The Importance of Being Causal

Iavor Bojinov^{1,2,3} Albert Chen⁴ Min Liu⁴

¹Department of Statistics, Faculty of Arts and Sciences, Harvard University, Cambridge, Massachusetts, United States of America,

²Technology and Operations Management Unit, Harvard Business School, Harvard University, Cambridge, Massachusetts, United States of America,

³Harvard Data Science Initiative, Harvard University, Cambridge, Massachusetts, United States of America,

⁴LinkedIn

Published on: Jul 30, 2020

DOI: <https://doi.org/10.1162/99608f92.3b87b6b0>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

Causal inference is the study of how actions, interventions, or treatments affect outcomes of interest. The methods that have received the lion's share of attention in the data science literature for establishing causation are variations of randomized experiments. Unfortunately, randomized experiments are not always feasible for a variety of reasons, such as an inability to fully control the treatment assignment, high cost, and potential negative impacts. In such settings, statisticians and econometricians have developed methods for extracting causal estimates from observational (i.e., nonexperimental) data. Data scientists' adoption of observational study methods for causal inference, however, has been rather slow and concentrated on a few specific applications. In this article, we attempt to catalyze interest in this area by providing case studies of how data scientists used observational studies to deliver valuable insights at LinkedIn. These case studies employ a variety of methods, and we highlight some themes and practical considerations. Drawing on our learnings, we then explain how firms can develop an organizational culture that embraces causal inference by investing in three key components: education, automation, and certification.

Keywords: causal inference, observational studies, cross-sectional studies, panel studies, interrupted time-series, instrumental variables

1. Introduction

The increasing abundance of data has enabled data scientists to uncover knowledge and insights that deliver a competitive advantage to firms in the private sector and agencies in the public sector (Hagiu & Wright, 2020). Data scientists extract value from data by identifying actionable and impactful opportunities, providing business ecosystem insights, and measuring the effect of innovations. While the actionability of an opportunity depends on organizational factors, its impactfulness can be forecasted through careful analysis, known as opportunity sizing, and a deep understanding of the ecosystem. Ecosystem insights come in many different flavors, but the objective is typically to define business metrics, develop new hypotheses, and gather data for future opportunity-sizing endeavors. Assessing the impact of innovations requires a retrospective analysis of how the project changed the relevant outcome (or business metric). In this article, we explain how LinkedIn's data science organization has expanded its ability to produce these three types of value-adding activities by investing in observational causal inference.

Causal inference enables the discovery of key insights through the study of how actions, interventions, or treatments (e.g., changing the color of a button or the email subject line) affect outcomes of interest (e.g., click-through rate, email-opening rate, or subsequent engagement; see Angrist & Pischke, 2009; Imbens & Rubin, 2015; Pearl & Mackenzie, 2018; and Rosenbaum, 2017, for comprehensive reviews). Over the past decade, most technology firms and a growing number of conventional firms have greatly expanded their

experimentation capabilities to measure the impact of product innovations (Thomke, 2020). In an experiment, the ability to randomize treatment assignment ensures that observed differences in outcomes between treatments are due to the intervention (Kohavi et al., 2020). Unfortunately, randomized experiments are not always feasible, as is often the case when it comes to opportunity sizing, understanding ecosystem insights, and measuring the impact of uncontrolled releases.¹ In these examples, directly attributing changes in the outcome to the treatment may lead to biased estimates as the differences may be due to a third factor (known as a confounder) that impacts both the treatment selection and the outcome. In these cases, we suggest an alternative approach that extracts causal claims directly from observational data; this practice is called observational causal inference (Rubin, 1974).

While observational studies greatly expand the possibilities for causal inference, there are two extremes to avoid. First, data should not be used to retrospectively justify decisions, as this builds false confidence and delivers little value. Second, a lack of observational data should not impede innovation, as some ideas may be so radical that existing data cannot fully capture the size of the opportunity. The ideal situation is somewhere between the two extremes, where data guides innovation without stifling it. For a firm with an established culture of rigorous experimentation, investments in well-designed observational studies extend its ability to uncover strategic opportunities, inform business intuition, and allocate resources optimally. It is important to remember that observational studies cannot replace experimentation; rather, they enhance it. While observational data does not provide perfect information about treatment effects, observational causal studies, when properly applied, make the best use of the available data to improve decision-making ability.

Methods for observational causal inference are not new to statisticians and econometricians, but its translation from research to industry applications remains challenging. To this end, we provide four case studies of how LinkedIn data scientists used observational studies to impact the firm's strategy. Along the way, we draw out themes, highlight critical considerations, and touch upon methods that every data scientist should know, with references for the eager reader. We then discuss how, by investing in education, automation, and certification, firms can develop a culture that embraces causal inference from observational studies. The goal of this article is to help data scientists and business leaders understand how observational causal studies can be applied to improve business and to catalyze firms to invest more in developing the infrastructure and culture that embrace this.

Example 1: Description vs. Prediction vs. Causal Inference. Most data scientists are familiar with prediction tasks, where outcomes are predicted from a set of features. This is fundamentally different from causal inference, which requires an understanding of how interventions will impact an outcome, rather than predicting in a constant state of the world (Hernán et al., 2019). Following we provide some examples of the different types of question data scientists have to answer (Leek & Peng, 2015).

Example Descriptive. What types of users have a large attrition (i.e., churn) rate?

Example Prediction. Can we predict who is likely to churn?

Example Causal. How do we reduce the likelihood of a user churning?

2. LinkedIn Case Studies

Each of our four LinkedIn case studies follows these four steps: first, we describe the business context and argue why an observational causal study is necessary; second, we provide the naive estimation strategy; third, we explain the choice of causal method; and finally, we share insights from the analysis. All effect estimates in the case studies are scaled for confidentiality.

About LinkedIn. LinkedIn’s vision is to create economic opportunity for every member of the global workforce. Through its website and mobile application, members can explore jobs, build meaningful relationships, and learn about opportunities to help advance their careers. LinkedIn offers enterprise applications that pair with this member ecosystem, to deliver value to both members and customers. LinkedIn Talent Solutions provides recruiting tools to help companies become more successful at talent acquisition, including promoting their company brand and engaging the right pools of qualified candidates. LinkedIn Marketing Solutions and LinkedIn Sales Solutions help customers engage a community of professionals in multiple ways (text advertising, sponsored messaging, lead generation, etc.) to improve brand awareness and build business relationships. LinkedIn Learning helps companies develop talent and helps employees keep vital business skills current with engaging online training and courses.

When to Use Observational Studies. Before starting an analysis, we determine if our question is indeed causal as opposed to descriptive or predictive; see Example 1 for a comparison. If the question is causal, then we decide whether we can answer it through an experiment; if we cannot, then we rely on an observational study. Broadly speaking, we use observational studies in three types of analysis:

1. *Opportunity Sizing.* Determines if any treatments out of a set of candidates are good business opportunities. Typically, the treatments are not fully implemented, and so we cannot use experiments to assess their feasibility.
2. *Ecosystem Insights.* Derived from analyzing all aspects of natural firm operations. Although it is possible to learn these through experiments, in practice, it is too costly and time-consuming to do so. Instead, we leverage observational causal inference methods for extracting practical insights.
3. *Uncontrolled Rollout.* Whenever the release process of innovations is outside of the control of the data scientists. In this case, an observational study is the best method to estimate the causal impact.

For these studies, suitable observational data must be available to test the hypothesis (e.g., the treatment must come before the outcome). The input data structure generally falls into one of four categories: cross-sectional, instrumental variable, panel, and interrupted time-series. Each category supports multiple methods to extract plausible causal estimates, each with its own set of assumptions, strengths, and limitations. One consistent thread across all approaches is that the results depend on the validity of the underlying assumptions. Therefore, data scientists should carefully apply diagnostic tools to assess violations of the underlying assumptions, and sensitivity analyses to measure the robustness of the findings.

2.1. Case 1. Opportunity Sizing: Job Postings

Talent Solutions is a suite of tools at LinkedIn that enables employers to attract job seekers by posting jobs on the site. An ideal job posting has a compelling description that provides all of the relevant information required by a job seeker to determine if they should apply. One way to ensure that a job seeker has all of the relevant information is to make some text fields mandatory at the time a job is posted. However, increasing the number of required fields also increases the complexity of posting a job, which can deter or delay new listings. Answering causal questions about the value of each job attribute can help product designers decide which fields should be required.

In this study, the units are job postings; the treatment is having versus not having an attribute, such as job title, function, industry, location, or employment type. Each attribute corresponds to a different treatment. Of the key business metrics we track, the most important is the view-to-apply rate, which is the probability a member applies to a job after seeing it. For simplicity, we focus our discussion on assessing the benefit of having a job title.

We can obtain a naive estimate by comparing the view-to-apply rate between the jobs that have a job title and those without one. Our naive approach yields an estimate of an approximately 10% difference in the view-to-apply rate. This, however, is likely an overestimate of the effect because of potential confounders. For example, listings from well-known companies typically have a higher view-to-apply rate and are more likely to post jobs with titles. The 10% difference could then be due to the effect of company popularity as opposed to the job title.

Whether or not a job posting has an attribute is determined at creation time, so each unit can only take the treatment once. This one-time treatment is a characteristic of cross-sectional studies. After the job is posted (i.e., the treatment is assigned), we monitor the view-to-apply rate for each posting. The temporal ordering of the intervention occurring before the outcome is vital in ensuring that it is plausible for the treatment to impact the outcome.

Cross-sectional study is the most prevalent class of observational studies. It usually provides a good starting point for answering many causal questions. Table 1 shows an example of the typical data

structure of cross-sectional studies. The data structure for a cross-sectional study includes a treatment label collected for each unit, an outcome collected after the treatment, and covariates collected before the treatment, as shown in Figure 1.

Table 1. Cross-sectional study data structure.

Unit	Covariate 1	Covariate 2	Other Covariates	Treatment	Outcome
	Job posting type	Job posting duration	...	1 if job has title, 0 otherwise	View-to-apply rate
1	Paid	6	...	0	0.1
2	Free	2	...	1	0.08
3	Paid	18	...	1	0.2
4	Paid	6	...	0	0.05
...	...	...	...	...	...

Figure 1. Cross-sectional timeline. First measure the covariates, then the treatment assignment, and finally the outcome or success metrics.

Popular strategies for analyzing cross-sectional observational studies follow a two-step process: design phase and analysis phase (Stuart, 2010). In the design phase, data scientists try to remove any systematic differences in the observed confounders between the treatment groups. A successful design phase creates a data set consisting of units that are alike in every measurable way, except for the treatment they took. Since the vital property of randomized experiments that allows us to identify the causal effect with the assumption that the only difference between the two groups is the treatment that they took, we can think

of the design phase as trying to approximate a randomized experiment. See Rosenbaum (2010) for a book-length review of cross-sectional observational study methods; common methods include matching-based methods (Iacus et al., 2009; Rubin, 1973, 2006; Stuart & Green, 2008) and weighting-based methods (Bang & Robins, 2005; Czajka et al., 1992).

To obtain a more accurate estimate, we matched treated and control job postings based on covariates (e.g., Table 1) that correlate with both the outcome value and treatment assignment. We used hundreds of categorical covariates, and even after removing outliers, our samples consisted of 16 million job postings. After matching, the treatment effect estimate reduced to 2.4%, a 76% reduction. For other treatments, such as having the job attributes function, skill requirements, and location, we found a reduction of opportunity size on the order of 38% to 56%.

Based on our opportunity-sizing analysis, we decided that job posting title and skills were the most impactful fields to include in a job listing. Our main recommendation was that product features should encourage job postings to contain these two attributes. Through this analysis, we also learned the importance of highlighting to job seekers the relevant skills they possessed for a particular job; this led to both user interface changes as well as an update in our recommendation algorithm.

2.2. Case 2. Business Ecosystem Insights: Value of Free Trials

LinkedIn is a complex ecosystem that hosts many subproducts, such as a social feed, notifications, member profiles, jobs, and online learning courses. LinkedIn has four companywide metrics and many more product-specific metrics that are sensitive to changes within a subproduct. A question that often arises is, are the product-specific metrics good surrogates for the company-wide metrics? In other words, does optimizing for the product-specific metrics necessarily lead to the optimization of company-wide metrics? To answer this question, we use causal inference methods to understand how changes in a product-specific metric impact the company metric. These results help each product area to set better goals and develop more accurate metrics.

LinkedIn Learning offers online education through video content. There are a variety of ways for members to access these courses, including organization-provided programs, individually paid subscriptions, and free courses. The product-specific success metric is the number of members whose video watch time in the past 30 days exceeded a certain threshold—we call these engaged learners. To determine whether this is a useful metric, we assess how increasing it impacts companywide metrics, in particular, we focused on revenue generated through purchasing a LinkedIn Learning subscription.

We can compute a naive correlational estimate of the impact by comparing revenue between engaged learners and everyone else. The average revenue generated by engaged learners is 94% higher than nonengaged learners. This estimate is likely to be much higher than the truth. To obtain a more accurate one, we can use the

results of past experiments that directly impacted product-specific metric (in this case, engaged learner) as an instrument.

Instrumental variable (IV) methods are ones that use a so-called instrument, a variable that affects the outcome, but only through changing the treatment (Angrist et al., 1996). In other words, an instrument has a direct impact on the likelihood of a user taking treatment, but does not directly impact the outcome. The exogenous variation introduced by the instrument allows us to isolate an estimate of the causal effect even in the presence of unobserved confounders. This method is particularly attractive for firms that run randomized experiments, as these create a class of natural instruments. While randomized experiments measure the net effect of experiment assignment on the outcome, in many cases, assignment affects the outcome indirectly by encouraging some “treatment.” If we are interested in the effect of this “treatment” on the outcome, regardless of how it is encouraged, we can use IV with the experiment assignment as the instrument. For a review of methods and applications, see Angrist & Krueger (2001). Table 2 shows the typical data structure, and Figure 2 shows the observation timeline.

Table 2. Instrumental variable data structure.

Unit	Covariate 1	Other Covariates	Instrument	Treatment	Outcome
	Total sessions	...	1 if shown banner ad, 0 otherwise	1 if is engaged learner, 0 otherwise	1 if subscribed, 0 otherwise
1	5	...	0	0	1
2	17	...	1	1	0
3	21	...	1	1	1
4	8	...	0	0	0
...	...	...	...	...	...

Figure 2. Instrumental variable observation timeline. First measure the covariates, then the instrument, then the treatment assignment, and finally the outcome or success metrics.

Nonsubscribers are sometimes given free access to learning courses for 24 hours. One experiment aimed to increase awareness of the free course access by notifying members in-product. The experiment randomly assigns members into two versions:

1. *Banner Ad.* The member is notified of the promotional trial period via a banner ad; this prompts them to start learning right away.
2. *No Banner Ad.* The member is not shown the banner ad but has access to the promotional trial.

This experiment directly impacted the proportion of engaged learners (the treatment) but did not have a direct impact on the likelihood of member conversion to paid subscribers (the outcome), other than through affecting engaged learner status. The IV study established that becoming an engaged learner increases the probability of conversion to a paid member by 46%, only half of the size of the naive estimate of 94%.

2.3. Case 3. Business Ecosystem Insights: The Value of Contributions

The LinkedIn platform is built to facilitate interactions between members either through public contributions (posting, commenting, or sharing in a social feed) or private contributions (messaging another member). The firm's working hypothesis is that contributions are a starting point for conversations and increase long-term value for the member. Typically, we measure this value by the member's engagement level, which is quantified through metrics such as time spent on LinkedIn.

Although the working hypothesis seems plausible, it does raise some questions: What is the impact of public contributions on subsequent member engagement? Which of the different member contributions drives retention the most? What is the relative importance of public versus private contributions? Answering these questions using an experiment is incredibly challenging because it is difficult to develop interventions that directly elicit member contributions; the member inherently decides which treatment to adopt. Even if we can modulate contribution behavior through user-interface changes, it is difficult to find such an experiment that does not also directly impact many other behaviors, such as scrolling patterns.

A naive approach that directly compares members who like a post to those who do not suggests that liking a post increases the likelihood of returning to LinkedIn in the following week by 80%. Casting this problem in the cross-sectional causal inference framework outlined in Case 1 yields an estimate of 34%, which represents a 57% reduction from the correlational estimate. However, we are still dropping a substantial amount of useful information.

Carefully examining our data, we notice that we observe each user taking multiple actions and see subsequently how their engagement changes in response to their behavior. This generates what is known as a panel data set; see Table 3 for an example of the data structure, and Figure 3 for the observation timeline. Notice that cross-sectional data is a particular case of panel data, with a single time step. Another important special case is when a single unit is observed over multiple time steps; this is known as a time series. In the next case study, we describe a specific application of causal inference for interrupted time series.

There are multiple strategies for analyzing **panel data sets**, see Table 3, and extracting causal effects that come with different assumptions. At a minimum, all strategies require that the outcome is observed after the treatment to ensure that the intervention can affect the outcome. The addition of a time component allows us to control for unobserved fixed confounders as we observe each user over time and see them taking both control and treatment—in a sense, each user can act as their own control (Imai & Kim, 2019). The applicability and reliability of these methods depend on the validity of the underlying assumptions and should always be combined with detailed diagnostics and sensitivity analysis methods (see, e.g., Chamberlain, 1982; Imai & Kim, 2019; Robins & Hernan, 2008; Sobel, 2012).

Table 3. Panel data structure.

Unit	Pre-treatment	Time Period 1		Time Period 2		Time Period 3	
	Covariates	Treatment	Outcome (Sessions)	Treatment	Outcome (Sessions)	Treatment	Outcome (Sessions)
1	3	2 Post	1	Nothing	0	1 Post	2
2	4	Nothing	0	1 Like	2	1 Like	1
3	3	1 Message	3	1 Message	1	Nothing	2

Figure 3. Panel observation timeline. A key feature of panel data is that each member is observed taking different treatments over time, followed by a measurement of their response.

To improve our estimate, we used a weighted linear fixed effects model, described in Bojinov et al. (2019).

The model allows each user to act as their own control, and the weighting further improves the performance by reducing the discrepancies between the treatment and control units. From our analysis, we concluded that the less engaged members see the most significant gain from contributing. If we had used a purely correlational analysis, we would have overestimated the effect by more than 75% and falsely concluded that highly engaged members have similar benefits. The magnitude of the estimated effect was also much closer to what we observe from typical experiments.

2.4. Case 4. Uncontrolled Rollout: Marketing Campaigns and Mobile App Preloads

Experimentation allows us to measure the impact of innovations as we control their release to our members. However, there are times when we are unable to run an experiment because we have no control over which members are exposed to the treatment. Uncontrolled rollout occurs in situations such as in marketing campaigns,² new mobile application releases, and mobile app preloads.³ LinkedIn, for instance, regularly targets specific cities with brand marketing campaigns involving both physical (billboards, radio, and television) and digital marketing channels to increase engagement. Similarly, LinkedIn uses mobile application preloading to increase sign-ups and engagement. Both use cases raise an essential question: What is the return on investment for these marketing activities?

In this section, we analyze marketing campaigns and the impact of app preloads using a fourth type of observational study format.

Interrupted time-series arise when we track an outcome of interest before and after an intervention (Lopez Bernal et al., 2016); see Table 4 and Figure 4 for the typical data structure and observation timeline. The classic example is a marketing campaign launched in a city. The effectiveness of these campaigns is hard to assess through standard experiments because randomization at the member

level is impossible. Randomization at the city level results in a small number of units, resulting in little power to detect effects without using more advanced tools.

Synthetic control methods are a subclass of interrupted time-series methods that we have found to be particularly useful when there are data on other units that were not affected by treatment (Abadie et al., 2010). In the marketing example, these could be the outcome metric in cities with no marketing campaign.

Synthetic control methods allow us to use a time series of observed outcomes before the intervention to generate a control group that is comparable to the treatment group under a particular model. We then estimate the causal effect by contrasting the observed time series against a counterfactual estimate of how the outcome time series would have evolved without the treatment, inferred from the synthetic control. For a popular method, see Brodersen et al. (2015). These methods naturally facilitate impact assessment over time.

Figure 4. Interrupted time-series observation timeline. The input data set for the interrupted time-series method consists of multiple time series over the same time period. One of them is the treatment's metric of interest. There can be any number of control time series, and the metric they measure is not required to be the same one as for treatment.

Table 4. Interrupted time-series data structure.

Time Step	Is the Focal Unit Treated?	Focal Unit Outcome	Control 1	[Control 2]	[Control 3...]
1	No	2	1	1	8
2	No	5	6	8	8

3	No	2	4	5	1
4	Yes	1	3	3	3

First, we analyze the impact of marketing campaigns. LinkedIn’s marketing campaigns are targeted at specific cities or geographical locations. A naive approach to estimate their impact is to count all traffic directed to LinkedIn from the marketing campaign (i.e., referred traffic). However, referral tracking is not always possible (especially for physical marketing campaigns), and even if it were, not all this traffic is incremental, as some members may have visited LinkedIn regardless. Another naive approach is a pre–post comparison of the metric value before and after the campaign. However, this conflates the natural change over time with the effect of the campaign itself.

A better approach is to analyze the data using an interrupted time-series method. We typically use synthetic control methods to measure the aggregate impact at the city level. The treatment region is the city exposed to the marketing campaign, and control regions are cities with comparable characteristics (e.g., sign-up penetration, member engagement, among many others) that were not exposed. Compared to the naive pre–post estimate, the synthetic control estimate can be higher or lower depending on time-series properties. In one such analysis, the naive approach estimated 20% lift, while the synthetic control estimate was 11%. The large causal impact influenced our decision to expand marketing to more cities. We identified demographic features of cities that responded well to our campaigns to inform strategic targeting of the next campaign.

On a practical note, even though this process does not require referral tracking, it does require having suitable controls. Marketing is an ongoing process, so it is important to launch campaigns in an orderly way that facilitates ongoing data collection and impact measurement. If cities are treated in a haphazard fashion, then it can be difficult to find controls that are not affected by intervention in the measurement period.

Next, we analyze the impact of mobile application preloading. A naive correlational study directly compares revenue from members who use a preloaded app against those who do not. However, along with typical confounding issues, some members who use the preloaded app would have installed it on their own, so the correlational study does not accurately measure the incremental revenue caused by the preload.

Just as in the marketing example, the data structure resembles an interrupted time-series, and so we again used synthetic controls to estimate the causal impact more accurately. Treatment comprises members who used a preloaded app, and we measure the revenue they generate over time. Our goal is to measure incremental revenue to assess return on investment. Rather than measure the impact on treatment as a whole, we split the treatment into multiple cohorts that past ecosystem analysis showed have different monetization values (segmented by previous app installation status, geographic region, etc.). We performed a separate analysis for each cohort and combined the final results in a meta-analysis. Modeling at the cohort level is similar to matching for cross-sectional data. It improves our model accuracy and yields insights into how value differs by

cohort. After defining the treatment cohorts, we defined the control cohorts. The ideal controls are members that were just as likely to be exposed to preload but were not. One approach would be to use propensity matching to identify a similar control set. We opted for a more straightforward approach, defining control cohorts that mirror the ones in treatment. For example, the treatment cohort with ‘First time app users on Android, in the United Kingdom’ was matched to a control cohort of United Kingdom members who never used the app before. We used an interrupted time-series method to estimate the causal effect.

The results were delivered to our business development teams as a self-serve calculator that estimates the return of a potential preload partnership according to its targeting criteria (e.g., geographic region). Business deals could be negotiated to break-even within a certain time horizon. Had we used naive correlation, our price targets would have been 50% to 250% higher than the true value, depending on the cohort. Causal analysis eliminated the guesswork so LinkedIn could negotiate preload deals with confidence.

In both these examples, observational causal methods deliver ongoing support to our marketing and business development teams to enable smart decision making in how we spend money to build our brand and engage our members. Moreover, as they rely on fewer ‘guesstimates’ than naive baselines, they give executives confidence that the marketing budget will be well-spent. Observational causal inference is the method of choice for accurate impact assessment of uncontrolled rollouts.

3. Organizational Adoption: Practical Lessons From LinkedIn

Observational causal studies provide an important class of tools for making well-informed data-driven business decisions; unfortunately, data scientists in many firms struggle to apply these in a business context. At LinkedIn, we identified three central components for building a culture that adopts and benefits from observational studies: education, automation, and certification.

3.1. Education

Data scientists. While many data science degree programs offer courses on causal inference, because of the breadth of the field, most new hires know little about experimental design and even less about observational studies. At LinkedIn, we created an internal education program, supplemented by external content, to develop causal evangelists who can then educate others. Because internal experts understand both domain context and statistical techniques, they are uniquely equipped to help teams apply methods for practical applications. Our training sessions covered when to use observational causal inference, the assumptions of different methods, proper analysis design, and how to choose the right method for the problem. To supplement our internal employee development, we look more broadly across disciplines when hiring, focusing especially in fields that deal with observational data, such as the social sciences.

Leaders. Data scientists cannot run observational studies in a vacuum. Leadership support is essential for ensuring adoption by decision makers, coordinating high-quality data collection across teams, and supporting

the necessary resource allocation. This can be done by focusing on champion use cases to demonstrate value and drawing on real-life illustrations of the dangers of mistaking correlation for causation. For instance, correlation shows, counterintuitively, that asthmatic patients have a lower probability of dying from pneumonia than nonasthmatics. But it would be wrong to conclude that the risk of death is lower for asthmatics; because of their high risk, they receive better treatment, leading to the surprising result of better outcomes (Caruana et al., 2015).

At LinkedIn, we proved the usefulness of observational causal studies to the business by answering a few top-priority strategic questions and comparing the results to correlational studies that would have yielded the wrong investment decisions (for example, Case Study 4). We also demonstrated that observational studies could yield results consistent with a randomized experiment,⁴ and therefore the results of a well-designed study could be reliably used for decision making. Because of our education efforts, LinkedIn's employees are now aware of the difference between correlation and causation, and that there is another option besides correlation and randomized experiments. Now, both data scientists and business leaders are quick to ask whether claims are 'causal' and advocate for rigorous, accurately communicated insights.

3.2. Automation

Even with a trained workforce, it takes significant time to design and run a proper observational study. In our experience, the first iteration of a study takes 2–3 weeks, followed by additional iterations to pass diagnostic tests. Analysis time depends on the complexity of the problem and the scale of the data, but when the cost is too high, few data scientists have the time to run observational causal studies.

At LinkedIn, we dramatically decreased analysis time while increasing quality by automating portions of the work. The main bottleneck in most observational studies is data collection, as this requires joining multiple data sets from different sources while carefully tracking the correct time index (or timestamp). The data sets for observational causal analysis typically have a similar structure: unique unit IDs, timestamps, treatment labels, response metrics, and confounders. Automating the data join overcomes one of the main hurdles to the adoption of observational studies. It also brings privacy advantages as individual data scientists do not need access to the original data.

To further improve reproducibility, productivity, and trust, we built a causal inference web platform, hosting all four categories of methods described in the case studies. Data scientists can execute an analysis at a click, with backend automation of computation and validation (Figure 5). Data scientists specify the analysis configuration (such as method, treatments, outcomes, and features) through a user interface, and the platform handles the data joins, analysis, and validation. The user interface guides the data scientist to develop a proper design that, for example, satisfies the temporal ordering of treatment and outcome. It also simplifies the process to refresh a previous analysis and quickly iterate on a new one: users can rerun any analysis with adjusted parameters. Automation reduced the time it takes to create the first iteration from a few weeks to a few hours. Furthermore,

centralization streamlines the review process and ensures high-quality analysis by integrating diagnostics and validation tests. Finally, the democratization of causal methods has enabled us to build a knowledge repository that simplifies the discovery of new insights through web pages that display analysis design, result, and approval status side-by-side. These web pages can be shared, searched, and organized. Thus, the observational causal platform serves not only as an analysis platform, but also as a repository for causal relationships and ecosystem insights.

Figure 5. Doubly Robust analysis design page
Video - click to play

The development team, consisting of a small group of data scientists and software engineers, owned the platform and methodology development. Other data scientists can build into the platform directly, with code review and guidance from our development team. For instance, the data scientists working on the brand marketing use case created a custom user interface for input and result visualizations. Our approach of developing a comprehensive platform while enabling customization for the top use cases met the general and specific needs of data scientists to facilitate adoption.

Like any automated statistical tool, without careful guardianship, there are ample opportunities for abuse. Observational studies are particularly prone to misuse as they rely on strong assumptions, some of which are unfalsifiable from the observed data alone. They often need in-depth domain knowledge to create a good design. Education and automation both act as safeguards to inform the proper design, but not every analysis run on the platform is guaranteed to be accurate. That is why we decided to rely on human certification to ensure that only valid results are called causal.

3.3. Certify

It is dangerous to assume correlational results are causal; it is even more hazardous to place confidence in poorly designed observational studies. To uphold a high standard, we established the Causal Data Analysis Review Committee to certify causal analyses and ensure the proper interpretation is communicated to business leaders. The committee holds office hours to help data scientists with analysis conception, study design, and result interpretation. During this process, we carefully assess the validity of assumptions, check the design, and ensure there was no abuse (akin to p-hacking or data-dredging) by examining the full history of analyses on the platform. Data scientists can request a review through our web platform. The approval status is then displayed within the report so that it is clear which results can be trusted and communicated to stakeholders.

The certification process also allows us to ensure that the results are properly communicated and interpreted in the business context. We typically require data scientists to present estimates along with confidence intervals and model assumptions in simple business terms (e.g., the population used for the study, the features included in the model, and limitations). In this way, without going into technical details, decision makers can understand the boundaries of the analysis. For some studies, notably opportunity sizing, we further push data scientists to think hard about the external validity of the estimates. Often, simple adjustments to align the composition of the sample to the population that will receive any future interventions will improve the quality of the study.

One challenge that certification cannot fix is that as results are broadly shared, problem-specific nuances are less understood, and as a result, there is a tendency to remember a single estimate without the details. Another problem is that users may drop ‘observational’ and instead say, ‘a causal study showed...’. To tackle both of these issues, we carefully educate stakeholders on what the ‘causal’ label means. We emphasize that observational studies cannot demonstrate causality as convincingly as an experiment, even though they provide a significant improvement over correlational studies. It is helpful to show the simple pyramid diagram in Figure 6.

Observational causal studies attempt to extract “causal claims” without a randomized experiment.

Figure 6. Pyramid diagram for the types of studies.

Although governance for certification and communication adds friction, it is vital to building trust. The committee is currently composed of reviewers from the development team, an impartial horizontal data science team with no personal stake in whether the result is positive or negative. As the firm matures, the certification process can be democratized. We have begun to scale the process by adding members from vertical data science teams to review analyses from other verticals. Through trustworthy results, we foster a culture that believes in following the evidence from data.

Acknowledgments

We thank our colleagues Xiaofeng Wang, Simon Yu, Vivek Agrawal, Guillaume Saint-Jacques, Jeremy Simpson, Alexander Ivaniuk, Maneesh Varshney, and Kinjal Basu from LinkedIn for building the observational causal inference platform; Fiona Li, Jackie Zhao, Ivan Chen, Iris Tu, Ming Wu, Aman Gupta, and Nan Wang for sharing their case studies; Weitao Duan, Shan Ba, Ken Soong, and Stephen Lynch for providing feedback on this article; Ya Xu, Parvez Ahammad, and Dan Antzelevitch for their continued support and investment to the observational causal studies initiative. We would also like to express our gratitude to users of the causal inference platform who provided us with many insights for improving it. Finally, we are immensely grateful to the *HDSR* editorial team for their comments on earlier versions of the article.

Disclosure Statement

The views and opinions expressed in this article are those of the authors only, and do not represent the views, policies, and opinions of any institution or agency, any of their affiliates or employees, or any of the individuals acknowledged below.

I. Bojinov, A. Chen, and M. Liu report no conflicts.

References

- Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490), 493–505. <https://doi.org/10.1198/jasa.2009.ap08746>
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444–455. <https://doi.org/10.1080/01621459.1996.10476902>
- Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4), 69–85. <https://doi.org/10.1257/jep.15.4.69>
- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press. <https://doi.org/10.1515/9781400829828>
- Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962–973. <https://doi.org/10.1111/j.1541-0420.2005.00377.x>
- Bojinov, I., Tu, Y., Liu, M., & Xu, Y. (2019). Causal inference from observational data: Estimating the effect of contributions on visitation frequency at LinkedIn. *arXiv*. <https://doi.org/10.48550/arXiv.1903.07755>
- Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1), 247–274. <https://doi.org/10.1214/14-aos788>
- Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining–KDD '15* (pp. 1721–1730). <https://doi.org/10.1145/2783258.2788613>
- Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, 18(1), 5–46. [https://doi.org/10.1016/0304-4076\(82\)90094-x](https://doi.org/10.1016/0304-4076(82)90094-x)

- Czajka, J. L., Hirabayashi, S. M., Little, R. J. A., & Rubin, D. B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business & Economic Statistics*, 10(2), 117–131. <https://doi.org/10.2307/1391671>
- Hagiu, A., & Wright, J. (2020). When data creates competitive advantage. *Harvard Business Review*, 98, 94–101. <https://hbr.org/2020/01/when-data-creates-competitive-advantage>
- Hernán, M. A., Hsu, J., & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *CHANCE*, 32(1), 42–49. <https://doi.org/10.1080/09332480.2019.1579578>
- Iacus, S. M., King, G., & Porro, G. (2009). cem: Software for coarsened exact matching. *Journal of Statistical Software*, 30(9), 1–27. <https://doi.org/10.18637/jss.v030.i09>
- Imai, K., & Kim, I. S. (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science*, 63(2), 467–490. <https://doi.org/10.1111/ajps.12417>
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press. <https://doi.org/10.1017/cbo9781139025751>
- Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy online controlled experiments*. Cambridge University Press. <https://doi.org/10.1017/9781108653985>
- Leek, J. T., & Peng, R. D. (2015). What is the question? *Science*, 347(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>
- Lopez Bernal, J., Cummins, S., & Gasparrini, A. (2016). Interrupted time series regression for the evaluation of public health interventions: A tutorial. *International Journal of Epidemiology*, 46(1), Article dyw098. <https://doi.org/10.1093/ije/dyw098>
- Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.
- Robins, J., & Hernan, M. (2008). Estimation of the causal effects of time-varying exposures. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 553–599). Chapman and Hall/CRC. <https://doi.org/10.1201/97814200115793>
- Rosenbaum, P. (2017). *Observation and experiment*. Harvard University Press. <https://doi.org/10.4159/9780674982697>
- Rosenbaum, P. R. (2010). *Design of observational studies*. Springer. <https://doi.org/10.1007/978-1-4419-1213-8>

- Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 29(1), 159–183. <https://doi.org/10.2307/2529684>
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701. <https://doi.org/10.1037/h0037350>
- Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge University Press. <https://doi.org/10.1017/cbo9780511810725>
- Sobel, M. E. (2012). Does marriage boost men’s wages? Identification of treatment effects in fixed effects regression models for panel data. *Journal of the American Statistical Association*, 107(498), 521–529. <https://doi.org/10.1080/01621459.2011.646917>
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1), 1–21. <https://doi.org/10.1214/09-sts313>
- Stuart, E. A., & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, 44(2), 395–406. <https://doi.org/10.1037/0012-1649.44.2.395>
- Thomke, S. H. (2019). *Experimentation works: The surprising power of business experiments*. Harvard Business Review Press.

©2020 Iavor Bojinov, Albert Chen, and Min Liu. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.

Footnotes

1. Uncontrolled release refers to any release that is not facilitated through an internal experimentation platform. For example, when the release is controlled by a third party such as Apple Store or Google Play. ↵
2. Exceptions include email campaigns sent by the marketer, or digital channels that can both randomize and report all metrics of interest. ↵
3. Companies can pay wireless carriers and original equipment manufacturers to preload applications on mobile devices. Preloads allow applications to be distributed with the hardware. ↵
4. For example, we measured the effect of receiving a connection request on member engagement using a cross-sectional causal inference method and found consistent estimates for the more active member cohorts, whose covariates are less sparse. Less engaged members were more difficult to model, but they do not

contribute much to the total metric value. By aggregating cohort-level effects into the overall effect, our estimate of the causal effect (3.2% to 4.3%) overlapped with the value reported by the randomized experiment (3.7%). [↩](#)