi-experiments

There are limitations to the generalizability of causal estimates made using quasi-experimental techniques. All of the methods we have reviewed estimate causal effects for specific populations. IVs estimate the causal effect for the individuals affected by the instrument[52]. RDD estimates the causal effect of individuals at the threshold[53], though extrapolation of causal effects away from the threshold is possible under stronger assumptions[54]. DiD estimates the causal effect for the selected treated group. However, we note that generalizability concerns extend to even randomized experiments, where participant demographics could limit causal findings to a specific population[55]. When applying quasi-experimental methods, like all causal analyses, we must be mindful of which populations the estimates are valid for.

A practical consideration is that quasi-experiments require a specific data-generating process or specific type of variable observed to be applied, such as an instrument strongly correlated with the treatment (IV), the presence of a threshold determining treatment (RDD) or the existence of an appropriate control group to compare to the treated group over time (DiD). Thus, the estimation of causal effects using quasi-experimental methods depends on whether the data fits into these frameworks, as one cannot usually retrofit existing data into one of these methods: "[study] design trumps analysis"[56]. In addition, there is some danger

in looking for data that fit these designs without properly considering the assumptions underlying the techniques, which could invalidate the analysis if not met. As discussed in the previous section, data that may seem appropriate at face value for the application of these methods may not work when carefully considering the assumptions.

Outlook

Data science can benefit from quasi-experiments.

Quasi-experimental strategies should be more broadly applied in data science settings (Table 1) as they are the only way of meaningfully estimating causality in many situations. For example, IV analysis has been used to estimate causal effects of recommender systems[57] by leveraging random shocks in product popularity as instruments[58]. A popular technique in epidemiology called Mendelian randomization is an application of IVs that uses genetic variation as an instrument[59,60]. Thresholds in medical practice abound and should be used more frequently for RDD analysis[61]. DiD techniques can be used to evaluate the effects of policies and actions over time across domains such as advertising[62] and public health[63]. Quasi-experimental strategies could estimate causal effects in these settings, providing useful feedback for decision-making.

Furthermore, we can combine these causal estimates from quasi-experiments with other data to improve performance on tasks across ML. For example, there has been work using overlapping experimental data to control for confounding in observational studies[64], which could be extended to leverage quasi-experimental data. Reinforcement learning agents perform interventions, make decisions and form policies when interacting with their environment, which are inherently causal tasks[11,65]. However, modern reinforcement learning agents are often expensive to train[66], so incorporating quasi-experimental estimates of causal effects may be a promising way to improve efficiency. In particular, RDDs have been applied to bandit problems for policy optimization[67]. There has been an exploding interest in explainable artificial intelligence[68] with a renewed focus on causality[69,70]: how does an ML model's prediction change when we change an input feature? Though some explainable artificial intelligence methods rely on perturbing the data to generate explanations[71,72], there is value in trying to exploit naturally occurring randomness in the data to estimate causal effects so that the feature distribution of the original data is preserved, as ML models can behave erratically when given out-of-distribution examples[73]. Combining quasi-experiments and their unbiased causal estimates with more flexible and powerful ML methods promises meaningful blending of the strengths of these approaches.

Quasi-experiments can benefit from ML.

Work in economics has traditionally focused more on parameter estimation (for example, estimating and interpreting regression coefficients) than prediction[74], which brings opportunities for ML to be incorporated into quasi-experimental methods. In certain contexts, we can replace parametric estimators traditionally used in a method with more flexible ML estimators. For example, the first stage regression in IV analysis is effectively a prediction task. Work has been done scaling IVs to high-dimensional data, nonlinear data domains using least absolute shrinkage and selection operator (LASSO), kernel methods

and deep learning[75–77]. There has also been growing interest in inferring causality from longitudinal data, with ensemble models as well as matrix completion methods being used for causal estimates in time series data[78,79]. Further crosstalk and application of ML techniques to causal time series may be promising future work.

Another area where ML can improve quasi-experiments is by refining the target population to which the causal estimates apply. The estimates made by quasi-experiments are only valid for specific groups, such as individuals who are affected by an instrument or threshold in the cases of IV and RDD analysis. Predicting which individuals are likely to respond to a treatment or an instrument[80–82] could increase interpretability of the causal estimates, as well as inform exclusion criteria for study designs. Furthermore, ML methods excel at optimizing loss functions, which can be used to refine treatment and policy strategies. For example, there has been work in optimizing the treatment thresholds of RDDs with respect to maximizing treatment effect or minimizing potential harm[83,84]. ML extensions of quasi-experiments show promise in characterizing the target population of the causal estimates and informing better policy choices.

Empirical economists are showing interest in using ML and data science[74,85,86]. In addition to the settings we have already reviewed, there have been methods developed for causal inference in other contexts, such as under interference (where the treatment assignment of individuals can affect others' outcomes)[87] or in network data[88]. As ML methods become commonplace, data scientists and ML researchers should look to opportunities for collaborations with the economics community.

We need to take causality seriously.

Much like how causality has been highlighted as the central issue in areas such as marketing[89] and neuroscience[90], there needs to be wider acknowledgement that many data science questions are causal in nature[3]. We advocate for data scientists to view the questions they are working on through a causal lens: can the question be answered through correlation or prediction, or is there an underlying causal mechanism that one is seeking to understand? If the goal is to understand decision-making, actions or interventions, a data scientist needs to carefully consider their data: is it observational or experimental, are there concerns about unobserved confounding variables and what strategies can be taken to address confounding? The quasi-experiments we have presented here are one set of methods that can successfully identify causal effects in the presence of confounding, and should be part of data scientists' toolkit much like how they have become part of economists' toolkit[45,91].

We have highlighted opportunities for both data science and the traditional users of quasi-experiments to build off each other. Causality needs to play a larger role in data science, and quasi-experiments provide practical methods for measuring causality in observational data. At the same time, innovations from ML and data science are being used in economics, and the continued application of these ideas to complex data domains can increase the scope of questions economists and other social scientists can ask. We envision a future where more data science endeavours are explicitly centered around causality, and where other fields such as economics can leverage advances in ML to help answer their own causal questions.

Acknowledgements

We thank R. Ladhania and B. Lansdell for their comments and suggestions on this work. We acknowledge support from National Institutes of Health grant R01-EB028162. T.L. is supported by National Institute of Mental Health grant R01-MH111610.

References

1. van Dyk D et al. ASA statement on the role of statistics in data science. Amstat News https://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science/ (2015).
2. Pearl J The seven tools of causal inference, with reflections on machine learning. Commun. ACM 62, 54–60 (2019).
3. Hernán MA, Hsu J & Healy B Data science is science's second chance to get causal inference right: a classification of data science tasks. Chance 32, 42–49 (2019).
4. Caruana R et al. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In Proc. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1721–1730 (ACM Press, 2015); 10.1145/2783258.2788613
5. Finkelstein A et al. The Oregon health insurance experiment: evidence from the first year. Q. J. Econ 127, 1057–1106 (2012). [PubMed: 23293397]
6. Forney A, Pearl J & Bareinboim E Counterfactual data-fusion for online reinforcement learners. In International Conference on Machine Learning (eds. Precup D & Teh YW) 1156–1164 (PMLR, 2017).
7. Thomas PS & Brunskill E Data-efficient off-policy policy evaluation for reinforcement learning. International Conference on Machine Learning (eds. Balcan MF & Weinberger K) 2139–2148 (PMLR, 2016).
8. Athey S & Wager S Policy learning with observational data. Econometrica (in the press).
9. Angrist JD & Pischke J-S Mostly Harmless Econometrics: An Empiricist's Companion (Princeton Univ. Press, 2008).
10. Imbens G & Rubin DB Causal Inference: For Statistics, Social and Biomedical Sciences: An Introduction (Cambridge Univ. Press 2015).
11. Pearl J Causality (Cambridge Univ. Press, 2009).
12. Hernán MA & Robins JM Causal Inference: What If (Chapman & Hall/CRC, 2020).
13. Pearl J Causal inference in statistics: an overview. Stat. Surv 3, 96–146 (2009).
14. Peters J, Janzing D & Schölkopf B Elements of Causal Inference: Foundations and Learning Algorithms (MIT Press, 2017).
15. Rosenbaum PR & Rubin DB The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55 (1983).
16. Chernozhukov V et al. Double/debiased machine learning for treatment and structural parameters. Econ. J 21, C1–C68 (2018).
17. Spirtes P, Glymour CN & Scheines R Causation, Prediction, and Search (MIT Press, 2000).
18. Schölkopf B Causality for machine learning. Preprint at https://arxiv.org/abs/1911.10500 (2019).
19. Mooij JM, Peters J, Janzing D, Zscheischler J & Schölkopf B Distinguishing cause from effect using observational data: methods and benchmarks. J. Mach. Learn. Res 17, 1103–1204 (2016).
20. Huang B et al. Causal discovery from heterogeneous/nonstationary data. J. Mach. Learn. Res 21, 1–53 (2020). [PubMed: 34305477]
21. Wang Y & Blei DM The blessings of multiple causes. J. Am. Stat. Assoc 114, 1574–1596 (2019).
22. Leamer EE Let's take the con out of econometrics. Am. Econ. Rev 73, 31–43 (1983).
23. Angrist JD & Pischke J-S The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. J. Econ. Perspect 24, 3–30 (2010).
24. Angrist JD & Krueger AB Instrumental variables and the search for identification: from supply and demand to natural experiments. J. Econ. Perspect 15, 69–85 (2001).

25. Angrist JD & Krueger AB Does compulsory school attendance affect schooling and earnings? Q. J. Econ 106, 979–1014 (1991).
26. Wooldridge JM Econometric Analysis of Cross Section and Panel Data (MIT Press, 2010).
27. Angrist JD, Imbens GW & Krueger AB Jackknife instrumental variables estimation. J. Appl. Econom 14, 57–67 (1999).
28. Newhouse JP & McClellan M Econometrics in outcomes research: the use of instrumental variables. Annu. Rev. Public Health 19, 17–34 (1998). [PubMed: 9611610]
29. Imbens G Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics Working Paper No. 26104 (NBER, 2019); 10.3386/w26104
30. Hanandita W & Tampubolon G Does poverty reduce mental health? An instrumental variable analysis. Soc. Sci. Med 113, 59–67 (2014). [PubMed: 24836844]
31. Angrist JD, Graddy K & Imbens GW The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. Rev. Econ. Stud 67, 499–527 (2000).
32. Thistlethwaite DL & Campbell DT Regression-discontinuity analysis: an alternative to the ex post facto experiment. J. Educ. Psychol 51, 309–317 (1960).
33. Fine MJ et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. N. Engl. J. Med 336, 243–250 (1997). [PubMed: 8995086]
34. Lee DS & Lemieux T Regression discontinuity designs in economics. J. Econ. Lit 48, 281–355 (2010).
35. Cattaneo MD, Idrobo N & Titiunik R A Practical Introduction to Regression Discontinuity Designs (Cambridge Univ. Press, 2019).
36. Imbens G & Kalyanaraman K Optimal Bandwidth Choice for the Regression Discontinuity Estimator Working Paper No. 14726 (NBER, 2009); 10.3386/w14726
37. Calonico S, Cattaneo MD & Titiunik R Robust data-driven inference in the regression-discontinuity design. Stata J. 14, 909–946 (2014).
38. McCrary J Manipulation of the running variable in the regression discontinuity design: a density test. J. Econ 142, 698–714 (2008).
39. Imbens G & Lemieux T Regression discontinuity designs: a guide to practice. J. Economet 142, 615–635 (2008).
40. NCI funding policy for RPG awards. NIH: National Cancer Institute https://deainfo.nci.nih.gov/grantspolicies/finalfundltr.htm (2020).
41. NIAID paylines. NIH: National Institute of Allergy and Infectious Diseases http://www.niaid.nih.gov/grants-contracts/niaid-paylines (2020).
42. Keele LJ & Titiunik R Geographic boundaries as regression discontinuities. Polit. Anal 23, 127–155 (2015).
43. Card D & Krueger AB Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania Working Paper No. 4509 (NBER, 1993); 10.3386/w4509
44. Ashenfelter O & Card D Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs Working Paper No. 1489 (NBER, 1984); 10.3386/w1489
45. Angrist JD & Krueger AB in Handbook of Labor Economics Vol. 3 (eds. Ashenfelter OC & Card D) 1277–1366 (Elsevier, 1999).
46. Athey S & Imbens GW Identification and inference in nonlinear difference-in-differences models. Econometrica 74, 431–497 (2006).
47. Abadie A Semiparametric difference-in-differences estimators. Rev. Econ. Stud 72, 1–19 (2005).
48. Lu C, Nie X & Wager S Robust nonparametric difference-in-differences estimation. Preprint at https://arxiv.org/abs/1905.11622 (2019).
49. Besley T & Case A Unnatural experiments? estimating the incidence of endogenous policies. Econ. J 110, 672–694 (2000).
50. Nunn N & Qian N US food aid and civil conflict. Am. Econ. Rev 104, 1630–1666 (2014).
51. Christian P & Barrett CB Revisiting the Effect of Food Aid on Conflict: A Methodological Caution (The World Bank, 2017); 10.1596/1813-9450-8171.

52. Angrist J & Imbens G Identification and Estimation of Local Average Treatment Effects Technical Working Paper No. 118 (NBER, 1995); 10.3386/t0118
53. Hahn J, Todd P & Van der Klaauw W Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica 69, 201–209 (2001).
54. Angrist J & Rokkanen M Wanna Get Away? RD Identification Away from the Cutoff Working Paper No. 18662 (NBER, 2012); 10.3386/w18662
55. Rothwell PM External validity of randomised controlled trials: "To whom do the results of this trial apply?". The Lancet 365, 82–93 (2005).
56. Rubin DB For objective causal inference, design trumps analysis. Ann. Appl. Stat 2, 808–840 (2008).
57. Chaney AJB, Stewart BM & Engelhardt BE How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In Proc. 12th ACM Conference on Recommender Systems 224–232 (Association for Computing Machinery, 2018); 10.1145/3240323.3240370.
58. Sharma A, Hofman JM & Watts DJ Estimating the causal impact of recommendation systems from observational data. In Proc. Sixteenth ACM Conference on Economics and Computation 453–470 (Association for Computing Machinery, 2015); 10.1145/2764468.2764488
59. Lawlor DA, Harbord RM, Sterne JAC, Timpson N & Smith GD Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat. Med 27, 1133–1163 (2008). [PubMed: 17886233]
60. Zhao Q, Chen Y, Wang J & Small DS Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization. Int. J. Epidemiol 48, 1478–1492 (2019). [PubMed: 31298269]
61. Moscoe E, Bor J & Bärnighausen T Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. J. Clin. Epidemiol 68, 132–143 (2015).
62. Blake T, Nosko C & Tadelis S Consumer heterogeneity and paid search effectiveness: a large-scale field experiment. Econometrica 83, 155–174 (2015).
63. Dimick JB & Ryan AM Methods for evaluating changes in health care policy: the difference-in-differences approach. JAMA 312, 2401–2402 (2014). [PubMed: 25490331]
64. Kallus N, Puli AM & Shalit U Removing hidden confounding by experimental grounding. Adv. Neural Inf. Process. Syst 31, 10888–10897 (2018).
65. Zhang J & Bareinboim E Markov Decision Processes with Unobserved Confounders: A Causal Approach. Technical Report (R-23) (Columbia CausalAI Laboratory, 2016).
66. Mnih V et al. Human-level control through deep reinforcement learning. Nature 518, 529–533 (2015). [PubMed: 25719670]
67. Lansdell B, Triantafillou S & Kording K Rarely-switching linear bandits: optimization of causal effects for the real world. Preprint at https://arxiv.org/abs/1905.13121 (2019).
68. Adadi A & Berrada M Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160 (2018).
69. Zhao Q & Hastie T Causal interpretations of black-box models. J. Bus. Econ. Stat 39, 272–281 (2021).
70. Moraffah R, Karami M, Guo R, Raglin A & Liu H Causal interpretability for machine learning—problems, methods and evaluation. ACM SIGKDD Explor. Newsl 22, 18–33 (2020).
71. Ribeiro MT, Singh S & Guestrin C 'Why should I trust you?': Explaining the predictions of any classifier. In Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1135–1144 (Association for Computing Machinery, 2016); 10.1145/2939672.2939778
72. Mothilal RK, Sharma A & Tan C Explaining machine learning classifiers through diverse counterfactual explanations. In Proc. 2020 Conference on Fairness, Accountability, and Transparency 607–617 (Association for Computing Machinery, 2020); 10.1145/3351095.3372850
73. Hooker G & Mentch L Please stop permuting features: an explanation and alternatives. Preprint at https://arxiv.org/abs/1905.03151 (2019).

74. Mullainathan S & Spiess J Machine learning: an applied econometric approach. J. Econ. Perspect 31, 87–106 (2017).
75. Belloni A, Chen D, Chernozhukov V & Hansen C Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80, 2369–2429 (2012).
76. Singh R, Sahani M & Gretton A Kernel instrumental variable regression. Adv. Neural Inf. Process. Syst 32, 4593–4605 (2019).
77. Hartford J, Lewis G, Leyton-Brown K & Taddy M Deep IV: a flexible approach for counterfactual prediction. In Proc. 34th International Conference on Machine Learning Vol. 70 (eds. Precup D & Teh YW) 1414–1423 (JMLR.org, 2017).
78. Athey S, Bayati M, Doudchenko N, Imbens G & Khosravi K Matrix Completion Methods for Causal Panel Data Models Working Paper No. 25132 (NBER, 2018); 10.3386/w25132
79. Athey S, Bayati M, Imbens G & Qu Z Ensemble methods for causal effects in panel data settings. AEA Pap. Proc 109, 65–70 (2019).
80. Kennedy EH, Balakrishnan S & G'Sell M Sharp instruments for classifying compliers and generalizing causal effects. Ann. Stat 48, 2008–2030 (2020).
81. Kallus N Classifying treatment responders under causal effect monotonicity. In Proc. 36th International Conference on Machine Learning Vol. 97 (eds. Chaudhuri K & Salakhutdniov R) 3201–3210 (PMLR, 2019).
82. Li A & Pearl J Unit selection based on counterfactual logic. In Proc. Twenty-Eighth International Joint Conference on Artificial Intelligence (ed. Kraus S) 1793–1799 (International Joint Conferences on Artificial Intelligence Organization, 2019); 10.24963/ijcai.2019/248
83. Dong Y & Lewbel A Identifying the effect of changing the policy threshold in regression discontinuity models. Rev. Econ. Stat 97, 1081–1092 (2015).
84. Marinescu IE, Triantafillou S & Kording K Regression discontinuity threshold optimization. SSRN 10.2139/ssrn.3333334 (2019).
85. Varian HR Big data: new tricks for econometrics. J. Econ. Perspect 28, 3–28 (2014).
86. Athey S & Imbens GW Machine learning methods that economists should know about. Annu. Rev. Econ 11, 685–725 (2019).
87. Hudgens MG & Halloran ME Toward causal inference with interference. J. Am. Stat. Assoc 103, 832–842 (2008). [PubMed: 19081744]
88. Graham B & de Paula A The Econometric Analysis of Network Data (Elsevier, 2019).
89. Varian HR Causal inference in economics and marketing. Proc. Natl. Acad. Sci. USA 113, 7310–7315 (2016). [PubMed: 27382144]
90. Marinescu IE, Lawlor PN & Kording KP Quasi-experimental causality in neuroscience and behavioural research. Nat. Hum. Behav 2, 891–898 (2018). [PubMed: 30988445]
91. Abadie A & Cattaneo MD Econometric methods for program evaluation. Annu. Rev. Econ 10, 465–503 (2018).
92. Huang A & Levinson D The effects of daylight saving time on vehicle crashes in Minnesota. J. Safety Res 41, 513–520 (2010). [PubMed: 21134518]
93. Lepperød ME, Stöber T, Hafting T, Fyhn M & Kording KP Inferring causal connectivity from pairwise recordings and optogenetics. Preprint at bioRxiv 10.1101/463760 (2018).
94. Bor J, Moscoe E, Mutevedzi P, Newell M-L & Bärnighausen T Regression discontinuity designs in epidemiology. Epidemiol. Camb. Mass 25, 729–737 (2014).
95. Chen Y, Ebenstein A, Greenstone M & Li H Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. Proc. Natl. Acad. Sci. USA 110, 12936–12941 (2013). [PubMed: 23836630]
96. Lansdell BJ & Kording KP Spiking allows neurons to estimate their causal effect. Preprint at bioRxiv 10.1101/253351 (2019).
97. Patel MS et al. Association of the 2011 ACGME resident duty hour reforms with mortality and readmissions among hospitalized medicare patients. JAMA 312, 2364–2373 (2014). [PubMed: 25490327]

98. Rishika R, Kumar A, Janakiraman R & Bezawada R The effect of customers' social media participation on customer visit frequency and profitability: an empirical investigation. Inf. Syst. Res 24, 108–127 (2012).

99. Butsic V, Lewis DJ, Radeloff VC, Baumann M & Kuemmerle T Quasi-experimental methods enable stronger inferences from observational data in ecology. Basic Appl. Ecol 19, 1–10 (2017).

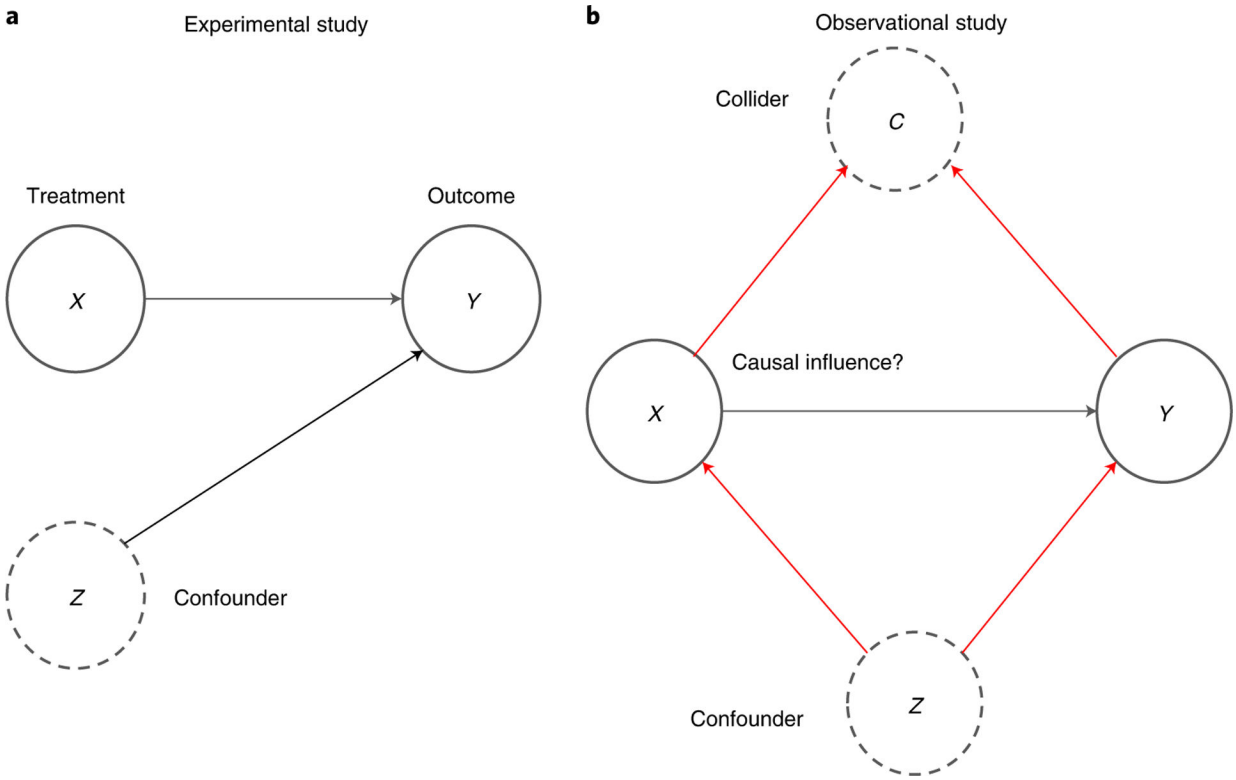

Fig. 1 |. Graphical representations for estimating causality in experimental and observational data.

a, Graphical representation of an experimental study, where there is no link between treatment X and confounder Z as X is randomized. **b**, Graphical representation of an observational study, where both colliders C and confounders Z can bias causal effect estimates.

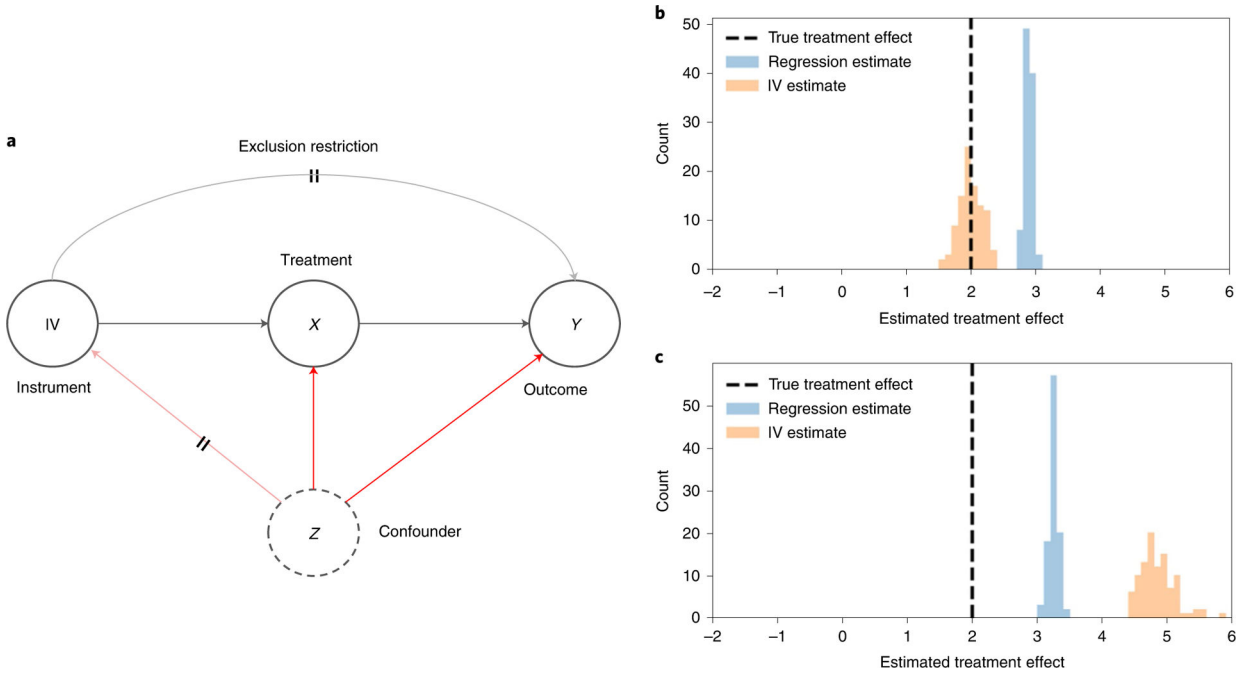

Fig. 2 |. DAG and plots of IVs.

a, Graphical representation of how IVs vary X to measure its causal effect on Y. The crossed edges depict assumptions needed for valid IV inference: the exclusion restriction (crossed grey edge) as well as no unmeasured confounding between IV and Y (crossed red edge). **b**, Histogram of causal effect estimates across 100 simulated datasets for both IV (orange) and typical regression (blue) when the true treatment effect (dashed line) of X on Y is confounded. **c**, Histogram of causal effect estimates across 100 simulated datasets for both IV (orange) and typical regression (blue) when the true treatment effect (dashed line) of X on Y is confounded and the exclusion restriction is violated.

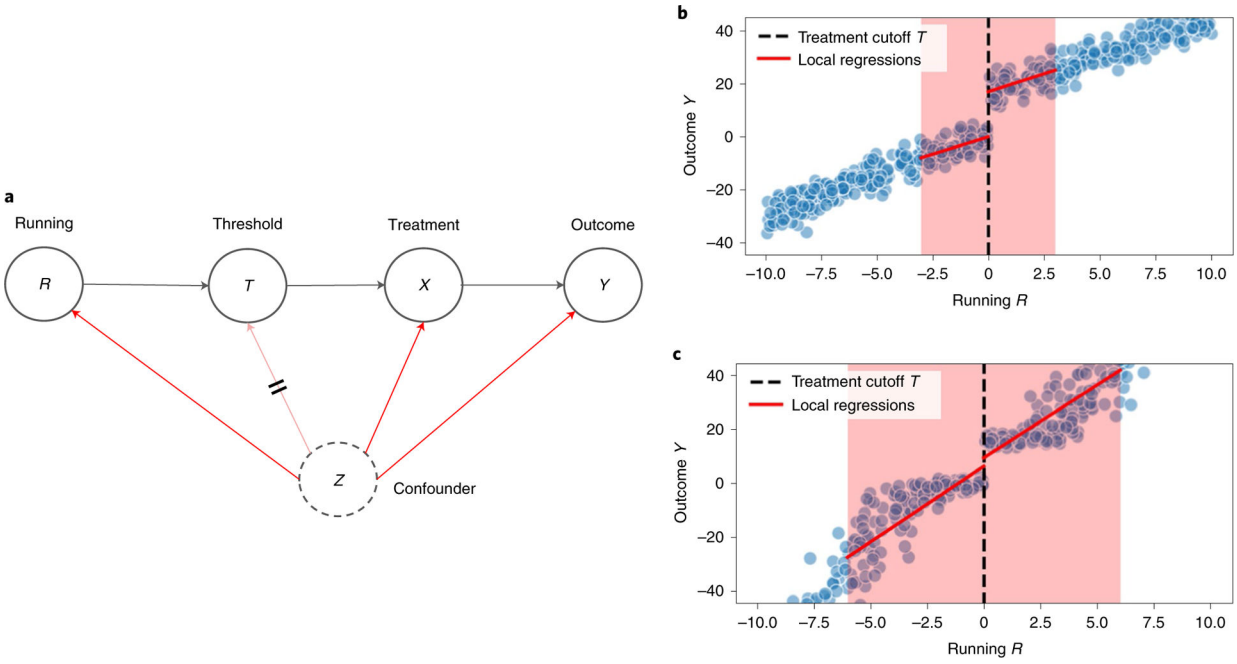

Fig. 3 |. DAG and plots of RDDs.

a, Graphical depiction of how RDDs use a threshold T in a continuous running variable R to effectively randomize treatment X (crossed red line) to estimate causal effects Y. **b**, Scatterplot (blue dots) and fitted local regressions (red lines) within the bandwidth (shaded red regions) of outcome Y on running variable R for simulated linear data to the left and right of the cutoff (dashed line). **c**, Scatterplot (blue dots) and fitted local regressions (red lines) within the bandwidth (shaded red regions) of outcome Y on running variable R for simulated nonlinear data to the left and right of the cutoff (dashed line).

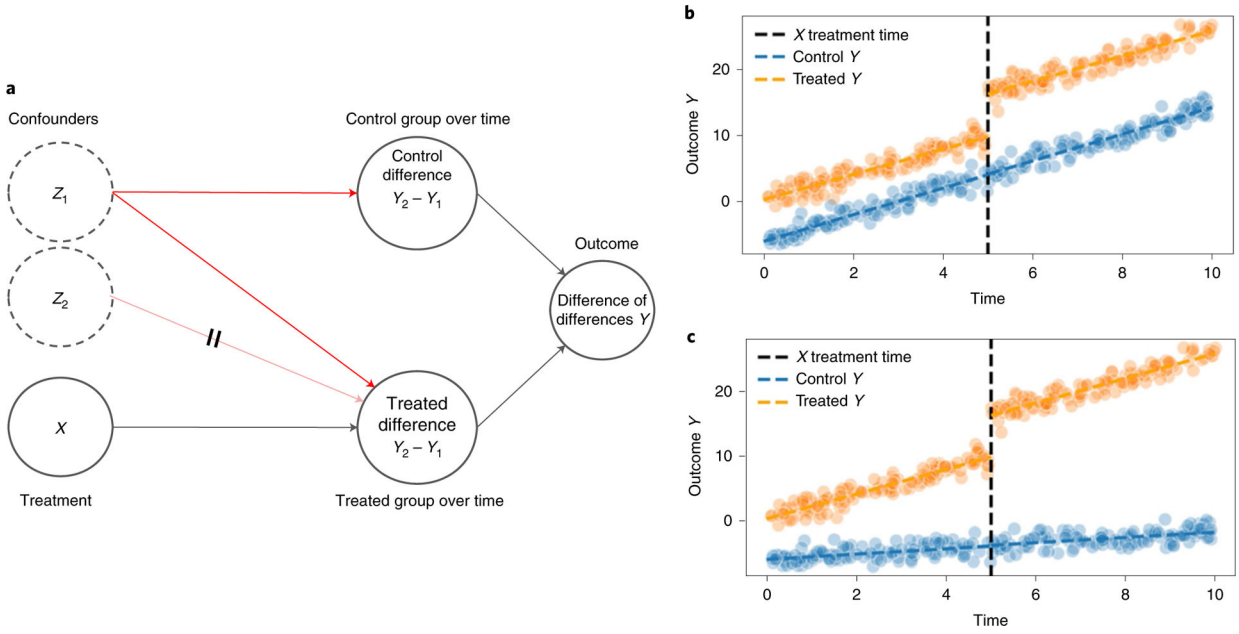

Fig. 4 |. DAG and plots of DiD.

a, Graphical representation of DiD, where both the control and treated groups are affected by the confounder Z_1, but the difference of their differences pre- and post-treatment yield unbiased estimates of the outcome Y. An appropriate control group must be chosen such that any confounders affect both groups (the crossed edge of Z_2 affecting only the treated group would be a violation of this). **b**, Scatterplot of outcomes Y for both the treated group (orange) and the control group (blue) over time in simulated data where parallel trends holds, with a treatment X being applied (dashed) to the treated group. **c**, Scatterplot of outcomes Y for both the treated group (orange) and the control group (blue) over time in simulated data where the parallel trends assumption is violated.

Table 1 |

Existing studies and opportunities for quasi-experiments

Instrumental variable				
Domain area	**Question**	**Treatment (*X*)**	**Instrument (*Z*)**	**Outcome (*Y*)**
Economics	Does required schooling influence future wages[25]?	Required schooling	Birth season	Future job earnings
Medicine	Using genome-wide association study data, what is the relationship between cholesterol and cardiovascular disease[60]?	High-density lipoprotein (HDL) cholesterol levels	Variation in thousands of genes that affect HDL cholesterol levels	Coronary artery disease and heart attack risk
Industry	How do product recommendations affect what products customers view[58]?	Product recommendation	Shock in popularity to product next to recommended product	Click-through rate on the recommended product
Society	Does sleep deprivation result in more car accidents[92]?	Sleep amount	Daylight savings time	Vehicle crash statistics
Neuroscience	Are neurons A and B causally connected to each other[93]?	Optogenetic stimulation of neuron A	random refractory periods of other neurons	Whether neuron B fires
Regression discontinuity designs				
Domain area	**Question**	**Treatment (*X*)**	**Running variable (*R*) and cutoff (*T*)**	**Outcome (*Y*)**
Education	How does academic recognition affect future academic success[32]?	Certificate of Merit recognition	NMSQT test score and minimum score needed for Certificate of Merit eligibility	Subsequent scholarship awards
Medicine	When should HIV patients get started on antiretroviral therapy[94]?	Antiretroviral therapy	CD4 T-cell count and threshold for treatment administration	Mortality
Industry	How much does being on the first page of search results influence click-through?	Website link on first page of search results	Search result ranking and first page display limit	Click-through rate on links
Society	How does air pollution influence life expectancy[95]?	Air pollutants from burning coal	Distance from river and free coal government policy north of the river	Life expectancy
Neuroscience	How do neurons learn to optimize their activity[96]?	Neuron spiking	Input neural drive and spiking threshold	Observed reward for neuron
Difference-in-differences				
Domain area	**Question**	**Treatment (*X*)**	**Parallel groups**	**Outcome (*Y*)**
Economics	Does increasing the minimum wage decrease employment[43]?	Increased minimum wage in New Jersey	Fast food workers in New Jersey and Pennsylvania	Employment after minimum wage increase
Medicine	How does limiting hospital work hours affect surgical outcomes[97]?	Limiting resident duty hours to 80 hours a week	Teaching and non-teaching hospitals, with/ without residents	Patient mortality, re-admission after duty hours limit
Industry	How does social media presence influence company profitability[98]?	Company launch of social media site	Participating and non-participating customers	Customer site visit frequency and profitability after social media site launch
Society	Does seeing advertisements make people unhappy?	Social media site incorporating ads into their app	Users and non-users of the social media platform	Well-being and sentiment measures after ads

Nat Comput Sci. Author manuscript; available in PMC 2022 June 04.

Ecology	How do wildfires affect plant species diversity[99]?	Wildfires burning plots of land	Burned and unburned geographically similar land plots	Measurements of plant species richness