Challenges in Statistics: A Dozen Challenges in Causality and Causal Inference*

Carlos Cinelli, Avi Feller, Guido Imbens, Edward Kennedy, Sara Magliacane, Jose Zubizarreta

Abstract. Causality and causal inference have emerged as core research areas at the interface of modern statistics and domains including biomedical sciences, social sciences, computer science, and beyond. The field's inherently interdisciplinary nature—particularly the central role of incorporating domain knowledge—creates a rich and varied set of statistical challenges. Much progress has been made, especially in the last three decades, but there remain many open questions. Our goal in this discussion is to outline research directions and open problems we view as particularly promising for future work. Throughout we emphasize that advancing causal research requires a wide range of contributions, from novel theory and methodological innovations to improved software tools and closer engagement with domain scientists and practitioners.

1. INTRODUCTION

Causal inference is an interdisciplinary field focused on understanding cause-and-effect relationships, such as learning the effects of interventions, uncovering causal

Carlos Cinelli is Assistant Professor, Department of Statistics, University of Washington, Seattle, WA, USA (e-mail: cinelli@uw.edu). Avi Feller is Professor, Goldman School of Public Policy and Department of Statistics, University of California, Berkeley, Berkeley, CA, USA (e-mail: afeller@berkeley.edu). Guido Imbens is Professor, Graduate School of Business, and Department of Economics, and Director of the Standard Causal Science Center, Stanford University, Stanford, CA, United States (e-mail: imbens@stanford.edu). Edward Kennedy is Associate Professor, Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA, United States (e-mail: edward@stat.cmu.edu). Sara Magliacane is Assistant Professor, Informatics Institute, University of Amsterdam, Amsterdam, Netherlands (e-mail: s.magliacane@uva.nl). Jose Zubizarreta is Professor, Departments of Health Care Policy and Department of Biostatistics, Harvard University, Boston, MA, USA (e-mail: email).

mechanisms, and providing explanations for observed phenomena. Today there is a vibrant research community studying causality in a diversity of domains—spanning biomedical sciences, the social sciences, computer science, and more—often leveraging insights from one application area to make advances in others. Statisticians have played a prominent role in causal inference, dating back to Fisher and Neyman a century ago. As methodologists actively working in this area, our goal in this paper is to highlight both the challenges and opportunities for the next decades of statistical research in causal inference.

We are optimistic about this future, in part because the field is just getting started: despite statistics' historic role in causal inference, the broader discipline's embrace of causal inference as a core subject is relatively new. As recently as the 1990s, there were very few scholarly articles in statistics journals that focused explicitly on causality. Statistics textbooks typically did not cover causal inference, beyond admonishing readers not to confuse correlation and causation, and any explicit mention of causality was typically limited to discussions of randomized experiments.[1] There were no graduate or undergraduate courses devoted to causal inference,[2] and few presentations at ma-

---

[1] For example, in his Nobel lecture Daniel McFadden wrote, "detection of true causal structures is beyond the reach of statistics" and recommends that "For these reasons, it is best to avoid the language of causality" ([McFadden, 2001] p. 369), despite the fact that his work is very much about what we would now describe as causal.

[2] There were courses devoted to design of experiments. As far as we know the first course devoted entirely to causal inference in exper-

jor statistics conferences. There was little explicit attention to causal inference in other fields where statistical methods are used, including social and biomedical sciences. This is even more remarkable because causal effects are often precisely the quantities of interest in these fields.[3]

This state of affairs has changed dramatically over the last three decades. Today many of the leading statistics journals routinely publish causal inference research (see Figure 1), and there are now two specialized journals devoted exclusively to the subject, the *Journal of Causal Inference* (since 2013) and *Observational Studies* (since 2015). Graduate and undergraduate education in statistics and substantive fields regularly includes methodological courses devoted solely to causal inference. There is a scholarly society, the *Society for Causal Inference*. There are multiple university centers devoted to causality and causal inference,[4] and there are annual academic conferences focused entirely on methods for and applications of causal inference.[5]

Given this rapid progress, the question is what challenges and open questions remain. We highlight a dozen areas below:

1. Complex Experiments and Experimental Design;
2. Interference and Complex Systems;
3. Heterogeneous Effects and Policy Learning;
4. Mediation and Causal Mechanisms;
5. Optimality and Minimaxity;
6. Sensitivity Analysis and Robustness;
7. Reliable and Scalable Causal Discovery;
8. Aggregation and Synthesis of Causal Knowledge;
9. Automation of the Causal Inference Pipeline;
10. Benchmarks, Evaluation, and Validation;
11. New Identification Strategies; and
12. Large Language Models and Causality.

For each area, we provide some background and pose outstanding challenges. We note that the challenges outlined here are not comprehensive, and, for example, Mitra et al. [2022] discuss some additional ones.

Fig 1: Proportion of articles with the word "causal" in the title, abstract, or keywords, in *Annals of Statistics*, *Annals of Applied Statistics*, *Biometrics*, *Biometrika*, *Journal of the American Statistical Association*, *Journal of the Royal Statistical Society*, and *Statistical Science* from 1970 to 2024.

What are causality and causal inference? Before turning to challenges and open questions, we need a working definition of causal inference and what separates it from other subfields of statistics; we give a more technical introduction in Section 2. We note here that we do not draw a sharp distinction between causality and causal inference, with perhaps the term causal inference making a slightly closer link to statistical methods. Broadly, causal inference is about answering "what if"[6] questions, such as learning the effect of (possibly hypothetical) interventions, uncovering causal mechanisms, or unveiling the causes of observed effects. For instance, typical examples include estimating the effect of taking a drug on some health outcome, understanding whether having a college degree was the cause of someone's earnings, or why someone had an adverse outcome. In all cases we are interested in comparing outcomes that we by definition cannot observe together—what Paul Holland in a widely quoted phrase described as "the fundamental problem of causal inference" [Holland, 1986]. Because the focus is on comparing (factual) outcomes that we see with (counterfactual) outcomes that we do not see, the challenges in causal inference differ conceptually from those in predictive modeling. Fundamentally, this comparison raises questions referred to as *identification* [Fisher, 1961]: even in large samples we may not be able to learn answers to our questions from the data at hand.[7] For instance, simply

---

imental and observational studies was a graduate course Imbens and Rubin co-taught at Harvard in 1996.

[3]Interestingly there is still pushback against this. For example, Bailey et al. [2024] question in general "whether the researchers are interested in estimating a causal effect at all" but they admit that "in some fields the default answer may be a clear 'yes' (for example, in economics)."

[4]For example, the CAUSALab at Harvard University, the Causal Artificial Intelligence Lab at Columbia University, the Stanford Causal Science Center at Stanford University, the Center for Causal Inference at the University of Pennsylvania.

[5]For example, the American Causal Inference Conference (ACIC), the European Causal Inference Meeting (EuroCIM), the Causal Learning and Reasoning conference (CLeaR), and the Causal Data Science meeting.
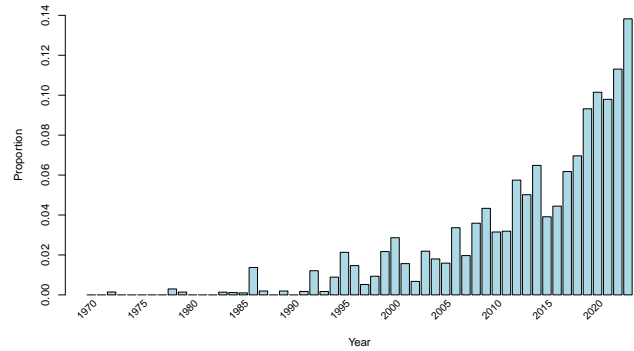
[6]The subtitle of one leading causal inference textbook, Hernan and Robins [2020].

[7]Exactly what identification means is not very well defined: Leamer [1983], in his seminal paper "Let's take the Con out of Econometrics" writes about a prominent researcher who "... mentions the phrase 'identification problem,' which, though no one knows quite

comparing observed outcomes for units exposed to different levels of a cause may tell us little about the causal effect due to *confounding*—units may differ across levels of the cause in ways related to the outcome. Moreover, these confounders are not always observed by the researcher, raising severe challenges for learning about the causal effects.

Dealing with these challenges requires assumptions that are substantive, or more precisely, causal—as opposed to solely statistical—in nature. Such identifying assumptions are by their nature untestable and require strong domain knowledge for drawing compelling, credible, causal conclusions. This makes causal inference intrinsically interdisciplinary, with models and methods developed by researchers in a variety of fields outside statistics for their specific substantive discipline, including computer science, political science, biomedical science, economics, and law. The interdisciplinary nature of causality and causal inference creates both challenges and opportunities. Differences in methods and terminology between fields can make it harder to recognize the commonalities in questions and problems. At the same time, insights from one discipline may have applications in other disciplines.

Where has causal inference been successful? Arguably, the most impactful research in causal inference has been on the design and analysis of randomized experiments, going back to the 1920s and 1930s. In particular, Fisher [1937] and Neyman [1923/1990] showed the power of randomization in inferring causal effects, initially in the context of agricultural experiments. These ideas ultimately led the Food and Drug Administration (FDA) in the United States to insist on randomized experiments as part of the drug approval process [Hill, 1990]. Today randomization remains the gold standard for causal inference. It has moved from being largely confined to biomedical settings to now being a staple of data analyses much more broadly, especially online in the private sector. For instance, large tech companies conduct hundreds of thousands of experiments every year to improve their products and services [Gupta et al., 2019].

More broadly, causal inference has had repeated success in sharpening substantive research questions, often by clarifying the underlying assumptions and the target quantity of interest. This has proved especially useful in applying novel methods from one field in others. A leading example is the *regression discontinuity design* (RDD), which was first developed by education researchers [Thistlethwaite and Campbell, 1960] and which remained a niche topic for decades [Cook, 2008]. Today, RDD is used widely across the social and biomedical sciences, largely driven by a common causal inference

framework that allows researchers in otherwise disparate domains to understand and apply the same methods. Another prominent example is *instrumental variables* methods, which originated in economics [Wright, 1928, Tinbergen, 1930][8], and spread throughout statistics in the 1990s [Angrist et al., 1996] and subsequently found new applications in biomedical settings through Mendelian randomization [Sanderson et al., 2022].

Finally, there are also many celebrated analyses that use the tools of causal inference. An early study in epidemiology, Snow [1856] inferred the causal mechanism for the spread of cholera through carefully documenting the location of cholera cases in relation to the pumps used for obtaining drinking water, in what would now be called a natural experiment.[9] Another classic example is assessing the causal effect of smoking on cancer, where a variety of studies documenting different pieces of evidence ultimately led to the consensus that there was indeed a causal relationship [Cornfield et al., 1959], despite the presence of detractors among leading statisticians, including Fisher [Fisher, 1958].

Where has causal inference failed? There have also been many failures of studies of causal inference. Often researchers have extrapolated insights from a study population to a target population that is quite different. For example, thalidomide had been prescribed as a sleep aid, but once it was prescribed to pregnant women it was found to lead to birth defects [Melchert and List, 2007]. These failures of *external validity* can plague both experimental and observational studies.

Another type of failure arises when researchers place excessive confidence in the assumptions required to identify causal effects. A prevalent example is assuming that all relevant confounders are measured without sufficient justification or adequate sensitivity analyses. Ed Leamer argued that "Hardly anyone takes data analyses seriously" Leamer [1983, p. 37], emphasizing the limited credibility of much empirical work in the social sciences, especially around the question of the deterrence effect of the death penalty on murder rates. The assumptions and models used to justify such estimates are often strong and arguably implausible; nonetheless, the full extent of this issue can go unnoticed.[10]

---

what he means, is said with such authority that it is totally convincing."

[8]Scholars debate whether the economist Philip Wright or his son, the geneticist Sewall Wright and inventor of path analysis, wrote the first exposition of instrumental variables. See Stock and Trebbi [2003] for stylometric evidence suggesting it was indeed Philip Wright.

[9]See Angrist and Pischke [2010] for a general discussion of natural experiments and the so-called credibility revolution in econometrics.

[10]This body of work is sometimes referred to as the "credibility crisis in economics" [Angrist and Pischke, 2010], a precursor to what Angrist and Pischke later termed the "credibility revolution." It is also related to the broader "replication crisis" in the social sciences, although that crisis raises other issues besides causality.

Finally, causal inference has frequently been used "for support rather than illumination," regulating science instead of supporting it [see, for example, Recht, 2025]. One consequence is that whole groups of researchers and modes of inquiry (*e.g.*, qualitative methods and case studies) have often been excluded from the field, leaving the discipline poorer and limiting the potential reach of causal inference ideas.

Organization. The rest of the paper is organized as follows. In Section 2 we describe current perspectives on causal inference and the state of the literature. In Section 3 we describe some of the main outstanding challenges for causal inference. In Section 4 we provide some concluding thoughts.

2. THE CURRENT STATE OF CAUSAL INFERENCE

To set the stage for the outstanding challenges and future directions for research on causal inference we briefly introduce the standard frameworks for causal inference, and then describe the current state of the field. More detailed discussions of these issues can be found in a quickly increasing number of textbooks on causal inference, including Cunningham [2018], Ding [2024], Hernan and Robins [2020], Huntington-Klein [2021], Imbens and Rubin [2015b], Morgan and Winship [2015], Pearl [2009], Pearl and Mackenzie [2018], Rosenbaum [2002, 2020] and Wager [2024].

2.1 Frameworks for Defining Causal Effects

One of the key achievements of modern causal inference research was the development of formal frameworks for articulating causal questions and causal assumptions, as well as deriving causal answers. Though their usage may differ to reflect traditions and idiosyncrasies of various fields, these frameworks are now broadly taught, acting as a common language for scientists across disciplines for understanding and discussing causal inference problems. Here we briefly introduce three closely related approaches, or languages, for causal inference used in the literature, highlighting some of their common aspects.

Potential Outcomes. Dating back to the 1920s [Neyman, 1923/1990], the *potential outcomes framework* originated in studies of randomized experiments. Donald Rubin subsequently generalized this framework to aid in the design and analysis of observational studies [Rubin, 1974]; see Imbens and Rubin [2015a] for a textbook based on this formulation.

This framework starts with the notion that each unit has multiple potential outcomes, each corresponding to a level of the treatment; these potential outcomes are the key objects in this approach. With a binary treatment $W_i \in \{0, 1\}$ and with some additional restrictions (see below), the two potential outcomes are $Y_i(0)$ and $Y_i(1)$, for the outcomes without and with treatment respectively.

Because the treatment can only take on one value for a particular unit (*e.g.,* for a specific individual at a specific time), at best only one of these potential outcomes can be observed, leading to Holland's celebrated "fundamental problem of causal inference." The realized and possibly observed outcome is the potential outcome corresponding to the treatment received, $Y_i \equiv Y_i(W_i)$. Causal effects are then defined as comparisons of these potential outcomes, *e.g.,* $Y_i(1) - Y_i(0)$. Assumptions are often formulated in terms of an *unconfoundedness* or *ignorability* assumption, [Rosenbaum and Rubin, 1983b]:

$$W_i \perp\!\!\!\perp \Big(Y_i(1), Y_i(0) \Big) \;\Big|\; X_i. \tag{1}$$

These assumptions entail conditional independence restrictions on the relationship between the level of the treatment and the potential outcomes, given some set of observed covariates or pretreatment variables X_i. For example, in a randomized experiment, the treatment would be guaranteed to be independent of potential outcomes by design, and thus the treatment assignment is ignorable or unconfounded. Other assumptions commonly used in the potential outcome framework include ruling out particular causal effects, so-called *exclusion restrictions*, or shape restrictions such as *monotonicity*. For example, in an instrumental variables setting [Imbens and Angrist, 1994, Angrist et al., 1996], one might start with potential outcomes $Y_i(w, z)$ indexed by the treatment W_i and an instrument Z_i, and make the assumption that the potential outcomes do not depend on the instrument, or $Y_i(w, z) = Y_i(w, z')$ for all z, z'. In addition, in the instrumental variables setting one postulates potential outcomes for the treatment, $W_i(z)$, for $z = 0, 1$, with the monotonicity assumption that $W_i(1) \geq W_i(0)$.

Structural Equation Models. Another framework for causal inference is to model mechanisms via a system of *structural equations*. This approach originated in the econometrics literature [Haavelmo, 1943, Strotz and Wold, 1960], with further development in the social sciences [Bollen, 1989, Duncan, 2014] and in computer science [Pearl, 2009]. Unlike the potential outcomes framework, this approach models the data generating process through a series of independent and local mechanisms, called structural equations. Interventions and causal effects in this framework are defined as the result of modifications on (parts of the) structural equations, while keeping the rest of the system untouched. While models were traditionally assumed to be linear, these can be fully nonparametric in modern treatments. Importantly, nonparametric structural equation models can be shown to be formally equivalent to the potential outcomes framework under some common assumptions [Pearl, 2009].

For instance, paralleling the previous example with unconfoundedness as in Equation (1), consider the following

nonparametric structural equation model:

$$X_i = f_x(U_{xi})$$
$$W_i = f_w(X_i, U_{wi})$$
$$Y_i = f_y(W_i, X_i, U_{yi}).$$

Here U_x, U_w and U_y stand for unobserved factors not modeled by the analyst, independent of each other. In this model, the potential outcome $Y_i(w)$ is defined as the solution of the system in which W_i is set experimentally to w, i.e., $Y_i(w) \equiv f_y(w, X_i, U_{yi})$. The assumption that unobserved variables are mutually independent implies the ignorability assumption (1). Exclusion restrictions in structural models are represented by the absence of a variable in the structural equation of another. Shape restrictions, such as linearity or monotonicity, can be encoded as constraints on the form of structural equations.

Graphical models. Graphical models or Directed Acyclical Graphs (DAGs) have emerged as a third language to encode causal assumptions, as well as to derive their logical ramifications, in an intuitive manner. Originating in the work of Sewall Wright [Wright, 1934], with additional discussions in economics, *e.g.,* Tinbergen [1940], Griliches and Mason [1972], their modern form is largely due to Peter Spirtes, Clark Glymour [Glymour et al., 1987, Spirtes et al., 2001], and Judea Pearl [Pearl, 1995, 2009]. In a causal graph, an arrow $W_i \to Y_i$ denotes that W_i may be a direct cause of Y_i. Its most general formulation does not impose any functional form or distributional assumptions about this causal relationship.

Figure 2 shows a typical causal diagram for an instrumental variables setting. The focus here is on the causal effect of the treatment W on the outcome Y. The complication is that there is an *unobserved confounder U* that affects both the treatment and the outcome, invalidating a direct comparison of units by treatment status. The instrument Z, which affects the treatment but has no direct effect on the outcome, can help in the identification of causal effects of the treatment on the outcome. The exclusion restriction is encoded by the absence of a direct path or arrow between Z and Y (although there is an indirect path from Z to Y through W); unconfoundedness of the instrument is encoded by the absence of arrows between U and Z and no unmeasured confounders of the instrument. Until recently, it has been difficult to graphically encode the monotonicity assumption, *i.e.,* that the effect of the instrument Z on the treatment W is nonnegative for all units; there are now several proposals [VanderWeele and Robins, 2010, Maiti et al., 2025], including by annotating arrows with a '+' sign. Finally, potential outcomes can also be directly represented in causal graphs; see for example, Twin Networks [Pearl, 2009], Single World Intervention Graphs (SWIGs) [Richardson and Robins, 2013], or Ancestral Multi-World Networks (AMWN) [Correa and Bareinboim, 2025].
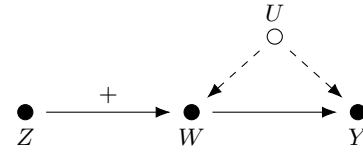

Fig 2: Example of a causal diagram for an instrumental variable model.

One way to connect the graphical model of Figure 2 with the previously discussed structural equations framework is to read it as encoding the following system of nonparametric structural equations,

$$Z_i = f_z(U_{zi})$$
$$W_i = f_w(Z_i, U_i)$$
$$Y_i = f_y(W_i, U_i),$$

along with the assumption that U_z and U are independent. Translating to potential outcomes, note this model encodes both the exclusion restriction, *i.e.,* $Y_i(w, z) \equiv f_y(w, U_i) = Y_i(w)$, as well unconfoundedness of the instrument, *i.e.,* $Y_i(w, z), W_i(z) \perp\!\!\!\perp Z_i$. The monotonicity assumption, denoted in the graph by $Z_i \xrightarrow{+} W_i$, can be interpreted as the functional constraint in the structural equations form, $f_w(1, u) > f_w(0, u), \forall\, u$, or, in the potential outcomes form, $W_i(1) > W_i(0), \forall\, i$.

Overall, similarly to how we can use different languages to express the same reality, these different languages for causal inference allow us to express the data generating process in various ways, and each of them can provide a more precise or convenient terminology depending on the setting. Understanding the connections between these frameworks, as well as possibly developing new frameworks and building blocks of causality will continue to be critical for advancing the field.

2.2 The central historical role of the case of binary treatments under unconfoundedness

Much of the causal inference literature from the 1980s onward has focused on a relatively narrow problem: how to estimate the average causal effect of a binary treatment on a scalar outcome in a setting where assignment to the treatment is ignorable; *i.e.,* a setting with no unobserved confounders. A key paper is Rosenbaum and Rubin [1983b], which introduced the ignorability assumption and the *propensity score*, the conditional probability of being exposed to the treatment given the confounders.[11]

---

[11]The Rosenbaum-Rubin paper has now, in 2025, almost 40,000 google scholar cites, accumulating currently at a still increasing rate of 3,000 per year.

This setting has led both to a vast and successful theoretical literature and to a huge empirical literature. Researchers have proposed a remarkably large number of different estimation approaches, including outcome modeling, inverse propensity score weighting, matching, and doubly robust methods based on combinations of outcome modeling and inverse propensity score weighting. Many of these are widely used in empirical practice, although very simple methods based on least squares analyses continue to be the most popular. See any of the modern textbooks for detailed discussions of these methods.

The corresponding theoretical literature has been very successful partly because it focused on a common problem. The drawback has been that it has also remained narrow. A key theme in our discussion is not only in pushing more causal inference research beyond this setting, but also in speeding up the practical adoption of existing causal inference advances beyond this binary case.

There are three ways in which this focus has been narrow. First, most of the work focused on the case with all confounders observed. This is unlikely the case in practice; we discuss many alternatives below. Second, most of the work has focused on the setting with a single binary cause. In practice, causes are often much richer, involving multiple causes, causes taking on many values, and complex systems. Third, the literature has also largely limited itself to estimating the effects of causes: what is the effect of a particular intervention. A different set of questions focuses on than causes of effects [Gelman and Imbens, 2013] or questions about attribution [Yamamoto, 2012]: why did this happen, or how much of what we see can we attribute to different causes. In practice there is also widespread interest in such questions.

3. CAUSAL INFERENCE: A LOOK TO THE FUTURE

In this section we discuss a dozen areas that we view as promising for future research on causality. In each case we lay out the general setting, and then discuss some specific challenges.

3.1 Complex Experiments and Experimental Design

3.1.1 Motivation For most of the 20th century, the causal inference literature on experimental design centered on agricultural and biomedical settings, typically with a modest number of fixed units (*e.g.*, patients or plots of land) and a small number of treatments (*e.g.,* active drug versus placebo); see Wu and Hamada [2011] for a modern discussion.[12] Over the past twenty years, however, the practice of experimentation has changed dramatically from these classical origins, driven in part by new technologies and the rapid growth of the tech sector.

Today, many experiments take place in online environments and are run by private organizations [Gupta et al., 2019]—a sharp departure from, among others, clinical trials that require FDA regulatory approval. Experimentation in offline environments can also be much more complex, such as *when* to treat in mobile health interventions [Nahum-Shani et al., 2018]. These modern experimentation settings raise new challenges and fundamental questions, and exhibit deeper connections with the parallel literature in machine learning.

3.1.2 Background.

Online and adaptive experimentation. In many modern settings units enter into the experiment sequentially, and outcomes can be measured soon after exposure. This combination of sequential enrollment and fast observation greatly expands the design space, especially by enabling researchers to adapt the design during the experiment. While these ideas have a long history in statistics [Efron, 1971] and in the bandit literature [Lai et al., 1985], in recent years there has been a large and very active literature on multi-armed bandits and sequential experimentation [Lattimore and Szepesvári, 2020]. For example, one prominent research direction is allowing the experimental design and analysis to vary across individual-level covariates, often referred to as "context" in this literature [Agarwal et al., 2014, Dimakopoulou et al., 2019]. Another recent push has focused on estimating average treatment effects after adaptive experiments [*e.g.,* Hahn et al., 2011, Kato et al., 2020]. Finally, there have been multiple extensions to settings with complex causal models, both with known [Lee and Bareinboim, 2018] and unknown causal structure [Lu et al., 2021, Bilodeau et al., 2022].

Complex experimental designs. A related challenge is that researchers today can control many more aspects of complex experiments. For instance, many modern experiments have a large number of possibly interacting treatments [Zhao and Ding, 2022], such as in conjoint experiments popular in survey research [Bansak et al., 2021]. This added complexity is also more common in non-randomized studies, in which greater care is needed to investigate interacting treatments [Yu and Ding, 2023]. Increasingly, researchers also have control over the timing of treatment, such as in mobile health interventions [Nahum-Shani et al., 2018, Kirgios et al., 2025] or when learning treatment policies [Nie et al., 2021]. Finally, the treatments themselves are becoming much more complex, especially with the increasing use of text and images as treatments [Egami et al., 2022]; see Section 3.12.

3.1.3 Challenges

Sequential treatments and increased adaptivity. The adaptive experimentation literature has largely focused on the case with a single treatment per unit. Current adaptive methods are largely confined to settings with few

---

[12]The literature on design of (often factorial) experiments in industrial settings was largely distinct.

constraints on the adaptivity. Much more challenging are settings with sequential treatments. The initial work in this area has relied heavily on detailed knowledge of the causal structure. In practice combining these questions with more limited knowledge about the causal structure will be important. Moreover, the adaptive nature of the experiments also raises questions about appropriately adapting always valid confidence sequences that are popular in online experimentation. Finally, there remain underexplored links to reinforcement learning [Sutton et al., 1998].

Platform trials. In connection with adaptive experiments, platform trials have gained considerable traction in the health sciences as a means to study new medical interventions in a timely manner. Unlike traditional trials that focus on specific treatments, platform trials center around particular diseases, allowing the evaluation of multiple interventions as they become available. These trials operate under a master protocol and often involve multiple sites, which helps in collecting broader samples and enhancing generalizability. Platform trials present interesting methodological challenges including the definition of estimands, robust estimation, and multiple comparisons [Santacatterina et al., 2025, Qian et al., 2024].

Automation and reducing frictions. In practice many frictions prevent researchers from implementing adaptive experimentation. Even if only some of these barriers are "statistical" in a traditional sense, our field is still responsible for working to reduce these barriers. One promising direction is the increased use of Large Language Models and AI agents to automate more complex designs; see Section 3.12. This introduces many additional challenges, including how to adapt pre-analysis plans and other best practices beyond simple settings.

3.2 Interference and Complex Systems

3.2.1 Motivation The classical causal inference paradigm assumes *no interference* between units [Cox, 1958]: a unit's outcome only depends on that unit's own treatment. This is typically formalized via the unfortunately-named Stable Unit Treatment Value Assumption [SUTVA; Rubin, 1980], that is, the potential outcomes depend only on the assignment for unit i: $Y_i(\mathbf{W}) = Y_i(W_i)$, where $\mathbf{W} = (W_i, \mathbf{W}_{-i})$ is the vector of treatment assignments for all units and $\mathbf{W}_{-i}$ is the corresponding vector excluding unit i. Historically, spillovers and other violations of the no-interference assumption were regarded as nuisances and were relatively rare. In agricultural experiments, plots of land do not interact with each other, at least after introducing sufficient space to avoid spillovers between adjacent plots; and in most classical medical settings, interventions applied to one patient do not affect other patients. Interference is a major concern in vaccine trials for infectious diseases: epidemiological concepts like *herd immunity* inherently violate the no-interference assumption. In such settings, interference need not be a nuisance but may be itself the quantity of interest.

In social science settings, however, spillovers and interference are often the rule, rather than the exception, putting the social into social sciences, and they are often the primary objects of interest. For example, tutoring some students in a class may affect their classmates not enrolled in the program. Providing information about new technologies to some farmers may lead them to share the information with other farmers not directly exposed to the new information. Changing the information about some rental properties in an online property rental environment may shift demand from control to treated properties, affecting the market equilibrium and so indirectly changing the experience for control properties.

3.2.2 Background Causal inference under interference is typically impossible to solve without additional structure: in a single experiment we observe only outcomes for a single vector of treatment assignments. Fortunately, there has been substantial progress on the vast area of restricted interference that lies between the extremes of no interference and arbitrarily complex interference, building off pioneering work from Sobel [2006], Hudgens and Halloran [2008] and Manski [2013]. One increasingly common technical tool is to define *exposure mappings* [Aronow and Samii, 2013, Manski, 2013], low-dimensional summaries of the full treatment assignment vector $\mathbf{W}$ and unit characteristics that capture how each unit is affected by the treatment status of other units. These summaries are then tailored to specific problems.[13] For instance, in a social network setting, interference could be well described by the average characteristics of an individual's friends [Ogburn et al., 2024]. In a marketplace, interference might be characterized by a summary measure like price [Munro et al., 2021]. In an education setting, spillovers can be summarized via the fraction of treated units in a peer group [Manski, 1993b, Carrell et al., 2013]. Finding effective structures for interactions and spillovers—and incorporating those structures or frameworks into other aspects of causal inference, such as experimental design—is an active area of research. Importantly, these questions typically involve delicate subject matter knowledge and are inherently interdisciplinary.

Example: Bipartite Graph. Consider the *bipartite graph* setting with two populations of units as illustrated in Figure 3; one prominent example is Papadogeorgou et al. [2019], which considers the impact of reducing power

---

[13]A complementary research thread instead asks what estimation is possible under potentially arbitrary interference or spillovers, viewing interference as a nuisance rather than a quantity of interest; see, for example, Sävje et al. [2021] and Leung [2022].
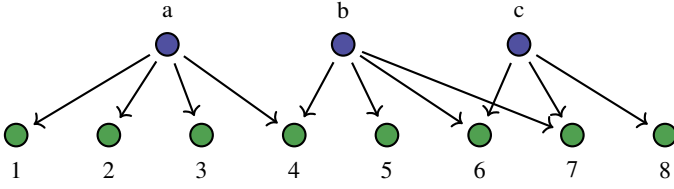
Fig 3: General bipartite interference graph with 3 interventional units (labeled a, b, c) and 8 outcome units (labeled 1, 2, 3, 4, 5, 6, 7, 8).

plant emissions on ambient ozone. The researcher can intervene on the first set of unit, the so-called *intervention units* (*e.g.,* power plants) by exposing them to different levels of treatment. Outcomes, however, are observed for a second set of units, the *outcome units* (*e.g.,* sensor locations). Intervention units may affect partially overlapping subsets of the outcome units. Designing experiments that allow researchers to learn effectively about policies of interest that have this form remains an open question [Lu et al., 2025].

Example: Experiments in Two-sided Marketplaces. Consider a marketplace where the researcher is interested in estimating the average effect of providing additional information to customers purchasing services or items. For example, Airbnb may provide videos or reviews to potential renters for some properties/providers. Marketplaces such as Airbnb in which many customers interact with many properties, have complex spillovers and are poorly suited to traditional experiments. Consider a traditional experiment that assigns some properties to the treatment group and some to the control group. Such experiments may lead customers to switch from control properties to treated properties without otherwise change a customer's overall rental level. Thus, simply reporting a difference in average outcomes could lead to misleading results with regard to the overall average effect of switching all properties between control and treatment. Bajari et al. [2023], Johari et al. [2022] instead consider a more complex randomization that assigns pairs of units—pairs consisting of one property and one customer—to treatment or control, as in the assignment matrix in Figure 4. Here some properties are assigned to a customer experiment where—for those properties only—all customers are randomly assigned to treatment or control. The remaining properties are assigned to a property experiment where the properties are randomly assigned to treatment or control. This randomization creates systematic variation in the fraction of treated properties for each customer and in the fraction of treated customers for each property, both of which enable direct estimation of the causal quantities of interest.

3.2.3 Challenges. Causal inference with interference is fundamentally an *Anna Karenina* problem, as in the

Fig 4: Possible randomization for customer and properties in a marketplace, following Bajari et al. [2023].

opening line of the Tolstoy classic: "All happy families are alike; each unhappy family is unhappy in its own way." Analogously, all no-interference applications are alike; each SUTVA violation violates SUTVA in its own way.

Balancing generality and practicality. One of the greatest challenge in this research area is to develop approaches that are general enough to apply to the many different types of interference applications, while also tackling important practical complications with each distinct type. Despite rapid progress in the past decade, the currently fragmented nature of the area has made it difficult for practitioners to see what is relevant for empirical work and to incorporate the latest advances into their research. For example, one divide is between interference that only depends on other units' treatment assignment, which is often the focus in the statistics literature [Hudgens and Halloran, 2008], and interference that depends on other units' treatments *and* outcomes, which is more common in the economics literature on peer effects [Manski, 1993a, Sacerdote, 2001]. Another divide is between research based primarily on potential outcomes rather than the extensive causal discovery and graphical causal model literature on interference, which primarily focuses on identification [see, for example Sherman and Shpitser, 2018, Sherman et al., 2020]. While there has been recent progress bridging these areas, we expect substantial gains from further work.

Incorporating substantive theory. Another central challenge is to better incorporate substantive theory into the analyses. For example, spillovers that affect units far away in some metric are rare but often of great importance—and finding such cascades has proved difficult empirically. Incorporating microfoundation network models of cascades could improve the ability to detect such spillovers.

One issue is that the theory often requires subject matter knowledge beyond the statistical expertise required for causal inference in settings without interference. A related concern is the difficulty of measuring spillovers, and reflecting any measurement difficulties in the resulting analysis [Bhattacharya et al., 2020].

Interference in observational studies. A third set of challenges concerns generalizing the methods that have been developed for experimental settings to the observational ones that are much more common. Understanding the experimental case with known assignment mechanisms is just the first step: spillovers are common in observational settings where challenges are bigger. For instance, in the standard no-interference case, addressing confounding typically involves adjusting for unit-specific characteristics. In the setting with spillovers the characteristics of other units may also matter, possibly making standard adjustment procedures insufficient.

Time evolution. The nature of interference may change over time. For example in Christakis and Fowler [2007], the authors have data on network connections at multiple points in time and find that the network linking individuals changes substantially. This may be in response to treatments, or in response to outcomes, in both cases requiring subtle adjustments to estimators for causal effects. See Gao [2024] for a discussion in the context of randomized experiments.

Statistical power. Statistical power remains a core issue with all but the largest empirical analyses of interference. A key challenge is to develop tools for applied researchers to better assess power in practice and to think about the appropriate notion of "replicates" under interference: intuitively, there is limited power to detect an interference pattern of interest that only has a few replicates. This question of replicates is closely related to the issue of external validity of causal inference studies with interference. While this is an issue in all studies, this is especially challenging without the standard formulation of *iid* draws from a common population. One interesting direction is to use adaptive designs based on the propensity score [Hahn et al., 2011], or, in network settings, to use graph-cutting techniques to balance bias and variance [Taylor and Eckles, 2018].

3.3 Heterogeneous Effects and Policy Learning

3.3.1 Motivation Much of causal inference has focused on simple summaries of causal effects, for example the average treatment effect, or average effect on the treated. Such global averages, however, can mask important heterogeneity; for example, if half of the subjects benefit and half are harmed, the average effect can be zero (see, *e.g.,* Figure 5).

On its own, effect heterogeneity is critical for scientific progress, as it can shed light on the mechanisms that

determine why some patients, but not others, respond to treatment. Effect heterogeneity is also central for targeting treatments or setting treatment policies, also known as *policy learning*. In the toy example above, if we knew which half of patients benefited from treatment, we could treat only them, avoiding harm to the other half. There are even more possibilities when the treatment takes on more than two levels, such as drug dosage, or when the treatment can vary over time.

3.3.2 Background For a binary treatment, effect heterogeneity is often quantified and explored via the conditional average treatment effect (CATE), defined as

$$\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x],$$

for covariates X_i (or possibly a subset of the available covariates). This estimand is a natural regression analogue of an average causal effect and plays a fundamental role in policy learning. Namely, in the binary treatment setting, let $\pi : \mathbb{X} \to \{0, 1\}$ denote some policy for assigning treatment based on covariates X_i, and let $Y_i(\pi(X_i))$ denote the potential outcome for unit i under policy $\pi(\cdot)$. The optimal treatment policy for maximizing the mean outcome is

$$\mathbb{1}\Big\{\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] > 0\Big\},$$

i.e., , the policy that simply thresholds the CATE at zero and treats everyone whose CATE is positive. Note that the maximization in the above is over all possible policies $\pi : \mathbb{X} \to \{0, 1\}$; however it is often of interest to only maximize over a smaller constrained class (possibly due to budget limits or interpretability), in which case the relationship to the CATE can be less clear. The field of policy learning is very large, spanning statistics, econometrics, machine learning, and more [Murphy, 2003, Hirano and Porter, 2009, Athey and Wager, 2021, Dudík et al., 2011], with connections to other large literatures on bandit problems and reinforcement learning [Bubeck and Cesa-Bianchi, 2012]. See Section 3.1 for related connections to experiments.

The CATE is fundamentally different from—and more difficult to estimate than—the ATE, analogous to how estimating a regression function is different from estimating a mean. For example, in an oracle setting where the potential outcome differences $Y_i(1) - Y_i(0)$ were actually observed, then the ATE could be estimated at parametric $(1/\sqrt{n})$ rates with just a sample average. However, estimating the CATE would require a regression of $Y_i(1) - Y_i(0)$ on covariates X_i; and classic no-free-lunch theorems from nonparametric regression [Stone, 1980, Györfi et al., 2006] tell us that no non-trivial rates of convergence can be achieved without extra assumptions about this regression function (*e.g.,* smoothness or discreteness of regressors). Differences become even more

stark outside the oracle setting, where potential outcomes must be estimated.

In early forays towards estimating the CATE, parametric and semiparametric models were popular, where the CATE was assumed to follow some particular (partly) parametric form (*e.g.,* constant, or linear in a subset of covariates) [Robinson, 1988]. More recently, methods using flexible nonparametric estimation and machine learning tools have become more popular, with a particular emphasis on double robustness [Athey and Imbens, 2016, Foster and Syrgkanis, 2023, Kennedy, 2023, Chernozhukov et al., 2018, Robins et al., 2008, Athey and Wager, 2021, Semenova and Chernozhukov, 2021, Van der Laan, 2006]. Similarly, earlier work on policy learning typically focused on randomized experiments with known propensity scores or parametric models where conditional effects can be estimated at parametric $1/\sqrt{n}$ rates [Kitagawa and Tetenov, 2018, Hirano and Porter, 2009]. Recent work has focused on the setting with unconfoundedness, allowing for flexible nonparametric models and for incorporating black-box machine learning tools [Athey and Wager, 2021, Luedtke and Chambaz, 2020].
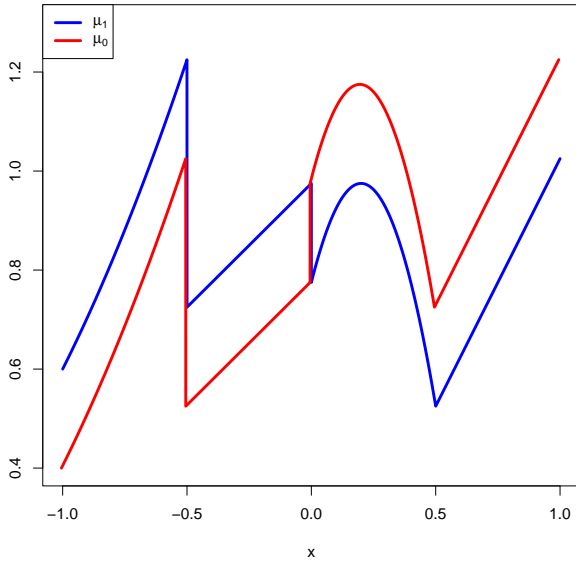

Fig 5: Example regression functions $\mu_w(x) = \mathbb{E}[Y_i \mid X_i = x, W_i = w]$ for which (i) the average causal effect $\mathbb{E}[\mu_1(X_i) - \mu_0(X_i)]$ is zero (*e.g.,* if $X \sim \text{Unif}[-1, 1]$), (ii) the individual regression functions μ_a are non-smooth and difficult to estimate accurately, and (iii) the CATE $\mu_1(x) - \mu_0(x) = -0.2 \times \text{sign}(x)$ is piecewise constant and very simple.

3.3.3 Challenges Much of the literature on heterogeneous effects and policy learning has focused on expectations, mostly in smoothness models, with relatively simple data structures (*e.g.,* single time-point binary treatments). There are some exceptions of course, including Wang et al. [2018], Leqi and Kennedy [2021], Bradic et al. [2019], Lewis and Syrgkanis [2021], and others. However there is much to explore outside this initial case.

Beyond expectations. Part of the focus on expectation-based contrasts in heterogeneous effect estimation stems from the connection to policy learning, and the fact that such contrasts are connected to mean-optimal treatment policies. However, these contrasts are often used in settings where explicit policy learning is not the main goal; thus there are potentially many other measures of heterogeneity that could be useful. These include heterogeneity based on variances and other features of the entire distributions. Further, it can be fruitful to explore policies that are optimal for objective functions that depend on other aspects of the distributions beyond means, such as inequality measures.

Beyond smoothness. Although smoothness models (where some number of derivatives are assumed bounded) are a standard framework from nonparametric statistics, many different forms of structure can be considered, each with their own nuances and advantages/disadvantages. Models relying instead on sparsity, or bounded variation, or neural network structure are underexplored for heterogeneous effect estimation, for example.

New data structures. Although studies with single time-point binary treatments are conceptually simple, and a natural starting point, real data are often much more complex, involving continuous or multivariate treatment options that may change over time, influenced by previous treatments and covariates that also change over time, for example. Further, covariates and outcomes can be high-dimensional and non-Euclidean (*e.g.,* graphs or images), there can be multiple studies from different sources, and subjects can be connected in complex networks [Ogburn et al., 2024, Kurisu et al., 2024]. There is much to do exploring heterogeneous effect estimation and policy learning in these non-standard but increasingly common data structures. Finally, a long line of work starting with Robins [1986] has focused on estimating optimal treatment regimes, that is, optimal dynamic policies; this is a much more challenging setting.

Inference. In nonparametric regression, there are fundamental challenges that come with inference (*e.g.,* constructing valid confidence bands) compared to estimation. These challenges are just as fundamental for the CATE, and so have led to a dearth of inferential versus estimation procedures. For example, optimizing mean squared error requires balancing squared bias and variance, but this complicates inference, since one is left with a non-trivial

bias that does not shrink when scaled by the standard error; Wasserman [2006] calls this "the bias problem". Further, although one can construct estimators that automatically adapt to unknown smoothness or other structure, there are strong negative results indicating how confidence bands cannot adapt in the same way [Low, 1997, Genovese and Wasserman, 2008]. These exact same issues arise for the CATE, and can be exacerbated due to the fact that the potential outcomes are not directly observed, and so there is additional interplay with nuisance estimation error. More work is needed to understand the role of these inferential challenges in causal-specific problems (*e.g.,* what are the implications for policy implementation), and to consider how some proposed fixes work for causal inference specifically (*e.g.,* bias reduction via debiasing or undersmoothing, settling for covering simpler CATE surrogates, or accepting weaker coverage guarantees).

New optimization goals. As described above, the classic setup in policy learning is to estimate policies $\pi(\cdot)$ that maximize the mean outcome $\mathbb{E}[Y_i(\pi(X_i))]$, often under assumptions of no unmeasured confounding and overlap, for example. However, practitioners often care about more than mean outcomes, and also want to balance optimizing outcomes with, for example, implementing policies that assign treatments in fair ways, or respect constraints on treatment availability, or other constraints. Further, in practice the typical identifying assumptions may be violated, or the goal may be framed in terms of the joint distribution of potential outcomes, in which case one could bound the optimal policy or identify a policy with minimal worst-case regret. Although there have been recent advances in this vein [Kallus and Zhou, 2021, Ben-Michael et al., 2024], there may be other important alternative approaches to explore, *e.g.,* in settings with distribution shift, where policies may be implemented in new populations that differ from that used in training.

Implementation, translation, and beyond. As is true in many areas of causal inference, for heterogeneous effect estimation and policy learning there is also a gap between modern theory and methods, and practical implementation. More work is needed studying how to apply state-of-the-art methodology in practice to substantive research problems, and making statistical tools more available and accessible.

Finally, one can construct heterogeneous treatment effect-based versions of many if not all of the challenges presented in this paper, to obtain other new and interesting challenges. For example, heterogeneous effects in complex experiments, with interference, combined with sensitivity analysis, from a discovery perspective, with automation, *etc.*, are all interesting and valuable areas to pursue. Compared to the versions of these problems involving more standard summary effects, versions focusing on heterogeneous effects and policy learning may come with substantial nuances and differences.
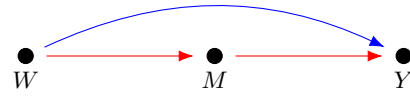

Fig 6: Directed acyclic graph for the standard mediation setup. M is a mediator variable on the path between treatment W and outcome Y. The blue path represents the *direct effect* of W, while the red path shows the *indirect effect* of W through M.

3.4 Mediation and Causal Mechanisms

3.4.1 Motivation. Much work in causal inference has focused on quantifying treatments' effects on an outcome. Even gold-standard experiments "reveal but do not explain causal relationships" [Bullock et al., 2010]. However, it is often of interest to go beyond a black-box assessment of *whether* treatment works, and to also learn *how* it works; this is the goal underlying mediation analysis.

Figure 6 shows a directed acyclic graph in a standard mediation setting. Here the goal may be not only to understand how treatment W affects outcome Y, but also how much W affects Y *directly* versus *indirectly* through a mediator M. Note the differences between Figure 2 with a DAG for an instrumental variables setting. Although in both cases we have a three variable setting with a causal path from the left to the right, the estimands and assumptions are quite different. In the instrumental variables case the direct link between the variable on the left (the instrument Z in Figure 2) and the outcome Y is absent, and there is an unobserved confounder affecting both the variable in the middle and the outcome. Understanding such mediation effects can be useful for many purposes. To make this more specific, consider Chen et al. [2019]. They study the effect of having a younger brother on a first child's education, surmising there may be a direct effect by taking away resources that are spent on the brother instead, and an indirect effect that having a brother may lead to fewer siblings overall (if there is a general desire for male children). Understanding these effects may help design educational interventions that affect children's outcomes.

Alternatively, furthering our understanding of the scientific mechanism of a treatment could be important as an end in and of itself.

3.4.2 Background. Mediation analysis has its origins in linear structural equation models [Wright, 1921] and was popularized by Baron and Kenny [1986] in a linear regression setting. Robins and Greenland [1992] and Pearl [2001] later formalized mediation analysis within an explicit, nonparametric causal inference framework. Since then, mediation analysis and adjacent areas have become a rich and active area of research, with important

applications in the health and social sciences, and emerging areas such as algorithmic fairness and policy evaluation. See VanderWeele [2015] for more of an overview.

Notation and common estimands. One of the earliest and most widely used estimands in mediation analysis is the *controlled direct effect* (CDE). Let $Y_i(w,m)$ denote the potential outcome that would be observed if both the treatment were set to $W_i = w$ and the mediator to $M_i = m$. The CDE of changing the treatment from w' to w, while holding the mediator fixed at m, is given by:

$$\text{CDE}_m(w,w') := \mathbb{E}\left[Y_i(w,m) - Y_i(w',m)\right].$$

This estimand captures the idealized setting of controlled experimentation, in which the scientist isolates the direct effect of the treatment by manipulating it while holding potentially mediating variables constant.

While conceptually straightforward, controlled effects can present limitations in answering mediation questions. Importantly, CDEs fail to capture a core objective of mediation analysis, which is indirect effects, *i.e.,* the effect of the treatment on the outcome that occurs only because the treatment influences the mediator. One can consider controlled effects of the mediator changing from m' to m, while holding the treatment fixed at w, *i.e.,*

$$\mathbb{E}\left[Y_i(w,m) - Y_i(w,m')\right],$$

but this can be non-zero even when the treatment does not affect the mediator, and so is not a proper indirect effect.

These considerations lead to the ideas of *natural* direct and indirect effects. Defining these objects require not only considering counterfactual mediators under different treatments, $M_i(w)$, but also nested counterfactuals, such as $Y_i(w, M_i(w'))$—that is, the potential outcome had the treatment been set to w, while letting the mediator take its natural value had the treatment been set to w'. Then, the natural direct effect of changing the treatment from w' to w, while letting the mediator take its natural level under w is defined as

$$\text{NDE}_w(w,w') := \mathbb{E}[Y_i(w, M_i(w)) - Y_i(w', M_i(w))].$$

Similarly, the natural indirect effect of holding the treatment fixed at w', while allowing the mediator to vary as if we had changed the treatment from w' to w, is

$$\text{NIE}_{w'}(w,w') := \mathbb{E}[Y_i(w', M_i(w)) - Y_i(w', M_i(w'))].$$

Importantly, now the average treatment effect can be decomposed into a sum of the natural direct and indirect effects,

$$\mathbb{E}[Y_i(w) - Y_i(w')] = \mathbb{E}[Y_i(w, M_i(w)) - Y_i(w', M_i(w'))]$$
$$= \mathbb{E}[Y_i(w, M_i(w)) - Y_i(w', M_i(w))]$$
$$\quad + \mathbb{E}[Y_i(w', M_i(w)) - Y_i(w', M_i(w'))]$$
$$= NDE_w(w,w') + NIE_{w'}(w,w').$$

Other decompositions of the ATE are possible, accounting for interactions between the treatment and the mediator [*e.g.,* VanderWeele, 2014].

Controlled and natural effects can both be useful in characterizing mediation, depending on the setting. Controlled effects can be viewed as representing direct manipulation, while natural effects can be more representative of underlying causal mechanisms [Pearl, 2001].

There are also randomized interventional versions of natural effects [Robins, 2003, Didelez et al., 2006, van der Laan and Petersen, 2008], where the mediator is not set to its actual values $M_i(w)$ or $M_i(w')$, but instead to random draws from the corresponding conditional distributions, *e.g.,* to some $G_i(w) \sim dP(M_i(w) \mid X_i)$.

Identification & estimation. Since controlled effects are solely effects of a joint exposure (W, M), their identification follows as in other causal settings, *e.g.,* via no unmeasured confounding in the form $(W, M) \perp\!\!\!\perp Y(w,m) \mid X$, along with consistency and positivity. For natural effects, identification is somewhat more involved. The terms in natural effects involving counterfactuals of the form $Y_i(w, M_i(w)) = Y_i(w)$ can be identified as usual, since this is just an effect of setting W. However, both natural direct and indirect effects also involve counterfactuals of the form $Y_i(w, M_i(w'))$, where the treatment is set to some w but the mediator is set to its counterfactual under a different treatment w'. The quantity $\mathbb{E}[Y_i(w, M_i(w')) \mid X]$ can be identified as

$$\int \mathbb{E}[Y \mid X, W = w, M = m]\, dP(m \mid X, W = w')$$

under the same consistency, positivity, and no unmeasured confounding assumptions used for direct effects, together with two additional assumptions. The first is no unmeasured confounding for the effect of treatment on mediator, $W \perp\!\!\!\perp M(w') \mid X$, which is analogous to standard assumptions. The second is known as *cross-world exchangeability*, $M(w') \perp\!\!\!\perp Y(w,m) \mid X$, which is a qualitatively different assumption since it concerns counterfactuals under different treatments w and w'. In particular, cross-world exchangeability cannot be enforced experimentally by randomization and cannot be falsified [Miles, 2023], although the randomized interventional versions of these effects can be identified without cross-world conditions.

After identification, estimation can proceed much as in other causal problems, for example via outcome modeling, inverse propensity score weighting, matching, semiparametric doubly robust methods, *etc*. Semiparametric theory was developed by Tchetgen Tchetgen and Shpitser [2012], who showed corresponding estimators are "triply robust". Although this sounds more favorable than "doubly robust", it is actually less so—in mediation there are three nuisances instead of two (usual treatment and outcomes distributions, as well as the mediator distribution),

and two of the three nuisances need to be estimated consistently (rather than one of two in the doubly robust case).

3.4.3 Challenges

Designing experiments for mediation analysis. While most mediation analysis are based on observational studies, experimental designs can provide stronger causal evidence of the underlying mechanisms. An interesting methodological challenge is the design of experiments that can uncover mediation effects. One promising approach is sequential randomization. For instance, participants might first be randomized to a smoking cessation program, after which their motivation is measured, and subsequently they can be randomized again to receive a motivational enhancement. This sequential assignment can help distinguish effects on the mediator from those on the outcome. Beyond sequential randomization, factorial designs allow simultaneous manipulation of both the treatment and the mediator, provided direct manipulation of the mediator is feasible. Encouragement designs can be employed when the mediator cannot be directly manipulated, by randomizing an encouragement or incentive intended to shift the mediator. These experimental strategies can be developed depending on the context, enabling stronger causal claims about mediation pathways.

Representing complex mediators. Another exciting challenge arises when dealing with high-dimensional, unstructured mediators, such as those present in multiomics, digital trace, or text data. In these contexts, there is often no clear mechanistic theory that specifies the mediators in advance; instead, the mediators must be learned, summarized, and interpreted directly from the data at hand. This raises interesting questions, including how to define appropriate causal estimands and articulate meaningful identification assumptions. Addressing this gap also presents opportunities to leverage advances in representation learning and related areas [Schölkopf et al., 2021] to detect and characterize mediators in a data-driven manner. The overarching goal is to develop causal models that can succinctly capture the effects of such mediators, enabling inference and interpretability in settings where there are many candidate mediating pathways both diffuse and overlapping in nature.

Decomposing causal pathways. Another relevant direction is the decomposition of causal effects, moving beyond traditional direct and indirect effects to assess, for example, disparities in health care access driven by gender or race [*e.g.,* Jackson, 2021]. This line of research is closely connected to issues of fairness and opportunity. However, these questions are particularly challenging, as causal effects and social constructs like race are deeply intertwined and often difficult to disentangle, even at a conceptual level. Nevertheless, this area is crucial for both policy and fairness considerations, making it essential to develop clear frameworks and language to rigorously discuss these complexities [Howe et al., 2022].

Studying time-varying mediation. From a dynamic viewpoint, it is important to understand how direct and indirect effects evolve over time [VanderWeele and Tchetgen Tchetgen, 2017]. A compelling emerging application is in digital health interventions, where app usage influences behavior change, which, in turn, affects health outcomes. Among others, these settings are challenging because of complex direct and indirect effects through multiple time points. Current methodological approaches include G-methods and Marginal Structural Models (MSMs) [*e.g.,* Aalen et al., 2020, Mittinty and Vansteelandt, 2020, Díaz et al., 2023]. Theoretical challenges involve defining estimands that are both meaningful and based on realistic assumptions, as well as decomposing effects through multiple mediator pathways. Practical challenges include sparse data across numerous causal pathways, issues of dropout and missing data, and the need to appropriately model time trends.

Generalizing mediation effect estimates. Often, just as it is crucial to understand the effects of an intervention in a target population, so too is uncovering the mechanisms underlying those effects in the population of interest. This requires generalizing or transporting estimates of direct and indirect (mediated) effects from a study population to a target population that may differ in covariate distributions, effect modifiers, and potentially in mediators themselves. Key challenges in this context include unmeasured confounding, the need for robust estimation approaches, and reliance on strong assumptions involving cross-world counterfactuals.

Conducting mediation analysis in practice. In a guide for applied researchers, Schuler et al. [2024] provide a thoughtful discussion of practical challenges encountered in mediation analysis. Among these, they highlight the importance of selecting mediation effects that accurately reflect the scientific question of interest, assessing the validity of the underlying assumptions of no unmeasured confounding, and addressing measurement error in the mediator. They also underscore the importance of comprehensive and transparent reporting of the results of mediation analyses. Addressing these challenges can greatly foster mediation analysis in practice.

3.5 Optimality and Minimaxity

3.5.1 Motivation Distinct issues arise in causal effect estimation, subsequent to identification. In particular, researchers often want to use flexible methods that do not rely on strong modeling assumptions. This raises important questions, such as: What kind of statistical structure are we willing to assume? How do we build estimators that work well under these assumptions? What is the *best possible performance* we could hope to achieve under our

assumed structure? The minimax framework is a natural and powerful one for considering these kinds of questions, with a long history in statistics going back half a century [Le Cam, 1973, 1986, Stone, 1980, Tsybakov, 2008].

Although minimax optimality is well understood in some settings, such as regression and density estimation, relatively little is known when it comes to causal inference problems. There is a pressing need to address this gap: it means that in causal inference we often do not know whether we could be making better, more efficient use of the data, whether there are better estimators out there waiting to be discovered, or if what we have is the best we could ever get. This can lead to a proliferation of different new methods, with little clarity about how to compare or benchmark them. And this lack of clarity is uniquely concerning in causal settings, where empirical benchmarking can be inherently limited due to lack of a ground truth [Holland, 1986]. In contrast, the minimax framework provides a path forward for *theoretical benchmarking* in causal inference—even if this path is rife with technical challenges.

3.5.2 Background We first outline the basic setup for minimax estimation; a very useful review can be found in Tsybakov [2008]. We start out with some desired statistical target, denoted ψ. This could be a function, like a conditional expectation $\mathbb{E}[Y \mid X = x]$, or a real-valued parameter (*i.e.,* functional), like the causally motivated quantity $\int \mathbb{E}[Y_i \mid X_i = x, W_i = 1] \, dP(x)$, which equals $\mathbb{E}[Y_i(1)]$ under no unmeasured confounding assumptions. Next, we decide what we want to assume statistically, which amounts to settling on a model $P \in \mathcal{P}$, *i.e.,* a set of distributions we assume contains the truth. Common examples include parametric models, where distributions only differ up to finitely many parameters, or nonparametric models, where distributional components are infinite-dimensional and only assumed to have some smoothness, sparsity, bounded variation, *etc*.

Given an assumed model $\mathcal{P}$, the minimax rate for estimating the statistical target $\psi = \psi(P)$ with loss ℓ is given by

$$R_n = \inf_{\widehat{\psi}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left\{ \ell \left(\widehat{\psi}, \psi(P) \right) \right\}$$

where the infimum is over all possible estimators, *i.e.,* all possible functions of the data. In words, this is the best possible (worst-case) estimation error, across all possible estimators. In practice, minimax rates are typically characterized in two steps: first we find a lower bound, showing that $R_n \geq C r_n$ for some constant C and sequence r_n, and second we find an upper bound, showing that $R_n \leq C' r_n$ for some possibly different constant C'. Then we can conclude the minimax rate is r_n. A lower bound on the minimax rate is a powerful result: it says

that no estimator can have risk (*i.e.,* expected loss) uniformly smaller than $C r_n$, no matter how creatively it is constructed, or how much computation it involves.

Minimax rates thus have crucial implications, both practical and theoretical. First, they give a precise benchmark for the best possible performance of a statistical task. If some estimator's risk does not match a minimax lower bound, then more work is needed—either at improving the estimator (or its analysis), or improving the bound. Alternatively, if some estimator's risk *does* match a lower bound, then it cannot be improved (at least in terms of rates), without adding or changing assumptions. Second, minimax rates precisely characterize the fundamental limits of estimation in a given problem. This allows us to compare different tasks, and order them in terms of their inherent statistical difficulty in interesting ways.
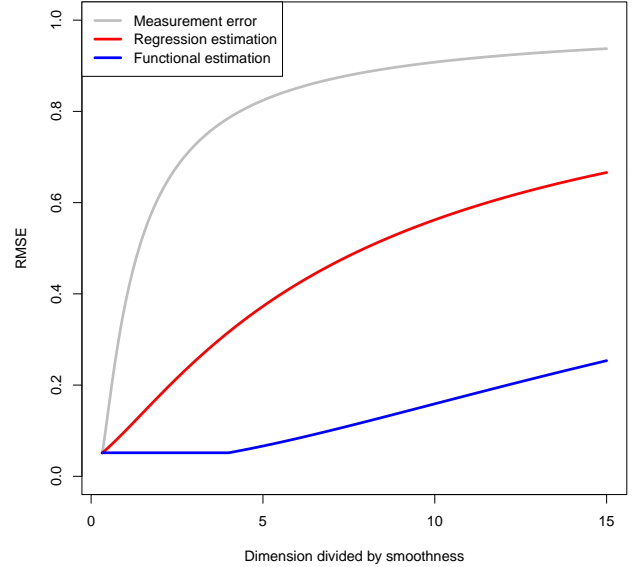

Fig 7: Minimax rates for three classical estimation tasks, as a function of covariate dimension d scaled by smoothness s: functional estimation (*e.g.,* estimating the ATE in causal inference); regression estimation (*e.g.,* the CATE); and density estimation with measurement error (*e.g.,* the counterfactual density when outcomes are measured with error). In causal analogues of these problems the rates further depend on the complexity of nuisance functions (*e.g.,* propensity scores, outcome regressions).

In some settings, much is known about minimax rates. For example, standard parametric models have been studied for decades and precise local rates are available [Van der Vaart, 2000]. For smooth nonparametric models, where functions are only assumed to have s bounded

derivatives, minimax rates for d-dimensional nonparametric regression and density estimation are of the form

$$n^{-1/(2+d/s)}$$

in terms of RMSE [Stone, 1980]; see Figure 7 for an illustration. This is like the parametric rate $1/\sqrt{n}$, but with an added penalty for how large the dimension d is relative to the smoothness s. For quadratic functionals in smooth nonparametric models, such as the expected density $\mathbb{E}\{p(X_i)\} = \int p(x)^2 \, dx$, the minimax rate has been shown to be

$$\max\left(\frac{1}{\sqrt{n}}, \ n^{-1/(1+d/4s)}\right)$$

[Birgé and Massart, 1995, Laurent, 1996]; see Figure 7 for an illustration (and note the "elbow" shape). This shows how estimating functionals of a density can be statistically easier than estimating the whole density itself: the rate is parametric when $s \geq d/4$, and otherwise the penalty driving the rate away from parametric is weaker than for density estimation, depending on $d/4s$ instead of d/s. For smooth density estimation with Gaussian measurement error, the minimax rate is

$$(\log n)^{-s/2}$$

[Carroll and Hall, 1988], which is very slow, indicating the inherent difficulty of the problem; see Figure 7 for an illustration. Thus in this problem one either needs to live with large estimation errors or else add more structure/assumptions.

Although minimax optimality has a relatively long history in the aforementioned settings, far less is known when it comes to causal inference. The causal inference setting where most is known about minimax rates is in estimation of sufficiently smooth (*e.g.,* pathwise differentiable) target parameters, when they can be estimated at parametric rates. In this setting, there are precise convolution theorems for regular estimators, and local asymptotic minimax rates available [Hájek, 1972, Bickel et al., 1993]. However, these rates are not always attainable, *e.g.,* if covariate dimension is too large or nuisance function smoothness is too low. Thus, over the last decade or so, Robins and colleagues have developed extensive results for pathwise differentiable functionals *even when parametric rates are not achievable*, in smooth nonparametric models [Robins et al., 2008, 2017].

3.5.3 Challenges

New parameters. Even for the most basic causal inference parameter $\psi(P) = \int \mathbb{E}[Y_i \mid X_i = x, W_i = 1] \, dP(x)$ (and even in classical smoothness models), there are still open and unsolved problems regarding the role of the covariate distribution. When one moves to other more complicated parameters, many more open problems arise. For example, stochastic intervention effects are by now of

wide interest [Kennedy, 2019, Díaz and van der Laan, 2012, Haneuse and Rotnitzky, 2013], but optimality has not been explored outside the $\sqrt{n}$ setting. Similarly, optimality for time-varying treatment effects outside the $\sqrt{n}$ regime is entirely open, regardless of the model setup. Kennedy et al. [2024] recently characterized minimax rates for heterogeneous effect estimation in smoothness models, but there is little work outside this regime. For similarly complex parameters like dose-response curves [Kennedy et al., 2017] or counterfactual densities [Kim et al., 2018], little is known.

Adaptivity. Most available results and methods in the context of optimality have been developed in settings where relevant complexity parameters, such as the amount of smoothness, are known (with few exceptions, *e.g.,* Mukherjee et al. [2015]). Adaptivity is a fundamental aspect of nonparametric estimation [Donoho and Johnstone, 1995, Efromovich and Pinsker, 1996, Low, 1997], but it has not been explored in detail in causal inference problems. Thus there are many fundamental open questions of how and to what extent it is possible to adapt to unknown smoothness, or more generally, across sets of different function classes.

New models. The large majority of work on minimax optimality in causal inference considers smoothness models, where nuisance functions are assumed to have some number of bounded derivatives [Robins et al., 2009, Kennedy et al., 2024]. (Recall functions with more bounded derivatives are less complex, and easier to estimate.) Smoothness models are a natural starting point: they are the prototypical nonparametric function class, for which classic kernel/series methods can work well, and which are infinite-dimensional but also structured enough to allow for fast convergence rates, depending on dimension/smoothness. But many other models are also of interest (and potentially more practical), including high-dimensional models [Bradic et al., 2019, Liu et al., 2023, Zeng et al., 2024], bounded variation models [van der Laan, 2017], structure-agnostic models [Balakrishnan et al., 2023, Jin and Syrgkanis, 2024], neural network models [Farrell et al., 2021], and more.

Implementation. In the relatively few settings where estimators are available that can improve on doubly robust-style methods (for example, using higher-order methods [Robins et al., 2008, 2017], undersmoothing [Newey and Robins, 2018, McGrath and Mukherjee, 2022, McClean et al., 2024], or exploiting linearity in high-dimensions [Athey et al., 2018, Smucler et al., 2019, Wang and Shah, 2024, Liu et al., 2023]), the estimators are often computationally intensive and require careful tuning. There are many opportunities to make these methods more practical, automatic, and user-friendly.

New frameworks. There are variations on the minimax optimality framework described above, explorations of which present new opportunities that may be particularly important in causal inference problems. For example, in many statistical problems it is common to use straightforward losses, *e.g.,* squared error $\ell(\widehat{\psi}, \psi_P) = (\widehat{\psi} - \psi_P)^2$, or integrated L_2 or supremum errors for functions. However, in causal problems it may be worthwhile to consider new losses, *e.g.,* that connect to the threat of unmeasured confounding, or are derived directly from feedback from policymakers in some way.

Similarly, expected loss may not always be most meaningful; recent work has explored minimax quantiles or confidence interval lengths, for example [Ma et al., 2024]. Another variant is to consider local minimax optimality, where the supremum in R_n is not over the whole model $\mathcal{P}$, but instead over a subset near some specific distribution P, *i.e.,* over all distributions Q such that $d(Q, P) < \epsilon$ for some distance d [Van der Vaart, 2000, Balakrishnan and Wasserman, 2019]. Local minimax rates are more nuanced and can illustrate how fundamental limits can change with a particular distributions P. Beyond these variations, it may be worthwhile developing other new frameworks for optimality, which depart even further from the minimax setup.

3.6 Sensitivity Analysis and Robustness

3.6.1 Motivation. Canonical methods used for drawing causal inferences from observational data almost by definition rely on strong, often untestable assumptions about the data generating process. For example, estimating causal effects through covariate adjustment assumes no unobserved confounders between the treatment and the outcome; instrumental variable methods require both an assumption of no unmeasured confounders between the instrument and the outcome, and an exclusion restriction that rules out direct effects of the instrument on the outcome; difference-in-differences designs assume that, in the absence of treatment, the average outcome for the treatment and control groups would have evolved in parallel over time.

While such assumptions may provide a useful benchmark and starting point for causal analysis, they are often unlikely to hold exactly in real-world settings. This raises several questions: how sensitive are causal inferences to violations of such key assumptions? Or, how large would deviations from such assumptions need to be to substantively change the main results of a study? And how plausible are such deviations in a given context? Probing the *robustness* of causal claims to departures from causal assumptions is an important and active area of research known as *sensitivity analysis*.

3.6.2 Background. The first—and arguably most successful—empirical application of sensitivity analysis is the seminal study by Cornfield et al. [1959], which assessed the causal nature of the association between cigarette smoking and lung cancer. In the 1950s, several observational studies found that smokers were substantially more likely than non-smokers to develop lung cancer [*e.g.,* Doll and Hill, 1950]. At the time, however, Sir Ronald Fisher and other prominent statisticians argued that, in the absence of experimental evidence, unobserved confoudners, such as an individual's genotype, could provide an alternative explanation for the observed association [Fisher, 1957, 1958].

The now-classic sensitivity analysis by Cornfield et al. [1959] demonstrated that, even if such an unobserved confounder existed, this hypothetical "smoking gene" would need to be at least nine times more prevalent in smokers than in non-smokers to fully account for the observed association—something deemed scientifically implausible by most experts then and now. Absent such a strong relation, they concluded that a causal link between cigarette smoking and lung cancer is still necessary to explain the observed association, even allowing for plausible amounts of unmeasured confounding. Building on Cornfield's seminal work, an extensive body of literature has developed quantitative frameworks for sensitivity analysis. Here we present a non-exhaustive sample of methods developed to perform sensitivity analysis to unobserved confounders.

A common thread of many approaches is to assume that unconfoundedness holds, but only after adjusting both for observed covariates X_i as well as for an unobserved covariate U_i, that is,

$$W_i \perp\!\!\!\perp \Big(Y_i(1), Y_i(0) \Big) \ \Big| \ X_i, U_i.$$

Given this augmented unconfoundedness assumption, researchers then impose limits on the association between the unobserved confounder and the treatment, or on the association between the unobserved confounder and the outcome, or on both. Alternatively, researchers can also compute the minimal strength of association of U_i such that it would invalidate the main conclusions of the original study.

One of the earliest works in this area is given by Rosenbaum and Rubin [1983a], who posited a parametric model for the relationship between potential outcomes and both observed and unobserved confounders, along with a parametric model for the propensity score as a function of these (observed and unobserved) confounders. Given specific values for the parameters attached to U_i governing these relationships, they then examined how variations in these parameters affect the estimated average treatment effect. Imbens [2003] and Cinelli and Hazlett [2020] extended this approach by linking the range of

values for the sensitivity parameters to estimable associations given observed confounders. Other related approaches can be found in Vanderweele and Arah [2011], Dorie et al. [2016], Altonji et al. [2005], and Veitch and Zaveri [2020].

For binary treatments, a common approach is to limit solely the strength of confounding by placing bounds on the odds ratio of the treatment assignment distribution, conditioning and not conditioning on potential outcomes. This model has been extensively developed by Rosenbaum [1987, 2009] and more recently by Tan [2006], Yadlowsky et al. [2018], Kallus and Zhou [2018], Zhao et al. [2019b], Jesson et al. [2021], Dorn and Guo [2023], Dorn et al. [2024]. Alternative approaches for binary treatments include Masten and Poirier [2018], who place bounds on the difference between these distributions, and Bonvini and Kennedy [2022], who limits the fraction of units for which treatment assignment deviates from being as good as random after conditioning on observed confounders. Closely related to the original result of Cornfield et al. [1959], Ding and VanderWeele [2016] derive general bounds for the causal risk-ratio, with sensitivity parameters also expressed in terms of risk-ratios relating the confounders with the treatment and the outcome.

Other approaches for sensitivity analysis directly specify a "tilting" or "bias" function that relates the conditional distribution of the outcome under treatment (or control) between treated and control units, or posits the magnitude of the difference of mean (or other contrasts of) potential outcomes. Earlier work in this area dates back to Robins [1999], Robins et al. [2000], Brumback et al. [2004], and Díaz and van der Laan [2013], with recent work by Franks et al. [2020] and Nabi et al. [2024]. In another vein, Cinelli and Hazlett [2020, 2025] derive omitted variable bias formulas for linear regression coefficients that depend on partial R^2 values describing how much residual variation unobserved confounders explain of the treatment and of the outcome. This omitted variable bias approach has been generalized to fully nonparametric settings in Chernozhukov et al. [2022a], along with doubly robust estimators, for a broad class of common causal parameters, including average treatment effects for binary treatments and average causal derivatives for continuous treatments.

3.6.3 Challenges. The sensitivity analysis literature presents somewhat of a conundrum. On the one hand, there is the canonical Cornfield et al. [1959] study, which has been extremely successful. There is also a clear sense that decision makers using causal analyses value information about the robustness of these estimates that goes beyond point estimates and standard errors. On the other hand—and despite the impressive amount of methodological work and the general consensus that causal analyses should be accompanied by some analyses to assess the

sensitivity to the fundamental assumptions that support them—such methods have not yet been widely adopted in the empirical literature. We now discuss some practical and theoretical open challenges of sensitivity analysis.

Robustness metrics. When traditional assumptions are met, there seems to be a rough agreement as to which statistics should be reported in a causal inference study: (i) a point estimate, reflecting our "best" guess of the target of inference; and (ii) a standard error, or confidence/credible interval, reflecting a measure of *statistical uncertainty* surrounding the point estimate. However, these statistics do not necessarily reveal how sensitive conclusions are to deviations from the very assumptions that justified them in the first place. An important line of research is to devise statistics that can be routinely reported to quickly summarize how sensitive (or robust) a result is to *systematic* biases. Examples include the E-value of Ding and VanderWeele [2016] and the robustness value of Cinelli and Hazlett [2020]. A closely related idea is the design of observational studies with such robustness in mind [Rosenbaum, 2004].

Connections to domain knowledge. One of the most challenging aspects of a sensitivity analysis is assessing the plausibility of violations that have been shown to be problematic. This usually requires making direct plausibility judgments on the magnitude of biases. Researchers, however, may leverage other types of domain knowledge to bound the strength of confounding. For example, Imbens [2003], Oster [2019], and Cinelli and Hazlett [2020], among others, propose comparing the relative strength of unobserved variables with the strength of observed variables. An example is provided in Figure 8. It remains an open challenge how to meaningfully exploit such types of domain knowledge for sensitivity analysis.

Other types of biases. While much of the sensitivity analysis literature has focused on biases due to unmeasured confounding, other types of biases may also pose substantial threats to valid causal inferences. For examples, the problems of sample selection, attrition, missing data, measurement error, or interference plague not only observational studies, but also randomized experiments. Performing sensitivity analysis to these types of biases, and perhaps handling all of them simultaneously, is an important open challenge.

Sharpness results. We say that a bound generated by a sensitivity analysis is sharp when it is the tightest possible bound given the sensitivity model and observed data. Obtaining sharpness results can often be difficult. For example, only recently have we obtained sharpness results for the sensitivity models of Tan [2006] and Ding and VanderWeele [2016]; see Dorn and Guo [2023] and Sjölander [2024].
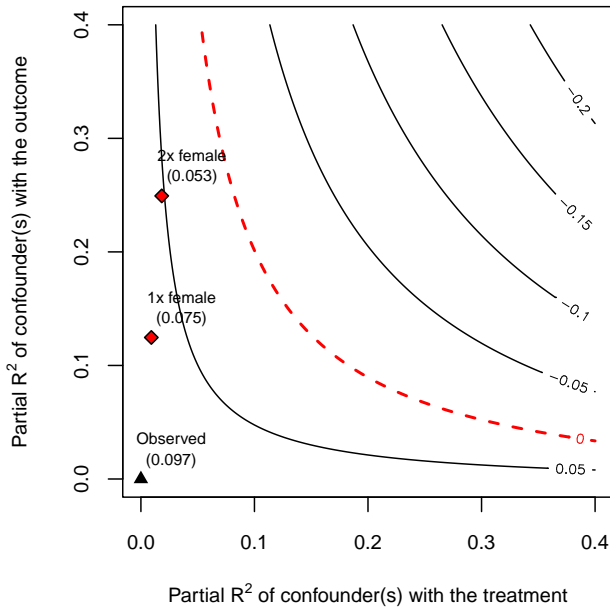
Fig 8: Example of a sensitivity contour plot with benchmarks [Cinelli et al., 2024]. The horizontal axis describes the strength of association of unobserved confounders with the treatment, while the vertical axis describes their association with the outcome. The black triangle shows the original treatment effect estimate, assuming no confounding. Contour lines show a bias-adjusted estimate, assuming confounding with the posited strength, and a direction of the bias that brings the estimate towards zero. The red diamonds show bounds on the strength of unobserved confounders if they were as strong as the observed covariate "female." In this example, the treatment effect estimate is still robust to unobserved confounding once or twice as strong as "female."

Change in culture. Current publishing practices may discourage researchers from questioning their own results. Thus, beyond technical challenges, advancing the adoption of sensitivity analysis requires a cultural shift on how researchers engage with imperfect assumptions and data. Rather than a threat to publication, sensitivity analysis should be seen as a tool for strengthening research, by drawing more robust and credible inferences under imperfect settings.

3.7 Reliable and Scalable Causal Discovery

3.7.1 Motivation. In most of the current causal inference literature we assume that we know at least qualitatively the causal relations between the treatment, the outcome and the covariates. However, in many real-world settings we do not know how the variables of interest relate to each other. For example, we might not even know which of the variables is a "treatment" and which an "outcome." Additionally, it might be impractical or unethical to intervene on any individual variable. This is common

in many disciplines, such as the biological, climate, and nutrition sciences. This challenge pertains to *causal discovery* [Spirtes et al., 2001, Glymour et al., 2019], *i.e.,* learning causal relations between the variables of interest from data.

For example, in system biology scientists are interested in discovering how cells communicate with each other and how the signals they exchange get decoded in each cell through *protein signalling networks*, networks of proteins that interact with each other [Sachs et al., 2005]. In this setting, we generally do not know which protein affects another, and our goal is to learn the causal relations between these variables from data, in order to be able to predict and control how certain signals affect a cell. This is particularly important in healthcare applications, *e.g.,* in cancer research, where cancerous cells do not respond to signals that would usually stop their growth.

Learning causal relations from data, especially observational data, is an extremely challenging task: correlation and causation are not the same. So, how can we actually learn such relationships? Intuitively, if we can assume that any causal relations between variables will make the affected variables associated each other, we can leverage patterns of association in the variables of interests to solve the inverse problem and narrow down the possible causal structures between them.

3.7.2 Background. Reichenbach's principle of common causes [Reichenbach, 1956] provides a first example of an assumption that relates dependence and causation. According to this principle, if two variables X and Y are dependent, then either X causes Y, Y causes X or there exists another variable Z that causes both X and Y. In this principle, saying X "causes" Y does not necessarily mean a direct causal effect, but possibly also an indirect causal effect. We can often additionally assume *causal faithfulness* [Spirtes et al., 2001], which in this case means that if two variables X and Y are independent, then none of these cases can happen.

How can this principle help us identify a causal structure from observational data? We show a simple example that can be identified using this principle. Let us assume that we have three variables X, Y and Z, for which we know that X and Y are independent, while the pairs X and X and Y and Z are dependent. Following Reichenbach's principle and the causal faithfulness assumption, and assuming that there are no other variables involved, we can leverage these dependences and independences to identify that the only possible causal structure between X, Y and Z is the rightmost structure in Fig. 9: because X and Y are independent, then X and Y do not cause each other, not even indirectly, nor are they both caused by another variable (in this case Z).

In this simple example, there is only one causal structure that fits the dependences and independences in the
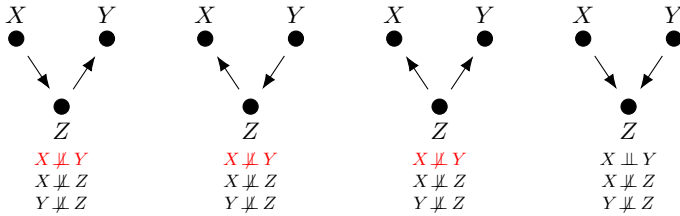
Fig 9: V-structure example: only the rightmost causal structure satisfies all the dependences and independences between the variables X, Y, Z (X and Y independent, X and Z dependent, and Y and Z dependent). The conditional independence statements in red are not consistent with the actual distribution described in the text.

distribution given our assumptions. This is not the case in general, however. There are typically multiple causal structures that fit a distribution; these structures constitute a *Markov Equivalence Class (MEC)*. If an edge between a variable X and Y are present in all of the structures in the MEC, we can say that the edge $X \rightarrow Y$ is identifiable, and X has a direct causal effect on Y.

Causal discovery [Glymour et al., 2019] provides a rich literature that allows us to learn causal relations in many complex settings, including under latent confounding. Extending Reichenbach's principle, *constraint-based causal discovery methods*, e.g. PC/FCI [Spirtes et al., 2001], leverage the assumption that conditional independence statements in the data correspond to certain patterns in the graphical structure that by the d-separation criterion [Pearl, 2000]. Through this correspondence, constraint-based methods identify the MEC by performing conditional independence tests on the data and using them as constraints on potential structures. Other methods instead focus on maximizing scores based on penalized likelihoods, e.g. [Chickering, 2002, Ramsey et al., 2017], or specific parametric assumptions [Hoyer et al., 2008, Shimizu et al., 2006].

This is just a glimpse of causal discovery—it is an exciting and rapidly evolving field, and there are many interesting challenges that we will not have space to touch upon. A particularly relevant topic is causal discovery on time series, which provides more background knowledge information, *e.g.,* the temporal order of the causal variables can preclude certain types of graph. This also comes with additional challenges, *e.g.,* subsampling, time-varying confounding, and other types of nonstationarity. We refer to Malinsky and Danks [2018], Glymour et al. [2019], Vowels et al. [2022], Zanga et al. [2022] for overviews of causal discovery.

3.7.3 Challenges Despite substantial recent progress in causal discovery in the last 30 years, many open challenges remain. We focus on three here.

Reliability in Complex Settings. Real-world scenarios often involve complex causal systems with many variables, some of which might not be observed (*i.e.,* latent confounding), as well as selection bias, measurement noise or feedback loops. Constraint-based methods provide some partial solutions to these issues as they can naturally model the presence of selection bias, latent confounding, or cyclic causal relations [Mooij and Claassen, 2020]. These methods can also integrate background knowledge, combine multiple datasets from different contexts [Mooij et al., 2020], and combine data with different sets of overlapping variables [Triantafillou and Tsamardinos, 2015b]. Additionally, unlike other causal discovery methods, constraint-based methods shift any parametric assumptions to the specific conditional independence tests used, which can then also be nonparametric [Gretton et al., 2007].

At the same time, conditional independence testing—which is at the heart of constraint-based methods—is a notoriously difficult statistical problem, even for one test at a time [Shah and Peters, 2020, Neykov et al., 2021]. Constraint-based methods compound these issues by using multiple tests on the same data, for which error dependence is difficult to model even for simple structures [Cornia and Mooij, 2014]. Moreover, measurement error can introduce spurious dependence that these methods cannot detect; there is only limited work in these settings [Zhang et al., 2018].

An open challenge is which assumptions or models we can use to improve the reliability of the estimation of causal relations from data, *e.g.,* by improving the reliability of conditional independence tests, even in these complex settings. Another direction is improving the integration of potentially unreliable background knowledge from experts or AI systems such as Large Language Models; see Section 3.12.

Scalability and Targeted Causal Discovery. Scalability is a major challenge for all causal discovery methods. Most algorithms are NP-complete and their computational complexity scales exponentially with the number of variables. In many cases, however, we are only interested in the causal effects between a few *target* variables and do not need to estimate all causal relations in a system. We then only need to identify causal relations between the target variables and their (optimal) adjustment sets, which we can then use to estimate the causal effects.

Local causal discovery methods [*e.g.,* Wang et al., 2014, Gupta et al., 2023] start to address this issue for a pair of target variables—a treatment and an outcome—using the parents of the treatment as an adjustment set. Other methods extend the types of adjustment sets, *e.g.,* by identifying groups of nodes that have certain properties in relation to this pair of targets [Maasch et al., 2024]. Watson and Silva [2022] propose an algorithm

to discover causal relations between multiple target variables under the restriction that the other variables are all non-descendants of the targets. Recently, Schubert et al. [2025] remove this restriction by proposing an iterative version of the classic PC algorithm that sequentially identifies the non-ancestors of the targets by increasing the order of the conditional independence tests and removes them from consideration. This allows us to learn a subset of the graph that is relevant for statistically efficient adjustment sets, including "optimal" adjustment sets, *i.e.,* sets that are optimal in terms of asymptotic variance [Henckel et al., 2022]. On the other hand, this iterative approach still reconstructs a MEC on the whole set of ancestors of the targets, which might be more than necessary to identify only the optimal adjustment sets, so an open challenge is then how to learn only the minimal part of the causal graph that is relevant to estimate the causal effects among a set of target variables with optimal adjustment sets in a computationally efficient way.

Post-selection inference in causal discovery. Another challenge in the integration of causal discovery and estimation is related to data "double dipping" [Gradu et al., 2022], *i.e.,* reusing the same data both for discovering the causal graph and for then estimating the causal effects, which invalidates the coverage guarantees of confidence intervals. Recently, Chang et al. [2024b] proposed a first approach to solve this issue by resampling test statistics, which can also provide guarantees in case of incorrectly learned causal graphs. While exciting, this work mostly focuses on constraint-based methods and on settings without latent confounding. An open challenge is to extend this framework to more general settings and different causal discovery approaches.

3.8 Aggregation and Synthesis of Causal Knowledge

3.8.1 Motivation A large body of causal inference research in the past few decades has focused on the design and analysis of a single study in isolation, often limited to binary treatments. However, a comprehensive understanding of causal relations ultimately relies on the integration, reconciliation, and synthesis of multiple study designs and data sources with multi-valued and multi-dimensional treatments. A multi-faceted strategy that integrates multiple data sources and complementary study designs is therefore often necessary for causal inference in the real world.

3.8.2 Background Consider, for example, a randomized experiment and an observational study. Randomized experiments control biases by design and therefore yield effect estimates with high internal validity. But such studies are typically conducted on narrow samples of available study participants, which can limit the generalizability or external validity of their findings. Conversely, observational studies tend to encompass broader, potentially representative samples and accumulate larger datasets with richer covariate and temporal resolutions, yet generally possess lower internal validity because the treatment assignment is not controlled by the investigator.

The causal inference literature has underscored the importance of integrating randomized and observational studies in enhancing the generalizability of experiments, improving their precision for targeted parameters, and reducing biases in observational studies [*e.g.,* Brantner et al., 2023, Colnet et al., 2024, Dahabreh et al., 2024]. Doing so rests on strong identification assumptions, typically ignorability and overlap conditions on the treatment assignment and sample selection mechanisms; see Hotz et al. [2005], Bareinboim and Pearl [2016] for a discussion of identification for transporting and combining effects across different populations and settings. Under these assumptions, estimation then proceeds via standard methods like treatment/selection weighting, outcome regression, and combinations thereof [see, *e.g.,* Degtiar et al., 2023]. Finally, this literature is distinct but related to research on inference from non-probability samples in sample surveys [*e.g.,* Elliott and Valliant, 2017], integration of historical controls in clinical trials [*e.g.,* Marion and Althouse, 2023], and prediction under covariate shift in machine learning [*e.g.,* Tibshirani et al., 2019].

3.8.3 Challenges
Estimating long-term effects. A common problem faced by investigators is estimating the effect of a treatment on a long-term outcome, when the investigators only observe one or more short-term outcomes that are related to the primary outcome of interest. Such short-term outcomes are often called "surrogates" for the long-term outcome [Prentice, 1989]. Athey et al. [2019] provide a framework for analyzing these issues, focusing on combining (1) an experimental study with the surrogate outcome only; and (2) a non-randomized study or descriptive data data set containing both the surrogate and primary outcome. This is a relevant and recurring setting. For instance, in cancer research, the ultimate outcome of interest is often survival, but there are surrogates that can be measured earlier, such as tumor size or other indicators of disease progression. Similarly, in economics, the objective may be to measure long-term employment or income, but these are only available in administrative datasets. In this setting, several challenges arise. For example, Athey et al. [2019] ask how to collectivize and systematize this scientific endeavor. An open challenge is building a public library of surrogate indices for long-term outcomes that could expedite the analysis of future interventions.

Constructing evidence factors. Drawing on Wittgenstein [1958, #265] regarding the critical distinction between new evidence and the same evidence repeated twice, Rosenbaum [2010, 2021] introduces a theory of evidence assembly in observational studies. This theory

aims to avoid the replication of biases that may have influenced earlier studies by isolating these biases, even though this process may expose them to new biases. The aim is to ensure that the findings mutually reinforce each other.

Rosenbaum does this within the context of a single dataset, from which different study designs using various identification strategies are leveraged. Each of these studies produces a largely independent test for the null hypothesis of no treatment effect, which is affected by different forms of biases. For instance, one can simultaneously exploit the assumption of unconfoundedness, an instrumental variable, and a dose-response relationship. This line of research offers promising opportunities, including the integration of alternative inferential approaches such as super-population sampling, the use of covariate adjustment techniques like weighting, and analytical tools from semiparametric efficiency theory.

Using historical controls in clinical trials. Another promising area that relates to the integration of diverse data sources is the use of historical controls. Randomization is the ideal method for evaluating treatments; however, in situations where it is unethical or impractical – such as when an effective treatment already exists, the disease is rare, or the study population is particularly vulnerable – investigators may have to rely on data from previous randomized experiments. These instances involve the use of historical controls, where previously collected data are used to construct comparison groups for evaluating new treatments.

This approach is gaining significant traction in pharmaceutical settings and raises several important methodological questions -— many of which intersect with broader issues in data fusion. For instance, how should we address covariate mismatch and non-compliance? How can we assess the validity of time-homogeneity assumptions that are typically in place? And how can we conduct efficient and unbiased estimation? It is important to emphasize that the use of historical controls constitutes a non-experimental approach; therefore, results derived from such analyses should not be regarded as equivalent to those from randomized studies.

Strengthening Causal Meta-Analyses. Another major challenge arises when we have aggregated information from various studies and we need to synthesize it. Meta-analysis has a long history in statistical research [*e.g.,* Hedges and Olkin, 1985, 2014], but it has developed largely independently from the field of causal inference, even though it is fundamentally concerned with causal questions. Notable exceptions include recent works by Dahabreh et al. [2020] emphasizing causal identification and semi-parametric estimation.

Nevertheless, many challenges remain, such as synthesizing information from multiple experimental and non-experimental studies, each valid under different identification assumptions. How can we combine this evidence when some studies provide disaggregated data while others offer aggregated estimates? And how can we use this information to estimate action-relevant parameters, for instance, for a specific target population of patients?

Analyzing complex systems. Traditionally, causal inference studies have prioritized learning a single, well-defined causal estimand. While this approach has produced valuable insights, a pressing challenge is how to causally describe an entire system, incorporating multiple dependencies and outcomes. Consider, for example, a model of the whole economy such as the Phillips Moniac (Monetary National Income Analogue Computer) [Bissell, 2007] or a mechanistic model of disease progression. These models encompass interacting variables and outcomes.

Key questions in these domains concern the use of multiple data sources and quasi-experimental estimates to learn such models. The challenge lies in reconciling often disparate data sources and study designs and synthesizing this information to accurately describe the entire system. Of relevance here is VanderWeele's [2017] emphasis on "outcome-wide" epidemiologic studies rather than "exposure-wide" approaches, acknowledging that public health recommendations should be informed by a holistic understanding of how various factors collectively influence health outcomes. By adopting such comprehensive approaches, we can better inform policy, ensuring that interventions are grounded in a robust understanding of the multitude of causal factors at play.

Generalizing effect estimates when (some) causal relations are unknown. While we often assume that we know the causal relationships between the covariates, the treatment, and the outcome, in some settings they need to be learned from the data at hand. This task is called *causal discovery*, as we described in Section 3.7. In the data fusion setting, all the challenges related to combining causal discovery and causal effect estimation in a statistically efficient way mentioned in Section 3.7 are compounded by additional challenges like multiple diverse populations with different distributions, or by datasets with covariate mismatch, *i.e.,* different variables measured in different settings.

Previous work already provides some partial answers to this challenge. Causal data fusion [Bareinboim and Pearl, 2016] is a principled framework to combine and transport causal effects across different populations. However, this approach requires a known causal graph and the knowledge of how distributions differ across the different populations. Although frameworks for causal discovery across multiple contexts [*e.g.,* Mooij et al., 2020], could be adapted to learn this knowledge from data, how to do this effectively is still an open question. Multiple methods have considered causal discovery in the challenging case of covariate mismatch or overlapping variables across datasets [Danks et al., 2008, Tillman and

Spirtes, 2011, Tillman and Eberhardt, 2014, Triantafillou and Tsamardinos, 2015a, Huang et al., 2020]. Another open question is how to combine causal discovery and causal effect estimation across populations. Recent work focused on observational data and experimental data [Triantafillou et al., 2023], but how to do this effectively across different populations and with covariate mismatch remains an open question.

3.9 Automation of the Causal Inference Pipeline

3.9.1 Motivation. Manual derivations have long been the primary approach for obtaining identification results for causal effects, as well as for devising efficient estimators and valid confidence intervals for the corresponding target estimands. However, this approach is labor-intensive, often slow, and error-prone, as each new set of assumptions and target query requires a new derivation from scratch. Furthermore, this process ultimately constrains empirical analysts to mold their investigations around well-established pre-existing results—both in terms of assumptions and research questions—which may not always align with their domain knowledge.

The goal of automation in causal inference is to provide scientists with methods that can flexibly adapt to their research needs. Automatic methods should handle various research questions, no matter how unconventional; leverage various types of domain knowledge, however imperfect; and, automatically handle model diagnostics as well as sensitivity analyses to key assumptions of the model. Solving these tasks requires developing *general purpose* tools that, given a set of causal estimands, assumptions and data, outputs (partial) identification results, along with finite sample estimates, confidence intervals, and statistical tests for any testable implications the model may entail.

3.9.2 Background. Automating causal inference tasks draws on various methods from symbolic and numeric computation, including techniques related to graph theory, matrix algebra, computer algebra, automated theorem proving, and optimization. Here we highlight a selected subset of these methods and applications.

One of the earliest automatic solutions to the problem of nonparametric identification originated in computer science with the development of the do-calculus and the ID algorithm, within the graphical models literature [Pearl, 1994, Tian, 2002]. The ID algorithm provides a *sound*, *complete* and *computationally efficient* solution for the problem of identifying interventional distributions from observational data, given assumptions encoded in a causal diagram [Huang and Valtorta, 2006, Shpitser and Pearl, 2006].[14] This method belongs to the class of *symbolic* approaches to identification, as it not only determines whether a causal effect is point-identified but also

produces the corresponding identification formula. Since then, several algorithms have been developed, handling a rich variety of complex problems, such as selection bias, missing data, surrogate experiments, and generalizing results across experimental settings and domains (see, e.g. Bareinboim and Pearl, 2016, Mohan and Pearl, 2021).

Other symbolic approaches to automate identification problems leverage tools from computer algebra and automated theorem proving. For example, it can be shown that identification in linear structural equation models (SEMs) with Gaussian errors reduces to solving a system of polynomial equations. A sound and complete solution to this problem is available using Gröbner bases and Buchberger's algorithm [Garcia-Puente et al., 2010]. Similarly, in nonparametric models with discrete variables, the problems of deriving testable implications as well as (partial) identification of causal parameters can be solved using quantifier elimination algorithms [Geiger and Meek, 2013]. As another example, logic-based approaches based on SAT (Boolean Satisfiability) have been used to tackle a variety of causal inference problems, such as causal discovery, including in complex settings with latent confounding, cycles and selection bias, e.g. [Hyttinen et al., 2014], and causal effect estimation with unknown graphs [Hyttinen et al., 2015]. However, in all these cases, the resulting algorithms are generally computationally inefficient.

Beyond symbolic methods, numeric methods have also been used to solve identification problems in causal inference. One such approach is to recast the problem of (partial) identification as a *constrained optimization problem*. For instance, in causal models with discrete observed variables, most queries, data, and assumptions can be expressed as polynomials of the probability mass function of discrete latent variables encoding all possible response types. The problem of (partial) identification is then equivalent to solving a *polynomial program*. While solving polynomial programs is in general NP hard, methods exist that yield anytime valid bounds [Duarte et al., 2023].[15] In contrast with the two previous examples, here the final identification result is *numeric* rather than symbolic. In certain special cases, however, the problem can be simplified to a *linear program*, for which efficient symbolic and numeric algorithms are readily available [Balke and Pearl, 1997, Sachs et al., 2023].

Once a parameter is (partially) identified, or testable implications are found, the next step in the causal inference pipeline is to draw inferences from finite data.

---

[14]Soundness ensures that when the algorithm provides an answer, it is correct, while completeness guarantees that if the algorithm fails to do so, the query is provably not identifiable without additional assumptions or data. The ID algorithm runs in polynomial time with the number of variables.

[15]That is, if the algorithm runs long enough, it converges to sharp bounds. If it is stopped at any time, the bounds remain valid, but are conservative.

The automation of this task requires the development of general-purpose tools that can accept as input potentially arbitrary identifying estimands—be it in symbolic form, or even defined implicitly, such as the solution to an optimization problem—and outputs (efficient) estimates as well as valid uncertainty quantification for those estimates. The automation of estimation and inference has a long history in statistics; well known and widely used examples include M-estimation (coupled with numerical differentiation, automatic differentiation or the bootstrap) and probabilistic programming [Efron and Tibshirani, 1994, Newey and McFadden, 1994, Carpenter et al., 2017]. In some cases, such tools are already capable of handling estimands obtained from identification algorithms. In many other cases, however, they either need to be adapted, or novel methods need to be developed.

3.9.3 Challenges Despite significant advances in the automation of identification and estimation tasks, several challenges remain. We now outline some of the most pressing issues.

Equality, inequality, and shape constraints. Algorithmic approaches to identification have focused on leveraging exclusion restrictions and ignorability assumptions (*e.g.,* absence of arrows in a graph). However, other types of constraints often arise in causal inference. For instance, many popular identification strategies make use of shape constraints such as monotonicity (*e.g.,* no "defiers"), equality (*e.g.,* "parallel trends"), or continuity (*e.g.,* continuity of potential outcomes at a cut-off point). Additionally, sensitivity analyses impose (in)equality constraints on certain parameters. Preliminary progress has been made in incorporating monotonicity and equality constraints into graph-based approaches. See VanderWeele and Robins [2010], Cinelli et al. [2019], Kumor et al. [2020], Zhang et al. [2021], Maiti et al. [2025] for examples with applications in (partial) identification and sensitivity analysis. In models with discrete variables, methods such as those discussed in the previous section can in theory handle such constraints, though in practice this has not yet been systematically studied, and issues of computational efficiency may arise. How to systematically and efficiently leverage shape, equality, and inequality constraints remains an important topic of research, with many important applications.

Scalability. Scalability is another significant challenge for many automated procedures. For example, while methods from computer algebra, such as Gröbner bases, offer a complete solution for identification in linear systems, their complexity is doubly exponential in the number of variables.[16] Similar efficiency issues arise in current complete solutions for identification in discrete systems. In many cases, it is possible to substantially speed up the performance of these algorithms by pre-simplifying equations or constraints. This is a pressing challenge to make such approaches more popular in practice. Developing algorithms that, while not complete, are efficient, also remains an important and active area of research for cases where a complete and efficient solution remains elusive or is outright known to be impossible.

Hardness results. Understanding the hardness of a task is important because it directly informs the computational resources required to solve it and guides algorithmic development. For example, if a task is known to be NP-hard, we may resort to approximation methods, heuristic approaches, or search for subsets of the problem for which an efficient algorithm may exist. Several identification problems still have unknown hardness. For instance, when the causal graph is known, the identification of interventional queries can be solved in polynomial time in nonparametric models; however, at the time of writing, whether such an algorithm exists for linear models remains an open question.

Knowledge representation. Automating identification tasks requires a precise description of the query, the data and the assumptions the researcher is willing to defend. How to systematically and efficiently represent the research question, the type of data collected, and various types of domain knowledge remains a significant challenge. For example, causal graphical models provide a natural and intuitive way of systematically representing assumptions of absence of direct effects (exclusion restrictions) or absence of confounding (ignorability restrictions); they have also been extended to encode missing data, measurement error, context specific independencies, or discrepancies across domains. On the other hand, many other types of domain knowledge remain under-explored, such as notions of variable importance, or shape constraints. Ideally, the profession should aim for a standard representation, in a machine-readable file, of estimands, types of data, and assumptions, encoding all the meta-information needed to replicate the formal aspects of a study.

Knowledge elicitation. Even when the problem of representation is solved, knowledge elicitation remains a challenge. For example, most identification algorithms assume that researchers start from a fixed, known, set of assumptions. In reality, this is rarely the case. Researchers may possess various types of domain knowledge, some of which may aid in identification while much of it may be completely inconsequential. Additionally, distinguishing between useful and irrelevant knowledge can be difficult. Thus, knowledge elicitation can be viewed as a search problem. Automatic systems should assist empirical analysts in this search by offering methods that interactively help them elicit relevant knowledge reliably and

---

[16]As an illustration, it can take hours or days to solve a five variable linear model using Gröbner bases. Note, however, that a human may also take hours, days, if not months or years to solve the same model—if ever.

efficiently. Having a set of sufficient conditions for identification may also help with this process, as it can provide a useful starting point for analysis and data collection, which can then be refined (see Sections 3.6 and 3.11).

Automatic inference with machine learning. The use of machine learning algorithms for causal effect estimation has gained significant attention. Constructing efficient estimators in these scenarios may require calculating a functional derivative known as the efficient influence function. Researchers have sought to automate these calculations through numerical [Frangakis et al., 2015, Luedtke et al., 2015, Carone et al., 2019, Jordan et al., 2022] and automatic [Luedtke, 2024] differentiation. Even when the analytical form of the efficient influence function is known, these methods depend on accurately estimating underlying nuisance parameters. To address this issue, loss functions have been developed to automatically estimate these nuisance parameters for regression functionals [Chernozhukov et al., 2022b,c, van der Laan et al., 2025]. Several open challenges remain in this area, including developing tailored loss functions that target general statistical functionals, automatically implementing estimators based on higher-order influence functions, or broadening the set of primitives to which automatic differentiation can be applied.

Software. Open-source software packages implementing many identification algorithms and estimation procedures are a fairly recent development. See, for example (non-exhaustive, and in alphabetical order): Ananke, autobounds, DoubleML, causaleffect, causaloptim, Causal Fusion, do-search, dagitty, EconML pcalg, tmle, among others [Textor et al., 2011, Gruber and Van Der Laan, 2012, Kalisch et al., 2012, Tikka and Karvanen, 2018, Karvanen et al., 2021, Syrgkanis et al., 2021, Bach et al., 2022, Jonzon et al., 2022, Lee et al., 2023, Duarte et al., 2023]. Notwithstanding this progress, software still remains an important bottleneck for the wide-spread adoption of algorithmic tools.

Artificial intelligence. While here we focused on the perspective of statisticians and data scientists who could benefit from using automated tools, the idea of automation of causal reasoning has its roots on the area of automated reasoning in artificial intelligence (AI). Integrating causal reasoning tools with modern AI systems, such as large language models, is an exciting direction of research that has not been fully explored, which we discuss in Section 3.12.

3.10 Benchmarks, Evaluation, and Validation

3.10.1 Motivation A key driver of the rapid progress in machine learning and AI has been a shared focus on community benchmarks, such as ImageNet [Deng et al., 2009], and the so-called common task framework [Donoho, 2023]. In causal inference, by contrast, there

have only been a handful of widely accepted community benchmarks. The few that exist, however, have had a major influence on the field—perhaps too much. We argue that developing new benchmarks and data challenges offers a promising opportunity for further advancement, with substantial demand from the causal inference community. Looking farther ahead, we also argue that the field needs to embrace more creative approaches for validating causal inference tools as applied in practice.

3.10.2 Background. The primary hurdle to developing benchmarks is Holland's fundamental problem of causal inference. Unlike in machine learning, where there is typically a "ground truth"—a correct image classification or future response variable—causality researchers can rarely have such access. This is because, with real-world data, we can observe only one of two or more counterfactual outcomes, making it difficult to precisely define a true causal effect. The field has therefore focused on several alternative approaches.

Replicating experimental benchmarks. In a hugely influential study, LaLonde [1986] sought to assess the empirical performance of regression methods then commonly used in economics. LaLonde began with a well-known experimental evaluation of a job training program, where the causal effect is estimated with high precision. He then replaced the experimental control group with individuals selected from surveys unrelated to the experiment. In the end, LaLonde found that regression using these non-experimental controls gave wildly different estimates than those from the original experiment itself, casting doubt on the suitability of these methods—and directly contributing to the rise in experimental evaluations of social policy. The primary challenge was the lack of overlap in covariate distributions in the two samples, see Figure 6, which displays histograms for the the log odds ratios, $\log(\hat{e}/(1 - \hat{e}))$. See Dehejia and Wahba [1999], Imbens and Xu [2024] for more details.

A cottage industry within causal inference has conducted LaLonde-style replication exercises in settings ranging from voter mobilization [Arceneaux et al., 2006] to online advertising [Gordon et al., 2019]. More broadly, there is a large and growing literature on *design replication* studies and *within-study comparisons* [Wong et al., 2021, Chaplin et al., 2018].

Synthetic and semi-synthetic benchmarks. A common alternative is to construct benchmarks with simulated causal effects. Kang and Schafer [2007] is a prominent, purely synthetic simulation benchmark that was developed to assess the performance of doubly robust methods. In another widely used benchmark, Hill [2011] constructed a semi-synthetic simulation, drawing the covariates, treatment, and control potential outcomes $Y_i(0)$ from the Infant Home Development Program (IHDP).

Fig 10: Assessing Overlap in Lalonde-Dehejia-Wahba (LDW) Data
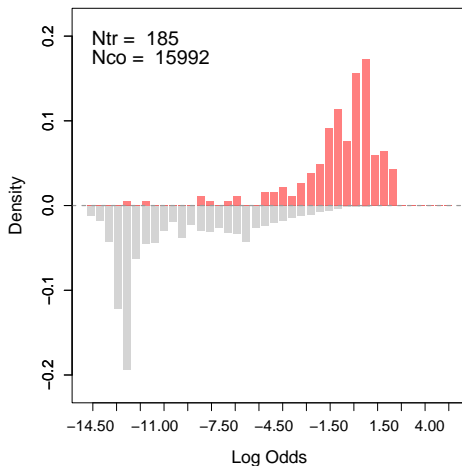
(a) LDW-CPS

Note: Histograms depict the log odds ratios, *i.e.,* $\log(\hat{e}/(1-\hat{e}))$, using propensity score estimated through generalized random forest. The data are the observational CPS sample and the experimental from the National Supported Work Demonstration sample, both available in the Lalonde-Dehejia-Wahba data in Dehejia and Wahba [1999], Imbens and Xu [2024].

More recently, the American Causal Inference Conference (ACIC) has hosted regular competitions where teams submit estimators with the goal of having the best predictions on a set of semi-synthetic datasets [Dorie et al., 2019].

Data challenges and common tasks. Finally, researchers have organized around analyzing a common dataset, seeking to aggregate insights across different methods and research teams—without a (synthetic) ground truth. This has been more prominent within causal discovery for biology, such as the DREAM 5 causal discovery challenge, with the goal of estimating protein signaling networks in cells [Hill et al., 2016].

3.10.3 Challenges. While existing benchmarks are widely used—each with thousands of citations—relying too heavily on a single one of them, such as LaLonde [1986], can distort incentives and lead the field to overadjust to specific attributes of the study design. This highlights the need for new data challenges and benchmarks to further accelerate causality research. In the short term, the causal inference community is actively working toward improved benchmarks and simulations that reflect the many and varied causal inference use cases [Curth et al., 2021]. A longer-term goal is to validate the use of causal methods more broadly. Here the question is more fundamentally about science—*e.g.,* explanation, interpretation, and understanding mechanisms or causal structure—rather than prediction. As such, this is less amenable to ML-style benchmarks but remains critically

important, especially after high-profile failures like those highlighted in the introduction. Among other considerations, it is important to evaluate the methods not only in terms of their final estimates, but also to the extent to which they approximate features of a hypothetical randomized experiment and the robustness of their estimates.

High-dimensional and complex data. Existing causal inference evaluations have typically been quite low-dimensional—most re-analyses of LaLonde [1986] only have ten covariates, though there have been some high-dimensional variants [Farrell, 2015]. Such evaluations are out of step with the rapid rise of the application of machine learning methods, such as neural networks, to causal inference. Developing benchmarks with high-dimensional, and complex, multi-modal data remains an open challenge.

Designed within-study comparisons. One promising generalization of LaLonde-style benchmarks is to explicitly incorporate within-study comparisons into a study design. For instance, Shadish et al. [2008] randomly assigned study participants to be in a randomized or nonrandomized trial, allowing for a direct assessments of non-experimental adjustment methods. Or Saveski et al. [2017] randomized whole groups into two different types of randomized designs to assess approaches for causal inference under interference.

Tailored simulations. A related issue is the gap between how causal inference methods are assessed when developed—typically in limited non-representative simulation studies and stylized data illustrations—and assessing the performance of that method in a specific application. One promising but underused strategy is to essentially design a simulation study tailored to specific applications, analogous to power calculations [Athey et al., 2021]. Conducting such simulations can be prohibitive for domain scientists and can dramatically slow practical statistical analysis. Moreover, designing simulation studies specifically for causal inference presents a unique set of challenges [Evans and Didelez, 2023]. Thus, automating this process and lowering implementation barriers for domain scientists is an important goal [Nance et al., 2024].

Developing practical tools. More broadly—and in the spirit of "frictionless reproducibility" in machine learning [Donoho, 2023]—the causal inference community has only recently developed easily-modified repositories. For instance, an important setup was the Dorie et al. [2019] R package for accessing semisynthetic data sets from the 2016 ACIC challenge; see also Lin et al. [2019]. Further developing these practical tools will be critical for broader uptake.

Ground truth validation in the biological and physical sciences. Developing frameworks for validation necessarily requires creative solutions—and continued engagement with domain scientists. One direction is from

biological applications, where researchers can sometimes pair data challenges and competitions with subsequent interventions to see which predictions were closest. In the context of gene knockout experiments, researchers might use a "training set" of interventions to predict promising interventions in a "test set"—and then confirm those predictions with actual experimentation [see Meinshausen et al., 2016]. In the context of physical systems, Weichwald et al. [2022] challenged causal researchers to deploy methods on novel data from real-world control systems; and Gamella et al. [2024] created computer-controlled laboratories that can assess and validate new causal methods. Finally, Runge et al. [2019] explored causal discovery methods in the context of climate data [see also Brouillard et al., 2024].

Improved success metrics. In many application domains, especially those in the medical and social sciences, the definition of "success" is open-ended. In addition to standard measures of statistical performance, other aspects, such as a strong study design and well-balanced covariates, can be crucial to the credibility and validity of a study; see, for example, Zhao et al. [2019a]. A close engagement with domain scientists is critical, as their specialized knowledge can guide the design and methods to address the most pertinent issues.

Professional hurdles to building better benchmarks. Even if the field agrees on the scientific goals, many professional and practical hurdles remain to building better benchmarks. Given the limited incentives for developing benchmarks, how can we secure funding to support their creation and maintenance? How can teaching efforts embed the development of benchmarks? And how can domain experts be actively involved in this effort (*e.g.,* following the example of several ACIC data challenges)?

As with any approach, there will be fundamental limits to such benchmarking and validation. Understanding these limits will be important for making these approaches actionable in practice.

3.11 New Identification Strategies

3.11.1 Motivation. As we have discussed, going from association to causation requires a deep understanding of the mechanisms that generated the data. Typically, this understanding is encoded in terms of identification assumptions, which allow us to use observed measurements within the data to identify or "see" unobservable causal quantities that go beyond the data. In an ideal scenario, these assumptions are inherently valid by design, such as in the case of a perfectly conducted randomized experiment. However, as we depart from this ideal—especially in observational studies—these assumptions must be grounded in substantive considerations about the data, often referred to as an research design or identification strategy [Angrist and Krueger, 1999].

There are now a fairly standard set of identification strategies widely used in the empirical social sciences, especially: adjusting or controlling for as many covariates as possible (see Sec 2.2); instrumental variables; regression discontinuity designs; and panel data strategies like difference-in-differences. This raises the question: besides these identification strategies, can we develop or uncover others?

3.11.2 Background.
Synthetic controls. A recent method that has gained much popularity is building synthetic controls [Abadie and Gardeazabal, 2003, Abadie et al., 2010]. This approach is often used in panel data settings to create a weighted average of control units that are balanced with respect to the pre-treatment outcomes of a focal treated unit. Since 2010, the number of applications of these methods in various forms has exploded in both academic settings and the private sector, especially in applications in which standard panel data methods are not appropriate. There has been an active methodological literature exploring and extending this framework, including for uncertainty quantification and bias correction.

Mendelian Randomization. In the context of instrumental variables, Mendelian randomization (MR) has emerged as a promising strategy for uncovering causal effects [Sanderson et al., 2022]. Based on Mendel's laws of inheritance, this strategy leverages the inherent randomness by which parents pass their genetic endowments to their offsprings. For example, MR has been used to explore the causal relationships between lifestyle factors such as alcohol consumption, diet patterns, and physical activity, and health outcomes such as cardiovascular disease. Despite major methodological improvements, some open challenges remain. Pleiotropic effects, where genetic variants influence multiple traits, can violate the exclusion restriction assumption, requiring the development of methods to detect and potentially adjust for these effects. Weak genetic instruments can produce biased estimates and reduce statistical power, requiring robust methods that strengthen the relationship between the instrument and the exposure. Furthermore, the selection and validation of genetic variants as instruments is crucial, requiring improved methods that integrate both biological and statistical perspectives. Finally, comprehensive guidelines that delineate best practices for MR estimation are key to guide researchers in conducting robust analyses.

Graphical models. Graphical models may uncover novel identification opportunities beyond popular strategies, some of which may still be underexplored. As an example of a graphically inspired identification result, Pearl [1995] proposed the front-door criterion. This strategy has a similar structure to instrumental variables, in which three variables are linked in a causal path. Here

the treatment of interest affects a mediator, the mediator affects the outcome, and, importantly, there is not direct causal path from the treatment to the outcome, other than through the mediator. The causal effect of the treatment on the outcome can then be identified through its mediating effects, even in the presence of unobserved treatment-outcome confounding, so long as the mediating paths are unconfounded. This causal model has not yet been widely explored in empirical social science, though see Bellemare et al. [2024] for a recent application. More broadly, if a novel graph presents new identification results, there is a question of how frequently such situations occur in real-world settings. It is an interesting direction where mathematics offers insights into the real world, as opposed to the conventional approach of formalizing theories from empirical observations.

3.11.3 Challenges.

Instrumental variables. There are many open challenges in extending instrumental variable approaches. Shift-share IVs, which have seen increased popularity in economics, exploit variation in the impact of external shocks across time and regions to identify causal effects [Borusyak et al., 2022]. Separately, bunching techniques detect behavioral responses to policy changes through clustering in outcome distributions [Kleven, 2016]. Both offer promising directions for further methodological research.

Panel data. Another relevant direction is panel data methods, which exploit repeated measurements of stable units across time points [Arkhangelsky and Imbens, 2024]. Historically, this literature has been largely distinct from the (mainly biostatistical) literature on time-varying treatments. Bridging these two literatures is a fruitful research direction. One recent example is for case-crossover designs, which provide a framework for examining transient exposures and acute events by comparing case (exposure) periods with control periods for the same individual [Shahn et al., 2023]. Exploring alternative identification strategies that involve forms of latent unconfoundedness offers interesting challenges for identification and inference [Athey and Imbens, 2025].

Proximal methods. Additionally, the use of negative controls is gaining traction as a method for detecting and addressing confounding bias and model misspecification by leveraging variables or populations that are not expected to have causal effects as a benchmark [Tchetgen Tchetgen, 2014]. These emerging methods collectively contribute to advancing proximal causal inference, a general framework that seeks to bridge different identification strategies [Tchetgen Tchetgen et al., 2024].

Combining strategies. As we have discussed, it is crucial to understand that the aforementioned strategies are not mutually exclusive. Hence, an important question, is how to substantively combine them for more robust causal inference.

3.12 Large Language Models and Causality

3.12.1 Motivation Large Language Models (LLMs) have burst onto the scene over the past several years, dramatically changing research across a wide range of domains in a short period. We are starting to see this impact in causal research as well. There is a growing literature at the interface of LLMs and causality, both on how to use LLMs to improve causal inference and on using ideas from causality for improving LLMs. Since this area is in flux, we offer a brief overview of some exciting topics here, recognizing that this literature is changing rapidly.

3.12.2 Background and Challenges
Incorporating complex data into the causal inference pipeline. Over the last decade, Natural Language Processing (NLP) methods have become increasingly common for leveraging text for causal inference, including as confounders, treatments, and outcomes [Veitch et al., 2020, Egami et al., 2022, Feder et al., 2022, Imai and Nakamura, 2024]. For example, researchers might be interested in the effect of reading different messages on a downstream outcome, and then try to isolate key attributes that explain these effects [Gui and Veitch, 2022, Lin et al., 2024b]. Building on these earlier ideas, the widespread availability of LLM tools has rapidly accelerated the use and scale of causal inference with text, as well as causal inference with other unstructured high-dimensional data, like images [Jerzak et al., 2023, Imai and Nakamura, 2025]. Many open questions remain on how best to incorporate LLMs into the causal inference pipeline, including developing best practices for architecture and representation learning and appropriately quantifying uncertainty. Additional issues arise when LLMs and other generative AI systems are used to generate the interventions of interest—even defining the appropriate causal questions here is challenging.

Synthetic experiments and Agentic Modelling. Artificial or synthetic experiments are one explicitly causal area that has already seen preliminary work. Suppose one is interested in the effect of a particular intervention, say exposing individuals to a social media post about vaccines, on the likelihood they would subsequently take the vaccine. One could simply ask an LLM what the effect of such a post would be. More recent work however, has addressed such questions by creating a sample of artificial agents, and running an experiment on these agents, just like a regular experiment on real humans. See, for example, Hewitt et al. [2024], De Bartolomeis et al. [2025], Chen et al. [2025], Wang et al. [2025a], Manning et al. [2024]. Such experiments seem to hold great promise in select settings given the speed with which one can conduct them and the avoidance of many (but not necessarily all) ethical issues.

At the same time, the performance of synthetic experiments in more realistic settings is an important open question; see Gui and Toubia [2023], Anthis et al. [2025]. For example, in the parallel literature on synthetic audiences—using artificial agents to approximate survey respondents—there has been growing concern about out-of-sample performance [*e.g.,* Li et al., 2025, von der Heyde et al., 2025]. Going forward, it does seem likely that LLMs will be very helpful in augmenting randomized experiments. This can be in the form of performing initial runs using artificial agents, followed by targeted experiments using real individuals, or by replacing actual experiments entirely in some special cases. One challenge is to validate the results from artificial experiments, and to assess the uncertainty surrounding their results. Another is to understand what features of such artificial experiments make them more or less reliable.

LLMs and causal reasoning. Whether LLMs have some version of causal reasoning capabilities is an active research area [Kiciman et al., 2023, Jin et al., 2023]. Examples include chain-of-thought reasoning [Wei et al., 2022] and hierarchical reasoning [Wang et al., 2025b]. These questions are closely related to causal discovery and suggest many promising directions for incorporating LLMs to better understand complex causal systems. For example, an interesting question is whether current LLMs can discover causal relations from correlations [Jin et al., 2024], or if they can be used to provide common sense knowledge *e.g.,* as experts in the loop in combination with classic causal discovery tools [Ankan and Textor, 2025] or as imperfect experts providing an initial causal ordering [Vashishtha et al., 2025]. Finally, there is also great interest in training *transformers, i.e.,* the backbone of many modern LLMs, as a foundation model to estimate causal effects [Ma et al., 2025, Balazadeh et al., 2025].

AI co-pilots for causal inference. Echoing the use of AI tools to automate data analysis (see Section 3.9), there have been several recent efforts to use AI tools as "co-pilots" in aiding applied causal inference [Alaa et al., 2024, Verma et al., 2025] and in study planning more broadly [Chang et al., 2024a]. This direction shows promise as a means for democratizing applied causal inference research especially.

Causality to improve LLMs. The alternative direction has also shown great promise: using the ideas of causality and causal inference to improve AI development and deployment. For example, causal reasoning has proved important in developing more trustworthy and reliable AI deployments [Binkyte et al., 2025] and in better understanding preference learning methods used to train LLM [Lin et al., 2024a, Kobalczyk and van der Schaar, 2025]. Causal inference is also critical for evaluating and monitoring deployed AI systems [Liu et al., 2025]. Another promising direction is the use of *causal abstraction,*

i.e., the theory describing the relations between different causal models at different granularities, as a way to interpret the internal mechanisms of the black-box LLMs, *e.g.,* [Geiger et al., 2021, Wu et al., 2023, Geiger et al., 2024].

Open questions. LLMs have burst onto the scene in a very short time, and have already had major impacts on many areas of study. The impact on causality and causal inference is still limited at the moment, but likely to be profound. Exactly where this will be is a difficult question to answer, and we are excited to see what develops.

4. CONCLUDING THOUGHTS

For centuries, the quest to establish causation has been central to scientists in their efforts to uncover the laws that govern the natural world. In today's society, a similar aspiration drives policy makers: to find which interventions work to build a better social world. We seek to understand causation, uncover natural laws, and learn which interventions work through data from the real world—a process known as causal inference. This is a perennial quest that will never end.

4.1 Many additional directions

In this paper, we explore this issue from the perspective of statistics and related fields, yet there is much to gain integrating diverse perspectives and fostering cross-disciplinary work. While we provide an overview and discuss some of the open challenges in causal inference from this perspective, this is merely a partial map of a much broader landscape. There are many others that we have not addressed.

Time and dynamics. One such challenge is incorporating more complex temporal dimensions. While we covered various principles and problems that generally apply to causal studies with increased temporal measurements, this dimension adds a specific complexity. Despite the 40 years of progress since Robins's [1986] foundational paper, integrating time-dependence—encompassing treatments, covariates, outcomes, and their feedback—into routine causal analyses remains a persistent challenge.

Bayesian causal inference. Another challenge involves using Bayesian methods for causal inference. Although Bayesian approaches have been proposed and adopted, especially in the literature on heterogeneous treatment effects and causal machine learning [*e.g.,* Hill, 2011, Hahn et al., 2020], the frequentist paradigm has traditionally dominated the field. Useful perspectives on Bayesian methods for causal inference are provided in, among others, Li et al. [2023] and Daniels et al. [2023].

Designing observational studies and target trial emulation. Furthermore, a core idea in causal inference is to cast causal questions in terms of concrete interventions and conceptualize them in the context of hypothetical randomized experiments. This notion has a long history, dating back at least to the work of Dorn [1953] and Cochran

[1965]. In this tradition, one widely adopted framework in the health sciences is that of target trial emulation [Hernán and Robins, 2016]. A key challenge is how to effectively apply this approach in other scientific domains and policy contexts.

Causes of effects. Finally, while we have focused on the question of the effect of causes, the reverse question of the causes of effects is also important [Dawid and Musio, 2022]. This type of questions emerges often in legal contexts, for instance, in determining whether a certain drug caused a health problem.

All these challenges and possible research directions intersect and compound each other, opening up new opportunities to advance causal inference.

4.2 Cross-cutting issues

Our main discussion raises several cross-cutting issues that we want to emphasize again here.

Bridging the gap between theory and practice. Every challenge we discuss above faces a familiar disconnect between applications and theory. There is an urgent need to make state-of-the-art methods accessible and applicable to address substantive research problems in practice and to facilitate their adoption. This requires deep engagement with substantive experts, the development of better software and practical diagnostics, and better guidance and outreach. In the other direction, such engagement will hopefully lead to new and exciting methodological challenges, following a long tradition of advances in causal inference.

Gains from incorporating machine learning and computational tools. Over the last two decades, the widespread availability of improved computation and machine learning methods has led to fundamental changes in the practice of causal inference and launched the increasingly influential subfield of "causal machine learning." Today, powerful LLM and AI tools offer similar promise, and we anticipate exciting uses for these new technologies in causality research.

Importance of building an open and inclusive research community. As we outline above, causality research has historically been highly fractured, with often distinct and isolated research communities. Over the last decade especially, there has been substantial progress in creating a "big tent" causal community. In addition to making the field more welcoming and inclusive, this has also led to important research advances and more rapid adoption of novel ideas. Furthering such efforts is critical for the continued growth and evolution of causality research.

4.3 Asking good questions

In the end, the central challenge in causality research is to continue to ask good questions. We have posed many questions for the causal community here. But as the field continues to grow, what questions are we missing or failing to ask?

ACKNOWLEDGMENTS

FUNDING

REFERENCES

O. O. Aalen, M. J. Stensrud, V. Didelez, R. Daniel, K. Røysland, and S. Strohmaier. Time-dependent mediators in survival analysis: modeling direct and indirect effects with the additive hazards model. *Biometrical Journal*, 62(3):532–549, 2020.

A. Abadie and J. Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93(-): 113–132, 2003.

A. Abadie, A. Diamond, and J. Hainmueller. 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, 2010.

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International conference on machine learning*, pages 1638–1646. PMLR, 2014.

A. Alaa, R. V. Phillips, E. Kıcıman, L. B. Balzer, M. van der Laan, and M. Petersen. Large language models as co-pilots for causal inference in medical studies. *arXiv preprint arXiv:2407.19118*, 2024.

J. G. Altonji, T. E. Elder, and C. R. Taber. Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of political economy*, 113(1):151–184, 2005.

J. D. Angrist and A. B. Krueger. Empirical strategies in labor economics. In *Handbook of Labor Economics*, volume 3, pages 1277–1366. Elsevier, 1999.

J. D. Angrist and J.-S. Pischke. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2):3–30, 2010.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

A. Ankan and J. Textor. Expert-in-the-loop causal discovery: Iterative model refinement using expert knowledge. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

J. R. Anthis, R. Liu, S. M. Richardson, A. C. Kozlowski, B. Koch, J. Evans, E. Brynjolfsson, and M. Bernstein. Llm social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*, 2025.

K. Arceneaux, A. S. Gerber, and D. P. Green. Comparing experimental and matching methods using a large-scale voter mobilization experiment. *Political Analysis*, 14(1):37–62, 2006.

D. Arkhangelsky and G. Imbens. Causal models for longitudinal and panel data: A survey. *The Econometrics Journal*, 27(3):C1–C61, 2024.

P. M. Aronow and C. Samii. Estimating average causal effects under interference between units, 2013.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

S. Athey and G. Imbens. Identification of average treatment effects in nonparametric panel models. *arXiv preprint arXiv:2503.19873*, 2025.

S. Athey and S. Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

S. Athey, R. Chetty, G. W. Imbens, and H. Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research, 2019.

S. Athey, G. W. Imbens, J. Metzger, and E. Munro. Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics*, 2021.

P. Bach, V. Chernozhukov, M. S. Kurz, and M. Spindler. Doubleml-an object-oriented implementation of double machine learning in python. *Journal of Machine Learning Research*, 23(53):1–6, 2022.

D. H. Bailey, A. J. Jung, A. M. Beltz, M. I. Eronen, C. Gische, E. L. Hamaker, K. P. Kording, C. Lebel, M. A. Lindquist, J. Moeller, et al. Causal inference on human behaviour. *Nature Human Behaviour*, 8(8):1448–1459, 2024.

P. Bajari, B. Burdick, G. W. Imbens, L. Masoero, J. McQueen, T. S. Richardson, and I. M. Rosen. Experimental design in marketplaces. *Statistical Science*, 1(1):1–19, 2023.

S. Balakrishnan and L. Wasserman. Hypothesis testing for densities and high-dimensional multinomials. *The Annals of Statistics*, 47 (4):1893–1927, 2019.

S. Balakrishnan, E. H. Kennedy, and L. Wasserman. The fundamental limits of structure-agnostic functional estimation. *arXiv preprint arXiv:2305.04116*, 2023.

V. Balazadeh, H. Kamkari, V. Thomas, B. Li, J. Ma, J. C. Cresswell, and R. G. Krishnan. Causalpfn: Amortized causal effect estimation via in-context learning. *arXiv preprint arXiv:2506.07918*, 2025.

A. Balke and J. Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American statistical Association*, 92(439):1171–1176, 1997.

K. Bansak, J. Hainmueller, D. J. Hopkins, T. Yamamoto, J. N. Druckman, and D. P. Green. Conjoint survey experiments. *Advances in experimental political science*, 19:19–41, 2021.

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27): 7345–7352, 2016.

R. M. Baron and D. A. Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.

M. F. Bellemare, J. R. Bloem, and N. Wexler. The paper of how: Estimating treatment effects using the front-door criterion. *Oxford Bulletin of Economics and Statistics*, 86(4):951–993, 2024.

E. Ben-Michael, K. Imai, and Z. Jiang. Policy learning with asymmetric counterfactual utilities. *Journal of the American Statistical Association*, pages 1–14, 2024.

R. Bhattacharya, D. Malinsky, and I. Shpitser. Causal inference under interference and network uncertainty. In *Uncertainty in Artificial Intelligence*, pages 1028–1038. PMLR, 2020.

P. J. Bickel, C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

B. Bilodeau, L. Wang, and D. Roy. Adaptively exploiting d-separators with causal bandits. *Advances in Neural Information Processing Systems*, 35:20381–20392, 2022.

R. Binkyte, I. Sheth, Z. Jin, M. Havaei, B. Schölkopf, and M. Fritz. Causality is key to understand and balance multiple goals in trustworthy ml and foundation models. *arXiv preprint arXiv:2502.21123*, 2025.

L. Birgé and P. Massart. Estimation of integral functionals of a density. *The Annals of Statistics*, pages 11–29, 1995.

C. Bissell. Historical perspectives-the moniac a hydromechanical analog computer of the 1950s. *IEEE Control Systems Magazine*, 27(1): 69–74, 2007.

K. A. Bollen. *Structural equations with latent variables*, volume 210. John Wiley & Sons, 1989.

M. Bonvini and E. H. Kennedy. Sensitivity analysis via the proportion of unmeasured confounding. *Journal of the American Statistical Association*, 117(539):1540–1550, 2022.

K. Borusyak, P. Hull, and X. Jaravel. Quasi-experimental shift-share research designs. *Review of Economic Studies*, 89(1):181–213, 2022.

J. Bradic, V. Chernozhukov, W. K. Newey, and Y. Zhu. Minimax semiparametric learning with approximate sparsity. *arXiv preprint arXiv:1912.12213*, 2019.

C. L. Brantner, T.-H. Chang, T. Q. Nguyen, H. Hong, L. Di Stefano, and E. A. Stuart. Methods for integrating trials and non-experimental data to examine treatment effect heterogeneity. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 38(4):640, 2023.

P. Brouillard, C. Squires, J. Wahl, K. P. Kording, K. Sachs, A. Drouin, and D. Sridhar. The landscape of causal discovery data: Grounding causal discovery in real-world applications. *arXiv preprint arXiv:2412.01953*, 2024. URL https://doi.org/10.48550/arXiv.2412.01953.

B. A. Brumback, M. A. Hernán, S. J. Haneuse, and J. M. Robins. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in medicine*, 23 (5):749–767, 2004.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and non-stochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 2012.

J. G. Bullock, D. P. Green, and S. E. Ha. Yes, but what's the mechanism?(don't expect an easy answer). *Journal of Personality and Social Psychology*, 98(4):550, 2010.

M. Carone, A. R. Luedtke, and M. J. Van Der Laan. Toward computerized efficient estimation in infinite-dimensional models. *Journal of the American Statistical Association*, 2019.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76: 1–32, 2017.

S. E. Carrell, B. I. Sacerdote, and J. E. West. From natural variation to optimal policy? the importance of endogenous peer group formation. *Econometrica*, 81(3):855–882, 2013.

R. J. Carroll and P. Hall. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83 (404):1184–1186, 1988.

S. Chang, A. Kennedy, A. Leonard, and J. A. List. 12 best practices for leveraging generative ai in experimental research. Technical report, National Bureau of Economic Research, 2024a.

T.-H. Chang, Z. Guo, and D. Malinsky. Post-selection inference for causal effects after causal discovery. *arXiv preprint arXiv:2405.06763*, 2024b.

D. D. Chaplin, T. D. Cook, J. Zurovac, J. S. Coopersmith, M. M. Finucane, L. N. Vollmer, and R. E. Morris. The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management*, 37(2):403–429, 2018.

S. H. Chen, Y.-C. Chen, and J.-T. Liu. The impact of family composition on educational achievement. *Journal of Human Resources*, 54 (1):122–170, 2019.

Y. Chen, Y. Hu, and Y. Lu. Predicting field experiments with large language models. *arXiv preprint arXiv:2504.01167*, 2025.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21 (1):C1–C68, 2018.

V. Chernozhukov, C. Cinelli, W. Newey, A. Sharma, and V. Syrgkanis. Long story short: Omitted variable bias in causal machine learning. Technical report, National Bureau of Economic Research, 2022a.

V. Chernozhukov, W. Newey, V. M. Quintas-Martınez, and V. Syrgkanis. Rieisznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. In *International Conference on Machine Learning*, pages 3901–3914. PMLR, 2022b.

V. Chernozhukov, W. K. Newey, and R. Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022c.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

N. Christakis and J. Fowler. The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine*, (357):370–379, 2007.

C. Cinelli and C. Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 82(1):39–67, 2020.

C. Cinelli and C. Hazlett. An omitted variable bias framework for sensitivity analysis of instrumental variables. *Biometrika*, 112(2): asaf004, 2025.

C. Cinelli, D. Kumor, B. Chen, J. Pearl, and E. Bareinboim. Sensitivity analysis of linear structural causal models. In *International conference on machine learning*, pages 1252–1261. PMLR, 2019.

C. Cinelli, J. Ferwerda, and C. Hazlett. sensemakr: Sensitivity analysis tools for OLS in R and Stata. *Observational Studies*, 10(2):93–127, 2024.

W. Cochran. The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, 128(2):234–266, 1965.

B. Colnet, I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang. Causal inference methods for combining randomized trials and observational studies: A review. *Statistical Science*, 39(1):165–191, 2024.

T. D. Cook. "waiting for life to arrive": a history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, 142(2):636–654, 2008.

J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1):173–203, 1959.

N. Cornia and J. M. Mooij. Type-ii errors of independence tests can lead to arbitrarily large errors in estimated causal effects: an illustrative example. In *Proceedings of the UAI 2014 Conference on Causal Inference: Learning and Prediction-Volume 1274*, pages 35–42, 2014.

J. D. Correa and E. Bareinboim. Counterfactual graphical models: Constraints and inference. In *Forty-second International Conference on Machine Learning*, 2025.

D. R. Cox. *Planning of experiments*. Wiley, 1958.

S. Cunningham. *Causal inference: The mixtape*. Yale University Press, 2018.

A. Curth, D. Svensson, J. Weatherall, and M. van der Schaar. Really doing great at estimating cate? a critical look at ml benchmarking practices in treatment effect estimation. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*, 2021.

I. J. Dahabreh, L. C. Petito, S. E. Robertson, M. A. Hernán, and J. A. Steingrimsson. Toward causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a new target population. *Epidemiology*, 31(3):334–344, 2020.

I. J. Dahabreh, A. Matthews, J. A. Steingrimsson, D. O. Scharfstein, and E. A. Stuart. Using trial and observational data to assess effectiveness: trial emulation, transportability, benchmarking, and joint analysis. *Epidemiologic Reviews*, 46(1):1–16, 2024.

M. J. Daniels, A. Linero, and J. Roy. *Bayesian nonparametrics for causal inference and missing data*. Chapman and Hall/CRC, 2023.

D. Danks, C. Glymour, and R. Tillman. Integrating locally learned causal structures with overlapping variables. *Advances in Neural Information Processing Systems*, 21, 2008.

A. P. Dawid and M. Musio. Effects of causes and causes of effects. *Annual Review of Statistics and Its Application*, 9(1):261–287, 2022.

P. De Bartolomeis, J. Abad, G. Wang, K. Donhauser, R. M. Duch, F. Yang, and I. J. Dahabreh. Efficient randomized experiments using foundation models. *arXiv preprint arXiv:2502.04262*, 2025.

I. Degtiar, T. Layton, J. Wallace, and S. Rose. Conditional cross-design synthesis estimators for generalizability in medicaid. *Biometrics*, 79(4):3859–3872, 2023.

R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

I. Díaz and M. van der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549, 2012.

I. Díaz and M. J. van der Laan. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *International Journal of Biostatistics*, 9(2):149–160, 2013.

I. Díaz, N. Williams, and K. E. Rudolph. Efficient and flexible mediation analysis with time-varying mediators, treatments, and confounders. *Journal of Causal Inference*, 11(1):20220077, 2023.

V. Didelez, A. P. Dawid, and S. Geneletti. Direct and indirect effects of sequential treatments. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 138–146, 2006.

M. Dimakopoulou, Z. Zhou, S. Athey, and G. Imbens. Balanced linear contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3445–3453, 2019.

P. Ding. *A first course in causal inference*. CRC Press, 2024.

P. Ding and T. J. VanderWeele. Sensitivity analysis without assumptions. *Epidemiology*, 27(3):368–377, 2016.

R. Doll and A. B. Hill. Smoking and carcinoma of the lung. *British medical journal*, 2(4682):739, 1950.

D. Donoho. Data science at the singularity. *arXiv preprint arXiv:2310.00865*, 2023.

D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, 90(432):1200–1224, 1995.

V. Dorie, M. Harada, N. B. Carnegie, and J. Hill. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in Medicine*, 35(20):3453–3470, 2016.

V. Dorie, J. Hill, U. Shalit, M. Scott, and D. Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

H. F. Dorn. Philosophy of inferences from retrospective studies. *American Journal of Public Health and the Nations Health*, 43(6):677–683, 1953.

J. Dorn and K. Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, 118(544):2645–2657, 2023.

J. Dorn, K. Guo, and N. Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *Journal of the American Statistical Association*, pages 1–12, 2024.

G. Duarte, N. Finkelstein, D. Knox, J. Mummolo, and I. Shpitser. An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*, pages 1–16, 2023.

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104, 2011.

O. D. Duncan. *Introduction to structural equation models*. Elsevier, 2014.

S. Efromovich and M. Pinsker. Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statistica Sinica*, pages 925–942, 1996.

B. Efron. Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417, 1971.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*, volume 57. Chapman & Hall/CRC, 1994.

N. Egami, C. J. Fong, J. Grimmer, M. E. Roberts, and B. M. Stewart. How to make causal inferences using texts. *Science Advances*, 8 (42):eabg2652, 2022.

M. R. Elliott and R. Valliant. Inference for nonprobability samples. *Statistical Science*, 32(2):249, 2017.

R. J. Evans and V. Didelez. Parameterizing and simulating from causal models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkad058, 2023.

M. H. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

M. H. Farrell, T. Liang, and S. Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

A. Feder, K. A. Keith, E. Manzoor, R. Pryzant, D. Sridhar, Z. Wood-Doughty, J. Eisenstein, J. Grimmer, R. Reichart, M. E. Roberts, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022.

F. M. Fisher. Identifiability criteria in nonlinear systems. *Econometrica: Journal of the Econometric Society*, pages 574–590, 1961.

R. Fisher. Cigarettes, cancer, and statistics. *The Centennial Review of Arts & Science*, 2:151–166, 1958.

R. A. Fisher. *The design of experiments*. Oliver And Boyd; Edinburgh; London, 1937.

R. A. Fisher. Dangers of cigarette-smoking. *British Medical Journal*, 2(5039):297, 1957.

D. J. Foster and V. Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.

C. E. Frangakis, T. Qian, Z. Wu, and I. Diaz. Deductive derivation and turing-computerization of semiparametric efficient estimation. *Biometrics*, 71(4):867–874, 2015.

A. Franks, A. D'Amour, and A. Feller. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, 115(532):1730–1746, 2020.

J. L. Gamella, J. Peters, and P. Bühlmann. The causal chambers: Real physical systems as a testbed for ai methodology. *arXiv preprint arXiv:2404.11341*, 2024.

M. Gao. Endogenous interference in randomized experiments. *arXiv preprint arXiv:2412.02183*, 2024.

L. D. Garcia-Puente, S. Spielvogel, and S. Sullivant. Identifying causal effects with computer algebra. *arXiv preprint arXiv:1007.3784*, 2010.

A. Geiger, H. Lu, T. Icard, and C. Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586. Curran Associates, Inc., 2021.

A. Geiger, Z. Wu, C. Potts, T. Icard, and N. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR, 01–03 Apr 2024.

D. Geiger and C. Meek. Quantifier elimination for statistical problems. *arXiv preprint arXiv:1301.6698*, 2013.

A. Gelman and G. Imbens. Why ask why? forward causal inference and reverse causal questions, 2013.

C. Genovese and L. Wasserman. Adaptive confidence bands. *The Annals of Statistics*, pages 875–905, 2008.

C. Glymour, R. Scheines, P. Spirtes, and K. Kelly. *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. Academic Press, 1987.

C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

B. R. Gordon, F. Zettelmeyer, N. Bhargava, and D. Chapsky. A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science*, 38(2):193–225, 2019.

P. Gradu, T. Zrnic, Y. Wang, and M. I. Jordan. Valid inference after causal discovery. *arXiv preprint arXiv:2208.05949*, 2022.

A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

Z. Griliches and W. M. Mason. Education, income, and ability. *Journal of political Economy*, 80(3, Part 2):S74–S103, 1972.

S. Gruber and M. Van Der Laan. tmle: an r package for targeted maximum likelihood estimation. *Journal of Statistical Software*, 51: 1–35, 2012.

G. Gui and O. Toubia. The challenge of using llms to simulate human behavior: A causal inference perspective. *arXiv preprint arXiv:2312.15524*, 2023.

L. Gui and V. Veitch. Causal estimation for text data with (apparent) overlap violations. *arXiv preprint arXiv:2210.00079*, 2022.

S. Gupta, R. Kohavi, D. Tang, Y. Xu, R. Andersen, E. Bakshy, N. Cardin, S. Chandran, N. Chen, D. Coey, et al. Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter*, 21(1):20–35, 2019.

S. Gupta, D. Childers, and Z. C. Lipton. Local causal discovery for estimating causal effects. In M. van der Schaar, C. Zhang, and D. Janzing, editors, *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pages 408–447. PMLR, 11–14 Apr 2023.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):1–12, 1943.

J. Hahn, K. Hirano, and D. Karlan. Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics*, 29 (1):96–108, 2011.

P. R. Hahn, J. S. Murray, and C. M. Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3): 965–1056, 2020.

J. Hájek. Local asymptotic minimax and admissibility in estimation. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 175–194, 1972.

S. Haneuse and A. Rotnitzky. Estimation of the effect of interventions that modify the received treatment. *Statistics in medicine*, 32(30): 5260–5277, 2013.

L. V. Hedges and I. Olkin. *Statistical methods for meta-analysis*. Academic press, 1985, 2014.

L. Henckel, E. Perković, and M. H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):579–599, 2022.

M. A. Hernán and J. M. Robins. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764, 2016.

M. A. Hernan and J. M. Robins. *Causal inference: What If*. CRC Boca Raton, FL:, 2020.

L. Hewitt, A. Ashokkumar, I. Ghezae, and R. Willer. Predicting results of social science experiments using large language models. *Preprint*, 2024.

A. B. Hill. Memories of the british streptomycin trial in tuberculosis: the first randomized clinical trial. *Controlled Clinical Trials*, 11(2): 77–79, 1990.

J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

S. M. Hill, L. M. Heiser, T. Cokelaer, M. Unger, N. K. Nesser, D. E. Carlin, Y. Zhang, A. Sokolov, E. O. Paull, C. K. Wong, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature methods*, 13(4):310–318, 2016.

K. Hirano and J. R. Porter. Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701, 2009.

P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

V. J. Hotz, G. W. Imbens, and J. H. Mortimer. Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, 125(1):241–270, 2005.

C. J. Howe, Z. D. Bailey, J. R. Raifman, and J. W. Jackson. Recommendations for using causal diagrams to study racial health disparities. *American Journal of Epidemiology*, 191(12):1981–1989, 2022.

P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

B. Huang, K. Zhang, M. Gong, and C. Glymour. Causal discovery from multiple data sets with non-identical variable sets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10153–10161, 2020.

Y. Huang and M. Valtorta. Pearl's calculus of intervention is complete. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence, 2006*, pages 13–16, 2006.

M. G. Hudgens and M. E. Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103 (482):832–842, 2008.

N. Huntington-Klein. *The effect: An introduction to research design and causality*. Chapman and Hall/CRC, 2021.

A. Hyttinen, F. Eberhardt, and M. Järvisalo. Constraint-based causal discovery: conflict resolution with answer set programming. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, page 340–349, Arlington, Virginia, USA, 2014. AUAI Press. ISBN 9780974903910.

A. Hyttinen, F. Eberhardt, and M. Järvisalo. Do-calculus when the true graph is unknown. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, page 395–404, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.

K. Imai and K. Nakamura. Causal representation learning with generative artificial intelligence: Application to texts as treatments. *arXiv preprint arXiv:2410.00903*, 2024.

K. Imai and K. Nakamura. Genai-powered inference. *arXiv preprint arXiv:2507.03897*, 2025.

G. Imbens and Y. Xu. Lalonde (1986) after nearly four decades: Lessons learned. *arXiv preprint arXiv:2406.00827*, 2024.

G. W. Imbens. Sensitivity to exogeneity assumptions in program evaluation. *The American Economic Review, Papers and Proceedings*, 93(2):126–132, 2003.

G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 61:467–476, 1994.

G. W. Imbens and D. B. Rubin. *Causal inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge, UK, 2015a.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015b.

J. W. Jackson. Meaningful causal decompositions in health equity research: definition, identification, and estimation through a weighting framework. *Epidemiology*, 32(2):282–290, 2021.

C. T. Jerzak, F. Johansson, and A. Daoud. Image-based treatment effect heterogeneity. *Proceedings of the Second Conference on Causal Learning and Reasoning (CLeaR), Proceedings of Machine Learning Research (PMLR)*, 213:531–552, 2023.

A. Jesson, S. Mindermann, Y. Gal, and U. Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. *arXiv preprint arXiv:2103.04850*, 2021.

J. Jin and V. Syrgkanis. Structure-agnostic optimality of doubly robust learning for treatment effect estimation. *arXiv preprint arXiv:2402.14264*, 2024.

Z. Jin, J. Liu, Z. Lyu, S. Poff, M. Sachan, R. Mihalcea, M. Diab, and B. Schölkopf. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*, 2023.

Z. Jin, J. Liu, Z. LYU, S. Poff, M. Sachan, R. Mihalcea, M. T. Diab, and B. Schölkopf. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*, 2024.

R. Johari, H. Li, I. Liskovich, and G. Y. Weintraub. Experimental design in two-sided platforms: An analysis of bias. *Management Science*, 2022.

G. Jonzon, M. C. Sachs, and E. E. Gabriel. Accessible computation of tight symbolic bounds on causal effects using an intuitive graphical interface. *arXiv preprint arXiv:2209.03657*, 2022.

M. Jordan, Y. Wang, and A. Zhou. Empirical gateaux derivatives for causal inference. *Advances in Neural Information Processing Systems*, 35:8512–8525, 2022.

M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann. Causal inference using graphical models with the r package pcalg. *Journal of statistical software*, 47:1–26, 2012.

N. Kallus and A. Zhou. Confounding-robust policy improvement. *arXiv preprint arXiv:1805.08593*, 2018.

N. Kallus and A. Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890, 2021.

J. D. Kang and J. L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, pages 523–539, 2007.

J. Karvanen, S. Tikka, and A. Hyttinen. Do-search: A tool for causal inference and study design with multiple data sources. *Epidemiology*, 32(1):111–119, 2021.

M. Kato, T. Ishihara, J. Honda, and Y. Narita. Efficient adaptive experimental design for average treatment effect estimation. *arXiv preprint arXiv:2002.05308*, 2020.

E. H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

E. H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245, 2017.

E. H. Kennedy, S. Balakrishnan, J. M. Robins, and L. Wasserman. Minimax rates for heterogeneous causal effect estimation. *The Annals of Statistics*, 52(2):793–816, 2024.

E. Kiciman, R. Ness, A. Sharma, and C. Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2023.

K. Kim, J. Kim, and E. H. Kennedy. Causal effects based on distributional distances. *arXiv preprint arXiv:1806.02935*, 2018.

E. Kirgios, S. Athey, A. L. Duckworth, D. Karlan, M. Luca, K. L. Milkman, and M. Offer-Westort. Does q&a boost engagement? health messaging experiments in the us and ghana. Technical report, National Bureau of Economic Research, 2025.

T. Kitagawa and A. Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

H. J. Kleven. Bunching. *Annual Review of Economics*, 8(1):435–464, 2016.

K. Kobalczyk and M. van der Schaar. Preference learning for ai alignment: a causal perspective. *arXiv preprint arXiv:2506.05967*, 2025.

D. Kumor, C. Cinelli, and E. Bareinboim. Efficient identification in linear structural causal models with auxiliary cutsets. In *International Conference on Machine Learning*, pages 5501–5510. PMLR, 2020.

D. Kurisu, Y. Zhou, T. Otsu, and H.-G. Müller. Geodesic causal inference. *arXiv preprint arXiv:2406.19604*, 2024.

T. L. Lai, H. Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

B. Laurent. Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24(2):659–681, 1996.

L. Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 1986.

E. E. Leamer. Let's take the con out of econometrics. *The American Economic Review*, 73(1):31–43, 1983.

J. J. Lee, R. Bhattacharya, R. Nabi, and I. Shpitser. Ananke: A python package for causal inference using graphical models. *arXiv preprint arXiv:2301.11477*, 2023.

S. Lee and E. Bareinboim. Structural causal bandits: Where to intervene? *Advances in neural information processing systems*, 31, 2018.

L. Leqi and E. H. Kennedy. Median optimal treatment regimes. *arXiv preprint arXiv:2103.01802*, 2021.

M. P. Leung. Causal inference under approximate neighborhood interference. *Econometrica*, 90(1):267–293, 2022.

G. Lewis and V. Syrgkanis. Double/debiased machine learning for dynamic treatment effects. In *NeurIPS*, pages 22695–22707, 2021.

A. Li, H. Chen, H. Namkoong, and T. Peng. Llm generated persona is a promise with a catch. *arXiv preprint arXiv:2503.16527*, 2025.

F. Li, P. Ding, and F. Mealli. Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A*, 381 (2247):20220153, 2023.

A. Lin, A. Merchant, S. K. Sarkar, and A. D'Amour. Universal causal evaluation engine: An api for empirically evaluating causal inference models. In *The 2019 ACM SIGKDD Workshop on Causal Discovery*, pages 50–58. PMLR, 2019.

V. Lin, E. Ben-Michael, and L.-P. Morency. Optimizing language models for human preferences is a causal inference problem. *arXiv preprint arXiv:2402.14979*, 2024a.

V. Lin, L.-P. Morency, and E. Ben-Michael. Isolated causal effects of natural language. *arXiv preprint arXiv:2410.14812*, 2024b.

L. Liu, X. Wang, and Y. Wang. Root-n consistent semiparametric learning with high-dimensional nuisance functions under minimal sparsity. *arXiv preprint arXiv:2305.04174*, 2023.

L. T. Liu, I. D. Raji, A. Zhou, L. Guerdan, J. Hullman, D. Malinsky, B. Wilder, S. Zhang, H. Adam, A. Coston, et al. Bridging prediction and intervention problems in social systems. *arXiv preprint arXiv:2507.05216*, 2025.

M. G. Low. On nonparametric confidence intervals. *The Annals of Statistics*, 25(6):2547–2554, 1997.

S. Lu, L. Shi, Y. Fang, W. Zhang, and P. Ding. Design-based causal inference in bipartite experiments. *arXiv preprint arXiv:2501.09844*, 2025.

Y. Lu, A. Meisami, and A. Tewari. Causal bandits with unknown graph structure. *Advances in Neural Information Processing Systems*, 34: 24817–24828, 2021.

A. Luedtke. Simplifying debiased inference via automatic differentiation and probabilistic programming. *arXiv preprint arXiv:2405.08675*, 2024.

A. Luedtke and A. Chambaz. Performance guarantees for policy learning. In *Annales de l'IHP Probabilites et statistiques*, volume 56, page 2162. NIH Public Access, 2020.

A. R. Luedtke, M. Carone, and M. J. van der Laan. Discussion of "deductive derivation and turing-computerization of semiparametric efficient estimation" by frangakis et al. *Biometrics*, 71(4):875–879, 2015.

T. Ma, K. A. Verchand, and R. J. Samworth. High-probability minimax lower bounds. *arXiv preprint arXiv:2406.13447*, 2024.

Y. Ma, D. Frauen, E. Javurek, and S. Feuerriegel. Foundation models for causal inference via prior-data fitted networks. *arXiv preprint arXiv:2506.10914*, 2025.

J. Maasch, W. Pan, S. Gupta, V. Kuleshov, K. Gan, and F. Wang. Local discovery by partitioning: Polynomial-time causal discovery around exposure-outcome pairs. In N. Kiyavash and J. M. Mooij, editors, *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 2350–2382. PMLR, 15–19 Jul 2024.

A. Maiti, D. Plecko, and E. Bareinboim. Counterfactual identification under monotonicity constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26841–26850, 2025.

D. Malinsky and D. Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470, 2018.

B. S. Manning, K. Zhu, and J. J. Horton. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research, 2024.

C. Manski. Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, 60(3):531–542, 1993a.

C. F. Manski. Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542, 1993b.

C. F. Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, 2013.

J. D. Marion and A. D. Althouse. The use of historical controls in clinical trials. *JAMA*, 330(15):1484–1485, 2023.

M. A. Masten and A. Poirier. Identification of treatment effects under conditional partial independence. *Econometrica*, 86(1):317–351, 2018.

A. McClean, S. Balakrishnan, E. H. Kennedy, and L. Wasserman. Double cross-fit doubly robust estimators: Beyond series regression. *arXiv preprint arXiv:2403.15175*, 2024.

D. McFadden. Economic choices. *American economic review*, 91(3): 351–378, 2001.

S. McGrath and R. Mukherjee. Nuisance function tuning and sample splitting for optimal doubly robust estimation. *arXiv preprint arXiv:2212.14857*, 2022.

N. Meinshausen, A. Hauser, J. M. Mooij, J. Peters, P. Versteeg, and P. Bühlmann. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368, 2016.

M. Melchert and A. List. The thalidomide saga. *The international journal of biochemistry & cell biology*, 39(7-8):1489–1499, 2007.

C. H. Miles. On the causal interpretation of randomised interventional indirect effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1154–1172, 2023.

N. Mitra, J. Roy, and D. Small. The future of causal inference. *American Journal of Epidemiology*, 191(10):1671–1676, 2022.

M. N. Mittinty and S. Vansteelandt. Longitudinal mediation analysis using natural effect models. *American Journal of Epidemiology*, 189(11):1427–1435, 2020.

K. Mohan and J. Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037, 2021.

J. M. Mooij and T. Claassen. Constraint-based causal discovery using partial ancestral graphs in the presence of cycles. In *Conference on Uncertainty in Artificial Intelligence*, pages 1159–1168. Pmlr, 2020.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21 (99):1–108, 2020.

S. L. Morgan and C. Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2015.

R. Mukherjee, E. Tchetgen Tchetgen, and J. Robins. Lepski's method and adaptive estimation of nonlinear integral functionals of density. *arXiv preprint arXiv:1508.00249*, 2015.

E. Munro, S. Wager, and K. Xu. Treatment effects in market equilibrium. *arXiv preprint arXiv:2109.11647*, 2021.

S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2): 331–355, 2003.

R. Nabi, M. Bonvini, E. H. Kennedy, M.-Y. Huang, M. Smid, and D. O. Scharfstein. Semiparametric sensitivity analysis: unmeasured confounding in observational studies. *Biometrics*, 80(4):ujae106, 2024.

I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, and S. A. Murphy. Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, pages 1–17, 2018.

N. Nance, M. L. Petersen, M. van der Laan, and L. B. Balzer. The causal roadmap and simulations to improve the rigor and reproducibility of real-data applications. *Epidemiology*, pages 10–1097, 2024.

W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

W. K. Newey and J. R. Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

M. Neykov, S. Balakrishnan, and L. Wasserman. Minimax optimal conditional independence testing. *The Annals of Statistics*, 49(4): 2151–2177, 2021.

J. Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5 (4):465–472, 1923/1990.

X. Nie, E. Brunskill, and S. Wager. Learning when-to-treat policies. *Journal of the American Statistical Association*, 116(533):392–409, 2021.

E. L. Ogburn, O. Sofrygin, I. Diaz, and M. J. Van der Laan. Causal inference for social network data. *Journal of the American Statistical Association*, 119(545):597–611, 2024.

E. Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019.

G. Papadogeorgou, F. Mealli, and C. M. Zigler. Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics*, 75(3):778–787, 2019.

J. Pearl. A probabilistic calculus of actions. In *Uncertainty in artificial intelligence*, pages 454–462. Elsevier, 1994.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4): 669–688, 1995.

J. Pearl. *Causality*. Cambridge University Press, New York, 2000.

J. Pearl. Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc., 2001.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2009.

J. Pearl and D. Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

R. L. Prentice. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8(4):431–440, 1989.

Y. Qian, Y. Yi, J. Shao, Y. Yi, N. Mayer-Hamblett, P. J. Heagerty, and T. Ye. From estimands to robust inference of treatment effects in platform trials. *arXiv preprint arXiv:2411.12944*, 2024.

J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics*, 3:121–129, 2017.

B. Recht. A bureaucratic theory of statistics. *arXiv preprint arXiv:2501.03457*, 2025.

H. Reichenbach. *The Direction of Time*. Dover Publications, Mineola, N.Y., 1956.

T. S. Richardson and J. M. Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30): 2013, 2013.

J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12): 1393–1512, 1986.

J. Robins, L. Li, E. Tchetgen Tchetgen, and A. van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

J. Robins, E. Tchetgen Tchetgen, L. Li, and A. van der Vaart. Semiparametric minimax rates. *Electronic journal of statistics*, 3:1305, 2009.

J. M. Robins. Association, causation, and marginal structural models. *Synthese*, 121(1):151–179, 1999.

J. M. Robins. Semantics of causal dag models and the identification of direct and indirect effects. *Highly structured stochastic systems*, pages 70–82, 2003.

J. M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155, 1992.

J. M. Robins, A. Rotnitzky, and D. O. Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. Springer, 2000.

J. M. Robins, L. Li, R. Mukherjee, E. J. Tchetgen Tchetgen, and A. W. van der Vaart. Minimax estimation of a functional on a structured high dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

P. Rosenbaum. *Replication and evidence factors in observational studies*. Chapman and Hall/CRC, 2021.

P. R. Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987.

P. R. Rosenbaum. Observational studies. In *Observational studies*, pages 1–17. Springer, 2002.

P. R. Rosenbaum. Design sensitivity in observational studies. *Biometrika*, 91(1):153–164, 2004.

P. R. Rosenbaum. Evidence factors in observational studies. *Biometrika*, 97(2):333–345, 2010.

P. R. Rosenbaum. *Design of observational studies*. Springer, 2020.

P. R. Rosenbaum and D. B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 212–218, 1983a.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55, 1983b.

P. R. Rosenbaum and J. H. Silber. Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association*, 104(488):1398–1405, 2009.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

D. B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371):591–593, 1980.

J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. Mahecha, J. Muñoz-Marí, E. van Ness, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10:2553, 2019.

B. Sacerdote. Peer effects with random assignment: results for dartmouth roommates. *Quarterly Journal of Economics*, 116(2):681–704, 2001.

K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

M. C. Sachs, G. Jonzon, A. Sjölander, and E. E. Gabriel. A general method for deriving tight symbolic bounds on causal effects. *Journal of Computational and Graphical Statistics*, 32(2):567–576, 2023.

E. Sanderson, M. M. Glymour, M. V. Holmes, H. Kang, J. Morrison, M. R. Munafò, T. Palmer, C. M. Schooling, C. Wallace, Q. Zhao, et al. Mendelian randomization. *Nature Reviews Methods Primers*, 2(1):6, 2022.

M. Santacatterina, F. M. Giron, X. Zhang, and I. Díaz. Identification and estimation of causal effects using non-concurrent controls in platform trials. *Statistics in Medicine*, 44(6):e70017, 2025.

M. Saveski, J. Pouget-Abadie, G. Saint-Jacques, W. Duan, S. Ghosh, Y. Xu, and E. M. Airoldi. Detecting network effects: Randomizing over randomized experiments. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1027–1035, 2017.

F. Sävje, P. Aronow, and M. Hudgens. Average treatment effects in the presence of unknown interference. *Annals of statistics*, 49(2):673, 2021.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

M. Schubert, T. Claassen, and S. Magliacane. Snap: Sequential non-ancestor pruning for targeted causal effect estimation with an unknown graph. In Y. Li, S. Mandt, S. Agrawal, and E. Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3340–3348. PMLR, 03–05 May 2025.

M. S. Schuler, D. L. Coffman, E. A. Stuart, T. Q. Nguyen, B. Vegetabile, and D. F. McCaffrey. Practical challenges in mediation analysis: a guide for applied researchers. *Health Services and Outcomes Research Methodology*, pages 1–28, 2024.

V. Semenova and V. Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021.

W. R. Shadish, M. H. Clark, and P. M. Steiner. Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American statistical association*, 103(484):1334–1344, 2008.

R. D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3), June 2020. ISSN 0090-5364.

Z. Shahn, M. A. Hernán, and J. M. Robins. A formal causal interpretation of the case-crossover design. *Biometrics*, 79(2):1330–1343, 2023.

E. Sherman and I. Shpitser. Identification and estimation of causal effects from dependent data. *Advances in neural information processing systems*, 31, 2018.

E. Sherman, D. Arbour, and I. Shpitser. General identification of dynamic treatment regimes under interference. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3917–3927. PMLR, 26–28 Aug 2020.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *AAAI*, pages 1219–1226, 2006.

A. Sjölander. Sharp bounds for causal effects based on ding and vanderweele's sensitivity parameters. *Journal of Causal Inference*, 12 (1):20230019, 2024.

E. Smucler, A. Rotnitzky, and J. M. Robins. A unifying approach for doubly-robust l1 regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737*, 2019.

J. Snow. On the mode of communication of cholera. *Edinburgh medical journal*, 1(7):668, 1856.

M. E. Sobel. What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407, 2006.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2001.

J. H. Stock and F. Trebbi. Retrospectives: Who invented instrumental variable regression? *Journal of Economic Perspectives*, 17(3):177–194, 2003.

C. J. Stone. Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pages 1348–1360, 1980.

R. H. Strotz and H. O. Wold. Recursive vs. nonrecursive systems: An attempt at synthesis (part i of a triptych on causal chain systems). *Econometrica: Journal of the Econometric Society*, pages 417–427, 1960.

R. S. Sutton, A. G. Barto, et al. *Reinforcement learning: An introduction*. MIT press, 1998.

V. Syrgkanis, G. Lewis, M. Oprescu, M. Hei, K. Battocchi, E. Dillon, J. Pan, Y. Wu, P. Lo, H. Chen, et al. Causal inference and machine learning in practice with econml and causalml: Industrial use cases at microsoft, tripadvisor, uber. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 4072–4073, 2021.

Z. Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476): 1619–1637, 2006.

S. J. Taylor and D. Eckles. Randomized experiments to detect and estimate social influence in networks. *Complex spreading phenomena in social systems: Influence and contagion in real-world social networks*, pages 289–322, 2018.

E. Tchetgen Tchetgen. The control outcome calibration approach for causal inference with unobserved confounding. *American Journal of Epidemiology*, 179(5):633–640, 2014.

E. J. Tchetgen Tchetgen and I. Shpitser. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics*, 40(3):1816, 2012.

E. J. Tchetgen Tchetgen, A. Ying, Y. Cui, X. Shi, and W. Miao. An introduction to proximal causal inference. *Statistical Science*, 39 (3):375–390, 2024.

J. Textor, J. Hardt, and S. Knüppel. Dagitty: a graphical tool for analyzing causal diagrams. *Epidemiology*, 22(5):745, 2011.

D. L. Thistlethwaite and D. T. Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 51(6):309, 1960.

J. Tian. *Studies in causal reasoning and learning*. University of California, Los Angeles, 2002.

R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32, 2019.

S. Tikka and J. Karvanen. Identifying causal effects with the r package causaleffect. *arXiv preprint arXiv:1806.07161*, 2018.

R. Tillman and P. Spirtes. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 3–15, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

R. E. Tillman and F. Eberhardt. Learning causal structure from multiple datasets with similar variable sets. *Behaviormetrika*, 41(1): 41–64, 2014.

J. Tinbergen. Determination and interpretation of supply curves: an example. *Zeitschrift fur Nationalokonomie*, 1(5):669–679, 1930.

J. Tinbergen. Econometric business cycle research. *The Review of Economic Studies*, 7(2):73–90, 1940.

S. Triantafillou and I. Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16(66):2147–2205, 2015a.

S. Triantafillou and I. Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *The Journal of Machine Learning Research*, 16(1):2147–2205, 2015b.

S. Triantafillou, F. Jabbari, and G. F. Cooper. Learning treatment effects from observational and experimental data. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 7126–7146. PMLR, 25–27 Apr 2023.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

L. van der Laan, A. Bibaut, N. Kallus, and A. Luedtke. Automatic debiased machine learning for smooth functionals of nonparametric m-estimands. *arXiv preprint arXiv:2501.11868*, 2025.

M. van der Laan. A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *The international journal of biostatistics*, 13(2), 2017.

M. J. Van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), 2006.

M. J. van der Laan and M. L. Petersen. Direct effect models. *The international journal of biostatistics*, 4(1), 2008.

A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

T. J. VanderWeele. A unification of mediation and interaction: a 4-way decomposition. *Epidemiology*, 25(5):749–761, 2014.

T. J. VanderWeele. *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press, 2015.

T. J. VanderWeele. Outcome-wide epidemiology. *Epidemiology*, 28 (3):399–402, 2017.

T. J. Vanderweele and O. A. Arah. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology (Cambridge, Mass.)*, 22(1):42–52, Jan. 2011.

T. J. VanderWeele and J. M. Robins. Signed directed acyclic graphs for causal inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(1):111–127, 2010.

T. J. VanderWeele and E. J. Tchetgen Tchetgen. Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):917–938, 2017.

A. Vashishtha, A. G. Reddy, A. Kumar, S. Bachu, V. N. Balasubramanian, and A. Sharma. Causal order: The key to leveraging imperfect experts in causal inference. In *The Thirteenth International Conference on Learning Representations*, 2025.

V. Veitch and A. Zaveri. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *Advances in Neural Information Processing Systems*, 33:10999–11009, 2020.

V. Veitch, D. Sridhar, and D. Blei. Adapting text embeddings for causal inference. In *Conference on uncertainty in artificial intelligence*, pages 919–928. PMLR, 2020.

V. Verma, S. Acharya, S. Simko, A. Haghighat, D. Bhardwaj, M. Sachan, D. Janzing, B. Schölkopf, and Z. Jin. Causal ai scientist: Facilitating causal data science with large language models, 2025. Preprint.

L. von der Heyde, A.-C. Haensch, and A. Wenz. Vox populi, vox ai? using large language models to estimate german vote choice. *Social Science Computer Review*, page 08944393251337014, 2025.

M. J. Vowels, N. C. Camgoz, and R. Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

S. Wager. *Causal Inference: A Statistical Learning Approach*. Cambridge University Press, 2024.

C. Wang, Y. Zhou, Q. Zhao, and Z. Geng. Discovering and orienting the edges connected to a target variable in a DAG via a sequential local learning approach. *Comput. Stat. Data Anal.*, 77:252–266, 2014.

D. Wang, T.-Y. Hsu, Y. Lu, H. Gu, L. Cui, Y. Xie, W. Headean, B. Yao, A. Veeragouni, J. Liu, et al. Agenta/b: Automated and scalable web a/btesting with interactive llm agents. *arXiv preprint arXiv:2504.09723*, 2025a.

G. Wang, J. Li, Y. Sun, X. Chen, C. Liu, Y. Wu, M. Lu, S. Song, and Y. A. Yadkori. Hierarchical reasoning model. *arXiv preprint arXiv:2506.21734*, 2025b.

L. Wang, Y. Zhou, R. Song, and B. Sherwood. Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, 113 (523):1243–1254, 2018.

Y. Wang and R. D. Shah. Debiased inverse propensity score weighting for estimation of average treatment effects with high-dimensional confounders. *The Annals of Statistics*, 52(5):1978–2003, 2024.

L. Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

D. Watson and R. Silva. Causal discovery under a confounder blanket. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

S. Weichwald, S. W. Mogensen, T. E. Lee, D. Baumann, O. Kroemer, I. Guyon, S. Trimpe, J. Peters, and N. Pfister. Learning by doing: Controlling a dynamical system using causality, control, and reinforcement learning. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 246–258. PMLR, 2022.

L. Wittgenstein. *Philosophical Investigations*. New York: MacMillan, 1958.

V. C. Wong, K. Anglin, and P. M. Steiner. Design-based approaches to causal replication studies. *Prevention Science*, pages 1–16, 2021.

P. G. Wright. *The tariff on animal and vegetable oils*. Number 26. Macmillan, 1928.

S. Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557, 1921.

S. Wright. The method of path coefficients. *The annals of mathematical statistics*, 5(3):161–215, 1934.

C. J. Wu and M. S. Hamada. *Experiments: planning, analysis, and optimization*. John Wiley & Sons, 2011.

Z. Wu, A. Geiger, T. Icard, C. Potts, and N. Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 78205–78226. Curran Associates, Inc., 2023.

S. Yadlowsky, H. Namkoong, S. Basu, J. Duchi, and L. Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*, 2018.

T. Yamamoto. Understanding the past: Statistical analysis of causal attribution. *American Journal of Political Science*, 56(1):237–256, 2012.

R. Yu and P. Ding. Balancing weights for causal inference in observational factorial studies. *arXiv preprint arXiv:2310.04660*, 2023.

A. Zanga, E. Ozkirimli, and F. Stella. A survey on causal discovery: theory and practice. *International Journal of Approximate Reasoning*, 151:101–129, 2022.

Z. Zeng, S. Balakrishnan, Y. Han, and E. H. Kennedy. Causal inference with high-dimensional discrete covariates. *arXiv preprint arXiv:2405.00118*, 2024.

C. Zhang, C. Cinelli, B. Chen, and J. Pearl. Exploiting equality constraints in causal inference. In *International Conference on Artificial Intelligence and Statistics*, pages 1630–1638. PMLR, 2021.

K. Zhang, M. Gong, J. D. Ramsey, K. Batmanghelich, P. Spirtes, and C. Glymour. Causal discovery with linear non-gaussian models under measurement error: Structural identifiability results. In *UAI*, pages 1063–1072, 2018.

A. Zhao and P. Ding. Regression-based causal inference with factorial experiments: estimands, model specifications and design-based properties. *Biometrika*, 109(3):799–815, 2022.

Q. Zhao, L. J. Keele, and D. S. Small. Comment: Will competition-winning methods for causal inference also succeed in practice? *Statistical Science*, 34(1):72–76, 2019a.

Q. Zhao, D. S. Small, and B. B. Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):735–761, 2019b.