Representation Learning Based Causal Inference in Observational Studies

Danni Lu

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Feng Guo, Chair
Fan Li
Inyoung Kim
Xinwei Deng

January 26, 2021
Blacksburg, Virginia

Keywords: Causal Inference, Representation Learning, Naturalistic Driving Study,
Propensity Score, Representation Balancing, Invariant Risk Minimization.

Copyright 2021, Danni Lu

Representation Learning Based Causal Inference in Observational Studies

Danni Lu

(ABSTRACT)

This dissertation investigates novel statistical approaches for causal effect estimation in observational settings, where controlled experimentation is infeasible and confounding is the main hurdle in estimating causal effect. As such, deconfounding constructs the main subject of this dissertation, that is (*i*) to restore the covariate balance between treatment groups and (*ii*) to attenuate spurious correlations in training data to derive valid causal conclusions that generalize. By incorporating ideas from representation learning, adversarial matching, generative causal estimation, and invariant risk modeling, this dissertation establishes a causal framework that balances the covariate distribution in latent representation space to yield individualized estimations, and further contributes novel perspectives on causal effect estimation based on invariance principles.

The dissertation begins with a systematic review and examination of classical propensity score based balancing schemes for population-level causal effect estimation, presented in Chapter 2. Three causal estimands that target different foci in the population are considered: average treatment effect on the whole population (ATE), average treatment effect on the treated population (ATT), and average treatment effect on the overlap population (ATO). The procedure is demonstrated in a naturalistic driving study (NDS) to evaluate the causal effect of cellphone distraction on crash risk. While highlighting the importance of adopting causal perspectives in analyzing risk factors, discussions on the limitations in balance efficiency, robustness against high-dimensional data & complex interactions, and the need for individualization are provided to motivate subsequent developments.

Chapter 3 presents a novel generative Bayesian causal estimation framework named *Balancing Variational Neural Inference of Causal Effects* (BV-NICE). Via appealing to

the Robinson factorization and a latent Bayesian model, a novel variational bound on likelihood is derived, explicitly characterized by the causal effect and propensity score. Notably, by treating observed variables as noisy proxies of unmeasurable latent confounders, the variational posterior approximation is re-purposed as a stochastic feature encoder that fully acknowledges representation uncertainties. To resolve the imbalance in representations, BV-NICE enforces KL-regularization on the respective representation marginals using Fenchel mini-max learning, justified by a new generalization bound on the counterfactual prediction accuracy. The robustness and effectiveness of this framework are demonstrated through an extensive set of tests against competing solutions on semi-synthetic and real-world datasets.

In recognition of the reliability issue when extending causal conclusions beyond training distributions, Chapter 4 argues ascertaining causal stability is the key and introduces a novel procedure called *Risk Invariant Causal Estimation* (RICE). By carefully re-examining the relationship between statistical invariance and causality, RICE cleverly leverages the observed data disparities to enable the identification of stable causal effects. Concretely, the causal inference objective is reformulated under the framework of invariant risk modeling (IRM), where a population-optimality penalty is enforced to filter out un-generalizable effects across heterogeneous populations. Importantly, RICE allows settings where counterfactual reasoning with unobserved confounding or biased sampling designs become feasible. The effectiveness of this new proposal is verified with respect to a variety of study designs on real and synthetic data.

In summary, this dissertation presents a flexible causal inference framework that acknowledges the representation uncertainties and data heterogeneities. It enjoys three merits: improved balance to complex covariate interactions, enhanced robustness to unobservable latent confounders, and better generalizability to novel populations.

Representation Learning Based Causal Inference in Observational Studies

Danni Lu

(GENERAL AUDIENCE ABSTRACT)

Reasoning cause and effect is the innate ability of a human. While the drive to understand cause and effect is instinct, the rigorous reasoning process is usually trained through the observation of countless trials and failures. In this dissertation, we embark on a journey to explore various principles and novel statistical approaches for causal inference in observational studies. Throughout the dissertation, we focus on the causal effect estimation which answers questions like *"what if"* and *"what could have happened"*. The causal effect of a treatment is measured by comparing the outcomes corresponding to different treatment levels of the same unit, *e.g. "what if the unit is treated instead of not treated?"*. The challenge lies in the fact that *i)* a unit only receives one treatment at a time and therefore it is impossible to directly compare outcomes of different treatment levels; *ii)* comparing the outcomes across different units may involve bias due to confounding as the treatment assignment potentially follows a systematic mechanism. Therefore, deconfounding constructs the main hurdle in estimating causal effects.

This dissertation presents two parallel principles of deconfounding: *i) balancing, i.e.,* comparing difference under similar conditions; *ii) contrasting, i.e.,* extracting invariance under heterogeneous conditions. Chapter 2 and Chapter 3 explore causal effect through balancing, with the former systematically reviews a classical propensity score weighting approach in a conventional data setting and the latter presents a novel generative Bayesian

framework named Balancing Variational Neural Inference of Causal Effects(BV-NICE) for high-dimensional, complex, and noisy observational data. It incorporates the advance deep learning techniques of representation learning, adversarial learning, and variational inference. The robustness and effectiveness of the proposed framework are demonstrated through an extensive set of experiments. Chapter 4 extracts causal effect through contrasting, emphasizing that ascertaining stability is the key of causality. A novel causal effect estimating procedure called Risk Invariant Causal Estimation(RICE) is proposed that leverages the observed data disparities to enable the identification of stable causal effects. The improved generalizability of RICE is demonstrated through synthetic data with different structures, compared with state-of-art models.

In summary, this dissertation presents a flexible causal inference framework that acknowledges the data uncertainties and heterogeneities. By promoting two different aspects of causal principles and integrating advance deep learning techniques, the proposed framework shows improved balance for complex covariate interactions, enhanced robustness for unobservable latent confounders, and better generalizability for novel populations.

Dedicated to my beloved parents and sister.

Acknowledgments

I would like first to express my sincere appreciation to my supervisor, Professor Feng Guo, whose expertise in both statistics and transportation was invaluable in formulating the research questions and methodologies. During the past five years, he has spent countless hours to guide my direction, sharpen my thinking, and polish my writing. His timely and scholarly advice not only to a great extent helped me accomplish this work but also brought the work to a higher level.

I want to express my gratitude to Dr.Fan Li, Dr.Inyoung Kim, and Dr.Xinwei Deng for serving on my committee, for giving great lectures, and for their valuable insights and detailed comments on my work. I am also very grateful to Dr.Chenyang Tao for many brilliant ideas on the data analysis and methodologies. This dissertation would not have been possible without many collaborative efforts from Dr.Fan Li and Dr.Chenyang Tao.

Additional acknowledgements to the Department of Statistics at Virginia Tech for providing valuable course resources and creating such a friendly and mutual-respectful community; to Virginia Tech Transportation Institute for providing the data, research opportunities, study space, and computing resources.

Throughout my journey at Virginia Tech, I am blessed with many friends from Statistics Department and Center for Sustainable Transportation Infrastructure. I am deeply grateful for their company and encouragement. Special thanks to my boyfriend Hao Zhang for being the inspiration along the wonderful journey, and for his love and supports.

Last but not the least, I would like to give my sincerest appreciation to my family for their unconditional love, support, and encouragement since my childhood. They taught me to be diligent, curious, optimistic, and confident. I wouldn't be able to chase my dreams without their immeasurable supports.

Contents

List of Figures

List of Tables

Chapter 1

Cause and Effect: An introduction

Reasoning cause and effect is the innate ability of a human. A toddler knows to cry in seeking attention as she thinks this behavior will draw her parents to attend to her needs since this has worked repeatedly in the past. A teenager chooses to actively engage in sports and bands as he believes it will make him a cool kid and become popular in school. As an adult, we constantly ask ourselves why things happen in a certain way, what are the real causes, and how can we make a change. Such reasoning of causation happens throughout our entire life to inform actions that optimize future outcomes. While the drive to understand cause and effect is instinct, a hallmark feature of human intelligence is that the logic behind causal reasoning is not only learned from their trials and errors but also, more importantly, through observations. With the advent of statistical science, such concepts have been formalized mathematically under the name observational causal inference, enabling machine intelligence to learn and reason like a human. In this chapter, we embark on a journey to the history and science of causality, concluding with some remarks on the frontier challenges this dissertation hopes to address.

1.1 Causality: a historical account

The earliest remarks on dissecting causation from association dates back to the early 1900s. As Fisher (1919) wrote in his paper: "*If we say, 'This boy has grown tall because he has been well fed', we are not merely tracing out cause and effect in an individual instance; we are suggesting that he might quite probably have been worse fed, and that in this case he would have been shorter*". Here, causation tries to emphasize the logic, or direction, of a relationship and emphasizes the comparison on the same unit. Therefore, we care about "what actual happened" as well as "what could have happened" if we changed the condition "well fed or not". This concept was not explicitly shaped until Neyman (1923) formalized it in the name of *potential outcomes*. In this example, the treatment of interest is "well fed or not", and the potential outcomes are outcomes under all possible treatments, *i.e.,* the height of the boy if he has been well fed and the height of the boy if he has not been well fed. Neyman (1923) implied that the consideration of treatment assignment and the stability of treatment effect is essential in order to estimate the effect. The recommendation of randomizing treatment to units in experiment design was first proposed by Fisher (1925), and is still considered the gold standard for statistical inferences in new drug approval today. While Fisher and Neyman have laid the foundation of causal inference, the concept of potential outcomes was rarely mentioned in the literature in the following decades. Researches continued to use observed outcomes to model treatment effects and causal inferences were made only in the context of randomized experiments.

It was not until a half century later, potential outcomes and the treatment assignment mechanism were re-introduced in the context of observational studies to adjust causal effect

estimation. When physical randomization is not an option, Rubin (1974b) pointed out, it is still possible to estimate causal effects through carefully controlled non-randomized data by enforcing covariate balance: treatment assignment becomes equally likely conditional on each covariate level after adjustment. For example, matching and covariate adjustment are two frequently adopted remedies to compensate the lack of randomization and improve the covariate balance across treatment groups (Rubin, 1976, 1977). Further, by explicitly incorporating treatment assignment mechanism in the Bayesian framework, Rubin (1978) proposed a unified perspective of causal effect estimation for both randomized and observational studies. These efforts constitute the primary features of *Rubin Causal Model* (RCM), whose name was first coined by Holland (1986). RCM, along with *structural equation models* (SEM) (Duncan, 2014, Goldberger, 1972) and graphical models (Pearl, 1995, 2000), represent the most distinguished contemporary advances in the statistical analysis of causality.

To be noticed, causality has a broader scope outside statistical sense. For example, from an engineering perspective, the cause of road crashes can be investigated through crash reconstructions (Davis, 2014, Limpert, 2020); the stealthy malware activities can be traced by temporal and semantic based traffic causality analysis (Zhang et al., 2014, 2016). This dissertation explores causality mainly from statistical perspective and builds on the work of RCM framework.

1.2 Concepts and assumptions

In this section, we briefly review the basic concepts and assumptions under the Rubin causal model (RCM) framework. We consider a basic setup where the treatment variable is binary (*i.e.*, 1: treated, 0: untreated), and seek answer to the intuitive question: *"What is the effect of applying the treatment to a population or a unit (relative to no treatment)?"*. This problem can be formalized in the language of *potential outcomes* (Neyman, 1923, Rubin, 1974b). Let $T_i \in \{0, 1\}$ be the treatment received by unit i, and $[Y_i(1), Y_i(0)]$ be the two potential outcomes with and without the treatment. A fundamental difference to a standard supervised learning setting is that, we only observe one of the potential outcomes that corresponds to the assigned treatment, $Y_i \triangleq Y(T_i) = T_i Y_i(1) + (1 - T_i)Y_i(0)$, known as the *factual data*, and we call the unobserved potential outcome $Y_i^{\mathtt{cf}} \triangleq Y(\neg T_i)$ *counterfactual data*, where $\neg$ is the negate operator.

In RCM, the treatment effect is mathematically defined as the contrast of potential outcomes. Specifically, the treatment effect, typically denoted as τ, can be measured as the difference in expectation for continuous outcomes, or in forms of risk difference, risk ratio, and odds ratio for binary outcomes. For example, the treatment effect for individual i can be defined as Eq. 1.1:

$$\tau_i \triangleq \mathbb{E}[Y_i(1) - Y_i(0)] \tag{1.1}$$

While the individual causal effect can not be identified due to the missing counterfactuals, the aggregated causal effect, *i.e, average causal effect* in the population, is identifiable under additional assumptions. The idea is to use the outcomes from other units to fill in

the missing puzzle.

Since causal effect is essentially the contrast between different treatments, it is intuitive that we should only borrow information from statistically similar units from the opposite treatment group. To inform such unit-level similarity, we collect additional unit characteristics and summarize them into covariate X_i. For example, in the investigation of the impact of cellphone distraction on motor vehicle crash, we are interested in the safety outcome difference with and without cellphone distraction. Ideally, the comparison is made for the same driver in the same driving context. A less convincing comparison is that between a teenager driver and an adult driver, as the difference in safety outcome can be a consequence of the driver's age difference rather than the distraction status. In statistical parlance, the driver's age in this case is called a *confounder*, which confusticates statistical analysis. More formally, *confounder* refers to a variable that affect both treatment (*i.e.*, cellphone distraction) and outcome (*i.e.*, vehicle crash). The causal effect is identifiable only when there is no unobserved confounder, or the unmeasured confounding is missing at random (Rubin, 1976). In the causality literature, this prerequisite is summarized by the assumption below.

Assumption 1.1 (**Unconfoundedness).** There is no unobserved confounder, the treatment is randomized given observed covariates $\boldsymbol{X}$:

$$\{Y(1), Y(0)\} \perp\!\!\!\perp T|\boldsymbol{X} \tag{1.2}$$

Under the unconfoundedness assumption, we have $Pr(Y(t)|\boldsymbol{X}) = Pr(Y|\boldsymbol{X}, T = t)$, which enables counterfactual reasoning based on the factual observations.

Apart from the confounders, another critical component in the identification of causal

effect is the treatment assignment mechanism. Heuristically, in order to make valid outcome contrasts, we must find statistical units from respective treatment groups that are similar. This requirement can be absorbed by the notation of *overlap*, which is guaranteed by a probabilistic treatment assignment mechanism if the following assumption wrt the observed covariates holds true (Rubin, 1977).

Assumption 1.2 (Overlap, positivity). Given the observed covariates and potential outcomes, the treatment assignment is probabilistic:

$$0 < Pr(T|\boldsymbol{X}, Y(0), Y(1)) < 1 \tag{1.3}$$

This assumption ensures that, given any unit, treated or otherwise, we have non-zero probability observing another unit from the opposite treatment group to make valid comparisons. A treatment assignment mechanism is called *strong ignorable* if both Assumptions 1.1-1.2 are satisfied (Rosenbaum and Rubin, 1983a).

Importantly, there are other seemingly trivial subtleties one must account for to ensure the validity of causal conclusions. When estimating average causal effect from multiple units, it is important to assume that there is no interference among each other (*e.g.*, competing for resources). Another trivial but worth mentioning prerequisite the treatment is uniquely defined (whose effect should be stable). These two conditions were summarized by Rubin (1980) and termed as *the stable unit treatment value assumption* (SUTVA):

Assumption 1.3. SUTVA:

(i) No interference: units do not interfere with each other; the treatment assignment of one unit does not affect the outcome of another unit.

(ii) There is only one version of treatment.

In the reminder of the dissertation, we will always assume SUTVA and only focus on the discussion of the strongly ignorability assumption.

More generally, causality is usually described in the language of *do*-calculus advocated by Pearl (2009). Concretely, to characterize the effect of treatment t on outcome y with *do*-calculus, we write

$$p(y|do(t)) \triangleq \sum_{\boldsymbol{x}} p(y|\boldsymbol{x}, t)p(\boldsymbol{x}), \tag{1.4}$$

where $\boldsymbol{x}$ denotes all other factors on the causal graph that points to y, with $p(\boldsymbol{x})$ as their marginal distribution. Eq. 1.4 is also known as the *G-formula*, and it takes an interventional perspective: fixing treatment at the specific value while let other variables occur in their natural frequency. This is fundamentally different from the conditional probability, which is given by

$$p(y|t) \triangleq \sum_{\boldsymbol{x}} p(y|\boldsymbol{x}, t)p(\boldsymbol{x}|t). \tag{1.5}$$

In general $p(y|do(t)) \neq p(y|t)$.

With the three assumptions listed above, the average causal effect in the population are identifiable from observational data. In a randomized controlled trial, the randomization and blocking guarantees the strong ignorability (Imbens and Rubin, 2015b). That is, Assumption 1.1 and 1.2 are automatically attained, so the simple difference of the treatment-conditioned outcome average can be directly considered as causal for *randomized controlled trials* (RCT). In observational studies, however, the treatment assignment mechanism is usually unknown and uncontrolled, necessitating additional statistical adjustments to de-bias

the causal estimates.

1.2.1 Target population and estimand

Since the causal effect can not be identified on the individual level, we must specify the target population of interest before deriving the population-level causal effect, also known as the *estimand*. Two of the most commonly used estimands are respcetively the *average treatment effect for the population* (ATE) and *average treatment effect for the treated* (ATT):

$$\tau_{ATE} = \mathbb{E}_i[Y_i(1) - Y_i(0)] \tag{1.6}$$

$$\tau_{ATT} = \mathbb{E}_i[Y_i(1) - Y_i(0)|T_i = 1]. \tag{1.7}$$

ATE considers the overall population the study sample represents, and it compares the potential outcomes between the scenarios where all subjects were treated and where all units were untreated. ATT instead evaluates the change in outcomes if those actually treated units were subjected to no treatment.

Sample representativesness is critical for statistical inference. For almost any research, the statistical conclusions should be generalizable to the population intended (usually the sampled population), and possibly to other similar population of interest as well. However, the representativeness of the sample is sometimes crippled by treatment assignment mechanism designed to control for confounding. Non-randomized studies, *e.g.* cohort studies, typically are more representative of the source population than experiments as the latter

are often conducted in constrained environments (Rubin, 1974b). Therefore, in addition to the treatment assignment mechanism, the discussion of sampling scheme with respect to the target population is equally important when seeking causal conclusions.

In recent years, there has been growing interest in more customized treatment effect estimation, which promises to enable fine-grained decision making in many critical applications. Formally known as the *conditional average treatment effect* (CATE), this causal estimand targets a population with a specific value $\boldsymbol{x}$ for their unit characteristics, *i.e.*,

$$\tau(\boldsymbol{x}) = \mathbb{E}[Y_i(1) - Y_i(0) | \boldsymbol{X}_i = \boldsymbol{x}] \tag{1.8}$$

With slight abuse of notation, $\tau(\boldsymbol{x})$ is also known as the *individualized treatment effect* (ITE), or *heterogenous treatment effect* (HTE) in the literature (Künzel et al., 2019, Shalit et al., 2017, Wager and Athey, 2018b).

1.2.2 Balancing score and deconfounding

Center to the estimation of causal effect is how to deconfound, that is to remove the correlation between the treatment and outcome due to the pre-treatment variables. Under the strong ignorability assumption, this can be achieved by adjusting for the treatment assignment mechanism (Rubin, 2001, Stuart, 2010). Intuitively, the goal of such adjustments is to create synthetic RCT based on observational data, such that the representations of treated and untreated groups are systematically balanced[1]. The solution is to find a *balancing score*,

[1]The covariates are called *balanced* when they are independent of the treatment assignment conditioned on some other variable.

conditional on which the link between pre-treatment variables and the treatment is broken.

Definition 1.4. A **balancing score**, $b(\boldsymbol{X})$, is a function of the observed covariates $\boldsymbol{X}$ such that the conditional distribution of $\boldsymbol{X}$ given $b(\boldsymbol{X})$ is the same for treated $(T = 1)$ and untreated $(T = 0)$:

$$\boldsymbol{X} \perp\!\!\!\perp T \mid b(\boldsymbol{X}) \tag{1.9}$$

Under the strong ignorability assumption, unbiased estimates of ATE can be attained by conditioning on the balancing score (Rosenbaum and Rubin, 1983a):

Theorem 1.5. *Suppose treatment assignment is strongly ignorable and $b(\boldsymbol{X})$ is a balancing score, then the expected difference in observed outcomes between treated $(T = 1)$ and untreated $(T = 0)$ groups at $b(\boldsymbol{x})$ equals to the average treatment effect at $b(\boldsymbol{x})$, that is,*

$$\mathbb{E}[Y(1)|T = 1, b(\boldsymbol{x})] - \mathbb{E}[Y(0)|T = 0, b(\boldsymbol{x})] = \mathbb{E}[Y(1) - Y(0)|b(\boldsymbol{x})] \tag{1.10}$$

The balancing score can be used in a number of ways, such as matching, weighting, and covariate adjustment, to restore the covariate balance for establishing causality. A direct application of Theorem 1.5 is matching-based adjustments. Conceptually, matching is very similar to blocking, where similar units are grouped together to make meaningful comparisons. Operational-wise, matching sets are identified based on the similarity in terms of $b(\boldsymbol{X})$, and within each matching set the data can be effectively considered as if they were coming from a small RCT. To derive causal estimate for a specific population, comparison results from many such synthetic RCTs are properly aggregated.

One straightforward choice of balancing score is the covariate itself $(b(\boldsymbol{X}) = \boldsymbol{X})$, which drives matched experiment designs like paired, case-control, and case-crossover studies. How-

ever, this choice is often considered suboptimal as matching is very sensitive to the curse of dimensionality: as the dimension of $b(\boldsymbol{x})$ or the noise level grows, it becomes progressively challenging to find close matches. As such, lower dimensional balancing scores are more popular in practice: via collapsing the the covariate into a more compact representation, we can expect improved representation overlap, lower-variance estimation, more flexible balance evaluation.

To maximally compress the representation of balancing score, we may explicitly account for the knowledge of treatment assignment, which is captured by the concept of *propensity score* (Rosenbaum and Rubin, 1983a) as defined below

Definition 1.6. Propensity score, $e(\boldsymbol{X})$, is the probability of receiving treatment ($T = 1$) given observed covariates $\boldsymbol{X}$:

$$e(\boldsymbol{x}) = Pr(T = 1|\boldsymbol{X} = \boldsymbol{x}) \tag{1.11}$$

Propensity score is by definition a balancing score, as it makes the treatment assignment independent of the covariates: units with the same propensity score are equally likely to receive the treatment, so the distribution of covariates across treatment groups are the same. The fact it is a one dimensional summary has made propensity score especially appealing in compensating representation imbalance, with applications in matching (Austin, 2009, Schwab et al., 2018a), stratification (Lunceford and Davidian, 2004), covariate adjustment (Athey and Imbens, 2017).

While playing a central role in causal inference, in observational settings, however, the propensity score is usually unknown and needs to be estimated from data. The effectiveness

of these balancing adjustments critically depends on if the accuracy of the propensity score model. Given the central role propensity score has played, research on strategies for handling extreme values, score estimation, and imbalance adjustment are the most active topics in observational causal inference.

1.3 Challenges and objectives

In this section, we want to point out some of the limitations of existing observational causal inference procedures, and highlight the new challenges arise in the wake of big data era, to motivate the our work in subsequent chapters.

A major concern for standard literature on causal inference with observational data is that they heavily rely on the unfoundedness assumption (Assumption 1.1), such that all confounders must be observed. This assumption is, however, untestable and oftentimes violated in practical settings. In general, estimating causal effects with unobserved confounding is infeasible (Pearl, 2009). Without additional assumptions, the observed data distribution can be compatible with many potentially contradictory causal explanations, that are indistinguishable based on the data in the eye of an investigator (D'Amour, 2019). This severely undermines the validity of causal estimates in empirical scenarios. An interesting question to ask is what can we do to reduce the sensitivity to the latent confounders; and more critically, is there any settings we can safely filter out the impact from them without observing the latent confounders?

Compounding the issue of compromised unfoundedness, we also recognize observed vari-

ables are possibly stochastic proxies to the true confounders (Maddala and Lahiri, 1992, Montgomery et al., 2000). For example, one can not directly measure the aggressiveness of a driver, and yet this might be indirectly assessed through alternative variables such as driving behaviors (*e.g.*, braking pattern, cruising speed, lane change, etc.), which are down stream to the unobserved confounding [2]. Caution that a common mispractice is to treat the proxies as if they are ordinary confounders, as bias can be expected (Fuller, 2009, Griliches and Hausman, 1986, Pearl, 2012, Rothman et al., 2008). While it has been established that causal identification is feasible when the proxy variables meet certain conditions (Cai and Kuroki, 2008, Kuroki and Pearl, 2014, Miao et al., 2018, Shi et al., 2018), these conditions are highly technical and the resulting solution only applies to very specific types of models. A more generic modeling approach that fully embraces the data uncertainty from the proxy observations is still missing in current literature.

In addition to the difficulties above, the explosive growth of modern data also adds new challenges to the adjustments strategies employed by conventional causal inference procedures. Notably, modern datasets often characterized by high dimensionality (Belloni et al., 2014), complex interactions (Zhang and Hyvärinen, 2009), and noisy observations. Unfortunately, classical balancing techniques like matching and weighting do not scale in such settings. In particular, matching is highly sensitive to the curse of dimensionality and observation noise, while the efficacy of weighting adjustment is compromised by extreme weights due to the lack of overlap and imbalance in high dimensions. These necessitate the integration of representation learning techniques to reduce the data to a more compact and balanced form, in order to properly assess causality.

[2] A related approach is to leverage instrumental variables to identify causal effects (Singh et al., 2019).

Switching gear, we refocus on a more fundamental obstacle in observational causal inference: the *counterfactual validation*. Since only partial observation of the potential outcomes, the one tied to the assigned treatment, have been made, direct validation of the estimated causal effect is not possible due to the missing counterfactuals. A popular practice is to base causal inference on the outcome models for each treatment arm, and then validate with the factual observations. As noted in Nie and Wager (2017) and Schuler et al. (2018), these factual validation targets do not effectively assess the counterfactual performance, resulting biased or unreliable results. This is especially concerning for working with either complex models or low-sample regimes, where over-fitting looms.

Finally, we want to take novel perspectives and explore an important new problem. In particular, we concentrate on the distribution shifts and data heterogeneity in our observational setup, and ask the question: can we build causal models that generalize well to the distributions that are different from the training data. Such scenarios arise frequently in real-world applications: (*i*) experimental designs preferentially target specific population different from the general public; and (*ii*) data collected from heterogeneous environments show distinct confounding patterns. Standard causal models are vulnerable in such settings, as they may unintentionally exploit unstable correlations specific to sub-populations that do not generalize well to others. New strategies are desired to enable better *out of distribution* generalizability for observational causal inference.

The overarching objective of this dissertation is to revisit the classical causal effect estimation approaches, to reconsider the role of uncertainty and heterogeneity in establishing a more robust and consistent causal effect model, and to explore the causal validation in the counterfactual reasoning process.

1.4 Dissertation structure

The rest of this dissertation is organized as follows. Chapter 2 and Chapter 3 focuses on the techniques of covariate balancing, from two perspectives: the original space of covariates and a latent representation space projected from the original space. Chapter 2 introduces propensity score weighting methods to balance observed covariates. Three weighting schemes corresponding to different target population are illustrated. The proposed approaches are applied in a *naturalistic driving study* (NDS) to estimate the causal impact of cellphone distraction on vehicle crash risk. Chapter 3 presents a novel generative Bayesian estimation framework for causal effect estimation that integrates representation learning and adversarial matching, termed as *Balancing Variational Neural Inference of Causal Effects* (BV-NICE). The approach is compared with state-of-art models with extensive experiments. Chapter 4 explores an innovative perspective of causal model generalizability by leveraging data heterogeneity, termed as *Risk Invariant Causal Estimation* (RICE). We demonstrate that RICE provides invariant causal estimation across heterogeneous sub-populations and yields better generalization for new samples through real and synthetic data.

Chapter 2

Deconfounding through Measured Confounders: Cohort Study using Propensity Score Method [1]

2.1 Introduction

Distraction is a major risk factor for driving safety (Dingus et al., 2016b). According to National Highway Traffic Safety Administration (NHTSA), human error was the major reason for 93% of crashes (NHTSA, 2013). The rapid increase in cellphone ownership over the last decade, especially for smartphones, makes cellphone distraction a major crash risk factor. A recent study by Insurance Institute of Highway Safety (IIHS) indicated that drivers were observed 57% more likely to be operating a cellphone in 2018 than in 2014 (IIHS, 2019). In

[1] This chapter is a published paper: Lu, Danni, Feng Guo, and Fan Li. "Evaluating the causal effects of cellphone distraction on crash risk using propensity score methods." *Accident Analysis & Prevention* 143 (2020): 105579.

addition, the complex functionality of smartphones imposes higher demand on both cognitive and visual-manual load for drivers. To accurately estimate the safety risk associated with cellphone use is critical for the development of safety improvement measures. Various methods and data sources have been used to investigate cellphone use behavior during driving, including surveys, experimental studies, etc. A well-designed survey can provide information on the pattern and characteristics of drivers' cellphone use habit, *e.g.*, percentage of time talking on a cellphone while driving and cellphone use pattern by demographic characteristics (Glassbrenner and Ye, 2007, McCartt and Geary, 2004) . Experiment studies on driving simulators and instrumented vehicles in controlled environments in general indicated deteriorated driving performance when carrying out designed tasks on cellphone (Leung et al., 2012, Li et al., 2016). However, crashes are usually not observed in such experiment settings, making it difficult to gauge the actual crash risk.

Crash databases, such as hospital records, the General Estimation System (GES), and the Fatality Analysis Reporting System (FARS), are major sources for crash risk evaluation. Using a case-crossover approach based on cellphone and hospital records, studies in Cantata and Australia showed that cellphone use was associated with a fourfold increase in crash risk (McEvoy et al., 2007, Redelmeier and Tibshirani, 1997a). Study using GES database showed that cellphone distraction resulted in significant higher risk (OR=11.56) of rear-end collision for teenage drivers (Neyens and Boyle, 2007). The crash database typically lacks detailed information on cellphone subtasks and precise time sequence of events.

The large scale *naturalistic driving study* (NDS) method provides high resolution, high frequency, and objectively collected real-world information on driver behavior, driving environment, and safety outcomes (Hankey et al., 2016). The NDS continuously records driving

data over an extended period of time in real-life driving conditions thorough advanced data acquisition system installed on participants' vehicle without intereferencing normal driving. Typical instrumentation includes multiple cameras, GPS, multidimensional accelerometers, and other devices in an integrated system. NDS provides valuable real-life information on driving behavior, crashes, and safety critical events, making it possible to evaluate the factors that increase driving risk and the causation of crashes (Guo, 2019a). The large scale NDS, such as the Second Strategic Highway Research Program (SHRP2) NDS, observed thousands of crashes for crash risk evaluation (Hankey et al., 2016).

Large scale NDS data have provided valuable insight into the impact of cellphone use on driving risk (Atwood et al., 2018, Farmer et al., 2015, Fitch et al., 2013, Klauer et al., 2014a). The results in general confirmed an increased risk associated with cellphone use while driving, *e.g.*, Dingus et al. (2016b) showed an *odds-ratio* (OR) of 3.6 associated with overall cellphone distraction. The continuous recorded videos and driving data also allow the evaluation of specific cellphone related tasks,*e.g.*, texting and talking. Majority of these studies indicated that cellphone related visual-manual tasks will increase crash risk considerably, such as texting, dialing, and reaching for cellphone, while with mixed results on cellphone talking (Atwood et al., 2018, Dingus et al., 2016b, Guo et al., 2017a).

The impact of cellphone distraction on crash risk, however, is potentially affected by other factors,*e.g.*,driving experience, age, and driver's intention to engage under different conditions. Klauer et al. (2014a) showed that the risk of crash and near crash doubled for novice drivers compared with experienced drivers. Evidences from observational surveys and driving simulator studies showed substantial difference in driving performance under cellphone distraction across age groups (Greenberg et al., 2007, McCartt and Geary, 2004) .

Using naturalistic driving data, Guo et al. (2017b) confirmed that both the prevalence and relative risk of distraction vary substantially by age group. Young drivers under the age of 30 years old and senior drivers older than 65 were at higher risk when using cellphone. Cellphone distraction risks can potentially be affected by complex driving environment factors. For instance, drivers' willingness to engage in cellphone use could be affected by driving speed, traffic flow, road type, lighting, and relationship to junction (Saito et al., 2011). At the same time, those factors can also contribute to crash risk. For example, high traffic flow density has been shown to be associated with higher crash risk and drivers tend to avoid cellphone use under high traffic flow conditions (Meng and Weng, 2011). Drivers' willingness to engage in cellphone use under various condition with different risk levels will potentially bias the estimation of causal effects and it is important to identify and adjust for those confounders (Cochran, 1983, Imbens and Rubin, 2015a).

Limited research has evaluated safety effect under formal causal inference framework although learning causal effects is one of the primary foci of traffic safety research. To evaluate the safety performance of road infrastructure, propensity score matching has been used to improve the balance between comparison groups (Karwa et al., 2011, Sasidharan and Donnell, 2013, Wood et al., 2015). Flannagan et al. (2019) and Karwa et al. (2011) use the propensity weighting method to evaluate the safety impacts of driver behavior on crash risk using NDS data. Wood and Donnell (2017) discussed generalizable safety effect estimates according to different target population. Apart from this stream of work, there is in general a lack of in-depth discussion on the assumptions and causal estimands behind each method and proper interpretation of causal effects in the context of traffic safety research. In particular, an important but less well-understood point is that in the

presence of treatment effect heterogeneity, different propensity score methods usually lead to causal effects on different target populations, and consequently, potentially different policy implications. The choice of propensity score method should be based on the specific research objective a priori and needs to be clearly communicated.

This chapter aims to systematically examine several common propensity score weighting schemes and elucidate their corresponding target population and causal effects for safety analysis based on NDS data. We implement three alternative methods on the SHRP2 NDS data and estimate the causal effects of cellphone distraction on crash risk. The rest of the paper is organized as follows. In Section 2.2, we introduce the SHRP2 NDS data and the case-cohort sampling design. In Section 2.3, we present three propensity score weights and models. In Section 2.4, we apply three propensity-score-weighting methods to the SHRP2 NDS data to evaluate the causal impact of cellphone distraction on crash risk. Discussions and conclusions are presented in Section 2.5.

2.2 SHRP2 NDS cohort data

The SHRP2 NDS includes more than 3,400 participating drivers from six sites across United States: Seattle (Washington), Tampa (Florida), Buffalo (New York), Durham (North Carolina), State College (Pennsylvania), and Bloomington (Indiana) (Hankey et al., 2016) . Participants' vehicles were instrumented with an advanced data acquisition system including four cameras, three dimensional accelerometers, GPS, forward radar, and alcohol sensor, *etc.* Driving data for each participant were continuously recorded from ignition-on to

ignition-off for up to two years. The study collected over one million hours of continuous driving data with 35 million vehicle-miles traveled (Dingus et al., 2016b).

A systematic approach was adopted to identify crashes by screening potential crashes by abnormal kinematic signatures followed by manual verification through the recorded videos. In total, 1,830 crashes were identified and were classified into 4 severity levels, from the most severe level-1 crashes to level-4 low risk tire strike crashes (Hankey et al., 2016). Since level-4 crashes were considered as minor crashes with low risk, we excluded them from the study and focused on 1,047 relatively severe crashes (level- 1, 2 or 3 crashes).

A case-cohort design was used to randomly sample control driving segments for risk evaluation (Guo, 2019a). The control driving segments are short, 6-second driving segments that represent driver behavior and traffic environments under normal driving conditions. Guo (2019a) showed that the OR under this control sampling scheme is an approximation to the crash rate ratio. A total of 19,798 control driving segments were randomly sampled from more than one million hours of continuous driving videos. The number of control samples for a driver is proportional to his/her total driving hours. A minimum of one control was included for each driver.

Driver behavior and environmental factors for both crashes and controls were extracted through the examination of videos by trained data reductionists following a rigorous data reduction protocol (Hankey et al., 2016). The reduced data include three types of variables: event variables, driver variables and environmental variables. Driver variables describe driver behavior such as impairments, secondary task presence, and visual obstructions. Environment variables include roadway characteristics, traffic flow, locality, lighting, weather

conditions, and *etc.* Besides information reduced from videos, the study also conducted a questionnaire survey on driver's demographic information.

2.2.1 Driver distractions and crashes

A driver is considered distracted if he/she engages in non-driving related secondary tasks. The driver's distraction status prior to a crash or during a control segment, including cellphone use and other non-driving related secondary task engagement is extracted by examining the recorded video. A driver's status can be undistracted, engaging in cellphone tasks, or engaging in other secondary tasks. Risk assessment depends on the reference level. Two widely used reference levels are model-driving and all-driving reference level: the model-driving refers to sober, alert, and attentive driving and the all-driving refers to not engaging in the distraction of interest but could be engaging in other distraction (Dingus et al., 2016b, Flannagan et al., 2019, Guo, 2019a, Owens et al., 2018). We adopt the model-driving as the reference level as it provides a stable and constant level for comparison (Guo, 2019a). After removing events with non-cellphone related distraction, 8,116 crashes and controls from 2,320 drivers were included in the analysis. The distribution of cellphone distraction and safety outcome by age group are shown in Table 2.1.

Table 2.1: Driver distraction and safety outcome sample size by age group

Age Group	Teenage (16-19 years old)			Young Adult (20-29 years old)			Middle Age (30-64 years old)		
Safety Outcome	Crash	Control	Total	Crash	Control	Total	Crash	Control	Total
Distractions									
cellphone overall	40	212	252	65	661	726	15	313	328
talking	9	75	84	22	323	345	6	187	193
visual-manual	31	141	172	49	346	395	9	128	137
Model driving	59	1,086	1,145	64	2,588	2652	60	2,953	3,013
Total Events		1,397			3,378			3,341	
Total Drivers		477			932			911	

Table 2.2 shows potential confounding factors for cellphone use including driver demographics, time of day, driving environment. Cellphone use while driving varies by many factors,*e.g.*, female drivers (14%) more than male drivers (11%); young adult drivers (21%) more than middle age drivers (10%); and drivers in the lower income group (15%) more than high income group (10%). In addition, drivers' willingness-to-engage in cellphone use depends on driving environment and context. For example, drivers tend to use cellphone more during dark lighting condition (16%) and in unstable traffic flow condition (16%).

The demographic characteristics and environmental factors not only affect cellphone use pattern but also could potentially contribute to driving risk (Guo et al., 2017b, Meng and Weng, 2011). As a result, the relationship among those factors could lead to confounding effects that will skew the causal effects of cellphone use on crash.

Table 2.2: The percentage of distractions by environmental factors

		Cellphone overall			Talking			Visual-manual		
Factor	Value	Perc.	Freq.	Total	Perc.	Freq.	Total	Perc.	Freq.	Total
Traffic	Free Flow	12%*	862	7,205	6%	403	6,731	7%	468	6,793
	Stable Flow	13%	453	3,407	7%	228	3,171	7%	228	3,174
	Unstable Flow	16%	35	222	8%	17	204	8%	17	207
Locality	Entrance/Parking	13%	166	1,327	7%	89	1,247	7%	81	1,242
	Intersection	13%	406	3,224	6%	195	3,008	7%	218	3,038
	Non-junction	12%	675	5,444	6%	318	5,070	7%	355	5,100
	Other	12%	103	839	6%	46	781	7%	59	794
Lighting	Darkness	16%	312	1,937	8%	145	1,769	9%	170	1,796
	Dawn or Dusk	14%	60	427	7%	26	387	8%	32	397
	Daylight	12%	978	8,470	6%	477	7,950	6%	511	7,981
Weather	Normal	13%	1231	9,711	7%	598	9,047	7%	642	9,088
	Rain or Fog	11%	112	1,060	5%	49	1,002	6%	65	1,020
	Snow	11%	7	63	2%	1	57	9%	6	66
Surface	Dry	13%	1145	8,969	7%	571	8,363	7%	584	8,369
	Wet	11%	205	1,865	4%	77	1,743	7%	129	1,805
Separation	Divided	13%	608	4,667	7%	294	4,340	7%	318	4,375
	Not Divided	12%	742	6,167	6%	354	5,766	7%	395	5,799
Intersection	No	12%	1042	8,408	7%	517	7,856	7%	535	7,878
	Yes	13%	308	2,426	6%	131	2,250	8%	178	2,296
Construction	No	12%	1289	10,385	6%	617	9,686	7%	684	9,765
	Yes	14%	61	449	7%	31	420	7%	29	409
Road Type	Highway	12%	408	3,398	6%	205	3,186	6%	206	3,193
	Local	13%	942	7,436	6%	443	6,920	7%	507	6,981
Income	Low	15%	435	2,956	8%	210	2,725	8%	227	2,755
	Middle	12%	603	4,912	6%	284	4,583	7%	327	4,622
	Other	11%	62	579	6%	35	547	5%	28	525
	High	10%	250	2,387	5%	119	2,251	6%	131	2,272
Gender	Female	14%	788	5,492	8%	408	5,097	8%	390	5,090
	Male	11%	562	5,342	5%	240	5,009	6%	323	5,084
Age Group	Teenage	18%	252	1,397	7%	83	1,228	13%	169	1,301
	Young Adult	21%	726	3,378	12%	345	2,993	13%	391	3,036
	Middle age	10%	328	3,341	6%	191	3,190	4%	140	3,144
	Senior	2%	44	2,718	1%	29	2,695	0%	13	2,693

2.2.2 Potential confounders

The selection of confounders, or more generally the covariates included in the propensity score and outcome model is crucial. In total, we investigated 39 potential confounders. An initial screening of the variables was conducted by combining both the opinion of domain experts and statistical variable selection using Elastic Net regulation method. Table 2.3 lists the 12 selected potential confounders for detailed analysis.

Since the listed covariates are all categorical, we use Cramer's V to measure the strength of association (Cramér, 1999). The Cramer's V ranges from 0 to 1 with higher value represents stronger association. Chi-square test was conducted to exam the association among variables. As shown in Table 2.3 the strength of association between covariates and cellphone use/crash varies.

Table 2.3: Associations and Chi-square tests

Covariate	Cellphone distraction		Crash		Include in PS model
	Cramer's V	p value	Cramer's V	p value	
Traffic density	0.03	0.10	0.06	<0.01	Yes
Locality	0.01	0.97	0.08	<0.01	Yes
Lighting	0.05	0.00	0.04	<0.01	Yes
Age group	0.24	0.00	0.08	<0.01	Yes
Weather	0.02	0.08	0.08	<0.01	Yes
Surface condition	0.02	0.02	0.06	<0.01	Yes
Separation	0.01	0.14	0.07	<0.01	Yes
Intersection	0.00	0.74	0.17	<0.01	Yes
Construction	0.01	0.47	0.02	0.03	Yes
Income	0.07	0.00	0.04	0.04	Yes
Road Type	0.01	0.36	0.07	<0.01	Yes
Gender	0.06	0.00	0.01	0.60	No

There are multiple alternative inclusion criteria for covariates in a propensity score model (Westreich et al., 2011). A widely adopted guideline advises the inclusion of all predictors of outcome regardless of their relationship with treatment (Brookhart et al., 2006, Westreich et al., 2011). This approach would reduce the variance of the treatment effect estimate compared to only including variables that are predictive of both outcome and exposure. In this study we include all covariates with statistically significant association with crashes (p-value smaller than 0.05) in the propensity score model as indicated in Table 2.3.

2.3 Propensity score method

Confounding is a major source of biased estimation of causal effects between a risk factor and disease outcomes (Hernán and Robins, 2020, Imbens and Rubin, 2015a). For example, age is highly correlated with both cellphone use and crash. As a result, the difference in crash rate between cellphone-distracted group and model-driving group may be simply due to different percentage of teenage drivers in the distracted group. To balance the distribution of confounders between the exposed and unexposed groups is essential for estimating causal effects (Rosenbaum and Rubin, 1983a). In randomized experiments, confounder balance is automatically attained through randomization (Rubin, 1990). For observational studies such as NDS, exposure such as cellphone use is decided by participants, and significant covariate imbalance between the exposed and un-exposed groups is the norm than exception. Imbalance in potential confounders should be adjusted using proper statistical methods for valid causal inference. Following the nomenclature in causal inference, we use the terms exposure and treatment interchangeably hereafter.

The propensity score plays a central role in covariate balancing and confounder adjustment (Rosenbaum and Rubin, 1983a). The propensity score is the probability of one unit being exposed (or receiving the treatment) conditional on observed confounders. It is a summary score of the multi-dimensional covariates. Propensity score can be used in various ways, including weighting, matching and stratification (Austin, 2009, Lunceford and Davidian, 2004). It is important to note that different propensity score methods lead to causal effects on different target population and thus consequently may have different policy implications. In practice, propensity scores are usually unknown in observational settings and need to be estimated first. Therefore, the propensity-score-weighting-based causal analyses typically involve three main steps: (1) estimate propensity scores based on selected covariates,*e.g.*, through a random-effects logistic regression; (2) evaluate the covariate balance given the selected propensity score method; (3) estimate the treatment effect based on the estimated propensity scores, *e.g.*, through weighted regression. The above Steps (1) and (2) are conducted in an iterative fashion, prior to Step (3), to ensure the propensity scores are well estimated. In the context of this paper, the treatment refers to the volitional cellphone use exposure by drivers.

2.3.1 Propensity score weights

Consider the collected data as a sample from I drivers. Let T_{ij} be a dichotomous variable indicating cellphone use exposure status of $j^{th}(j = 1, \cdots, N_i)$ event (either a crash or a control) of driver $i(i = 1, \cdots, I)$:

$$T_{ij} = \begin{cases} 1, & \text{cellphone use} \\ 0, & \text{model-driving} \end{cases} \tag{2.1}$$

Let $\boldsymbol{X}_{ij}$ denotes the selected covariates. We assume that the positivity and unconfoundedness assumptions are satisfied (Hernán and Robins, 2020, Imbens and Rubin, 2015a, Rosenbaum and Rubin, 1983a). The positivity assumption, also known as the overlap assumption, requires in each event the driver has a non-zero probability of being either exposed or in the reference group. Unconfoundedness, also known as exchangeability or ignorability, assumes there is no unmeasured confounders other than the observed covariates.

The propensity score measures cellphone use tendency under the context described by the observed covariates. The propensity score for the k^{th} event of driver i, e_{ij}, is the conditional probability of engaging cellphone use given corresponding covariates $\boldsymbol{X}_{ij}$, *i.e.*,

$$e_{ij} = e(\boldsymbol{X}_{ij}) = Pr(T_{ij} = 1 \mid \boldsymbol{X}_{ij}) \tag{2.2}$$

A high propensity score indicates drivers have a high probability using cellphone and vice versa. Rosenbaum and Rubin (1983a) showed that propensity score has two theoretical properties underpinning its central role in causal inference. First, balancing the scalar propensity score leads to the balance in the multi-dimensional covariates. Second, if the treatment/exposure assignment is unconfounded conditional on a set of covariates, it is also unconfounded conditional on the propensity score. Therefore, the confounding effect can be addressed by adjusting the one-dimensional propensity score instead of the set of potentially high-dimensional covariates.

In this section, we focus on the propensity score weighting approach for its clear definition of target population. The key idea of weighting is that each event is re-weighted by a function of its propensity score, denoted as w_{ij}, to create a target population where the exposed and reference groups are balanced in the distribution of the measured confounders. Then the weighted difference in the outcomes between the exposed and unexposed groups is interpreted as the causal effect on the corresponding target population. There is a general class of balancing weights (Li et al., 2017b) , each corresponds to a specific target population and causal estimand.

The choice of weights, and equivalently target population and causal estimand, is crucial in applications; statistical, scientific and policy considerations all may come into play. In this paper we compare three commonly used balancing weights: the *inverse probability weight* (IPW), the ATT weights, and the *overlap weights* (OW), which correspond to three different target population and consequently estimands: the average treatment effect on the whole population (ATE), the average treatment effect for the treated population (ATT), and the average treatment effect for the overlapped population (ATO), respectively. The IPW, corresponding to the ATE estimand, is defined as:

$$w_{ij}^{\text{ATE}} = \begin{cases} \frac{1}{e_{ij}}, & \text{if } T_{ij} = 1 \\ \frac{1}{1-e_{ij}}, & \text{if } T_{ij} = 0 \end{cases} \tag{2.3}$$

where e_{ij} is the propensity score for the k^{th} event of driver h. The target population of ATE is the combined exposed and reference population represented by the study cohort. In the context of this paper, the population is the entire SHRP2 NDS case-cohort sample

events, including both crashes and 6-second baseline driving segments. By using a weight reciprocal to each event's assignment probability to its observed group, IPW up-weights the under-represented events and down-weights the over-represented events. For example, if a driver who has a low probability of using cellphone but indeed used cellphone (*i.e.*, in the exposed group), then the IPW scheme up-weighs this subject to represent more subjects.

The ATT weight, corresponding to the ATT estimand, is defined as:

$$w_{ij}^{\text{ATT}} = \begin{cases} 1, & \text{if } T_{ij} = 1 \\ \frac{e_{ij}}{1-e_{ij}}, & \text{if } T_{ij} = 0 \end{cases} \tag{2.4}$$

The target population of ATT is the exposed group, which is the subpopulation of all cellphone-use subjects in our application. Under this weighting scheme, reference events are weighted to resemble the exposed population. ATT answers the question "what is the effect of the exposure on the exposed events, compared with the scenario if all the exposed events would have been un-exposed". The ATT estimand is similar to the estimand in the common propensity score matching methods where each exposed unit is matched to a set of control units (Rosenbaum, 2002).

The overlap weights (OW) (Li et al., 2017b), corresponding to the ATO estimand, is defined as:

$$w_{ij}^{\text{ATO}} = \begin{cases} 1 - e_{ij}, & \text{if } T_{ij} = 1 \\ e_{ij}, & \text{if } T_{ij} = 0 \end{cases} \tag{2.5}$$

OW weights each event by its probability of being assigned to the opposite group. OW smoothly up-weights the events in the center and down-weights the events in the tails of

the propensity score distribution. Specifically, events with propensity scores close to 0.5, *i.e.*, equally likely to be either exposed or non-exposed, have the largest contribution. The target population of OW are the events with the most overlap in observed characteristics between the exposed and non-exposed groups. This population resembles the common target population in randomized trials, referred to as the events in "clinical equipoise" in the context of clinical trials. In our application, they are the events who are "on the fence" in terms of their cellphone-use behavior, and thus would most likely shift as new information about the effect of the exposure is obtained.

Both IPW and ATT weights use reciprocals of propensity scores, and thus are sensitive to extreme scores that are close to 0 and 1, which can lead to excessive variance and large bias. A common remedy is to trim the events with propensity scores exceeding a threshold (Crump et al., 2009, Stürmer et al., 2010). However, such methods are often sensitive to the choice of cutoff points and can discard a large proportion of the sample events (**?**). In contrast, OW are bounded between 0 and 1, thus automatically overcome the large uncertainty issue caused by extreme propensity scores, and also avoid the arbitrary choice of a cutoff point. Moreover, **?** theoretically proved that OW based on the propensity score estimated from a logistic model leads to exact balance between exposure groups for all the covariates in that logistic model.

Due to different target population and interpretation, the selection of a particular weighting scheme,or estimand, depends on the objective of the analysis. Because there might be diverse objectives in a particular driving safety analysis, we implemented and compared all three estimands. In randomized experiments, as propensity scores are constant across units, estimates of the three estimands are expected to be the same, which is usually not

the case in observational studies. Indeed, with the same observational data, estimates of the three estimands can be vastly different, which would indicate significant heterogeneity across events.

2.3.2 Estimate the propensity score and balance check

The propensity scores are unknown and must be estimated from the observed data, denoted by $\hat{e}$. Because there are multiple events from the same participant, we adopt the random-effects logistic regression model to accommodate the potential correlation of observations within the same driver:

$$\begin{aligned} T_{ij} &\sim Bernoulli(e_{ij}) \\ Logit(e_{ij}) &= \boldsymbol{X}_{ij}\boldsymbol{\beta} + \delta_i \end{aligned} \tag{2.6}$$

where $\boldsymbol{X}_{ij}$ is the covariates matrix for j^{th} event of driver h, $\boldsymbol{\beta}$ is a vector of the coefficients of the fixed effects, and $\delta_i \sim N(0, \sigma_\delta^2)$ is the driver-specific random effect. Due to the different sampling scheme of cases and controls, the estimation of propensity score is not straight forward. As crashes are rare outcomes, the covariate distribution of control samples approximates the entire population. The propensity score estimated from control samples approaches the propensity score of the population (Månsson et al., 2007). Therefore, we fit the above model to the control samples to estimate the propensity score.

Covariate balance should be checked for propensity score analysis, for two reasons. First, the goodness-of-fit of a propensity score model is determined by how well the estimated propensity scores balances the covariates. Second, balancing all potential confounders is

necessary to ensure the confounding effects are removed so that the difference in outcomes between the exposed and unexposed groups can be interpreted as causal effects. In our context, balanced cellphone use distribution for each covariate level is a precondition to make valid causal inference. A common statistic for checking covariate balance is the *standardized mean difference* (SMD) between the exposed group and reference group for each covariate. For a continuous covariate, the SMD is defined as:

$$d = \frac{\bar{x}_{exp} - \bar{x}_{ref}}{\sqrt{0.5 \times (s^2_{exp} + s^2_{ref})}} \tag{2.7}$$

where $\bar{x}_{exp}$ and $\bar{x}_{ref}$ are the sample mean of exposed group and reference group with respect to covariate x. s^2_{exp} and s^2_{ref} are sample variance of the covariate x for exposed and reference group, respectively. For a categorical variable, the SMD is defined for each categorical level. The SMD for a specific level is defined as:

$$d = \frac{\bar{p}_{exp} - \bar{p}_{ref}}{\sqrt{0.5 \times (\bar{p}_{exp}(1 - \bar{p}_{exp}) + \bar{p}_{ref}(1 - \bar{p}_{ref}))}} \tag{2.8}$$

where $\bar{p}_{exp}$ and $\bar{p}_{ref}$ are the percentage of observations with the specific level of the covariate for exposed and reference groups respectively. For example, to evaluate SMD for daytime, the $\bar{p}_{exp}$ is the percentage of events under daytime condition among exposed group (i.e. use cellphone while driving); and $\bar{p}_{ref}$ is the percentage of events under daytime condition among model-driving events. A positive SMD indicates that drivers are more prone to use cellphone at this covariate level.

Note that the sample mean and sample variance in Eq. (2.7-2.8) vary according to different weighting schemes. The three weighting methods as introduced in Section 3.3, the

ATE, ATT and ATO, will assign weight to each observation according to the corresponding target population (Eq. (2.3-2.5)). The different weight for each observation will lead to different sample mean and variance,*e.g.*, $\bar{x}_{exp}$ and $\bar{p}_{exp}$. Therefore, weighting method-specific standard difference should also be examined.

A smaller $|d|$ indicates better balance and the standard threshold for significant imbalance is 0.1 (Austin, 2009). In case the standard difference of a covariate is larger than the threshold value. The propensity score model as in Eq. (2.6) should be improved, for example, by adding quadratic terms or interaction terms. When balanced cannot be achieved by the model, stratified analysis or adding it as a predictor in the outcome model can also be considered.

We use the average of absolute value of SMD and maximum of the absolute value of SMD to assess the overall balance status for all covariates. High maximum SMD value indicates that at least one individual covariate is not unbalanced well.

2.3.3 Estimating treatment effect: causal crash OR

We considered three types of treatment effect: ATE, ATT, and ATO. The treatment effect is estimated by weighted outcome regression. To accommodate the clustering structure of the data, *i.e.*, multiple observation from the same driver, a random-effects logistic regression outcome model is adopted. The probability of being a crash for k^{th} event of driver i, μ_{ij}, is modelled as:

$$\begin{aligned} Y_{ij} &\sim Bernoulli(\mu_{ij}) \\ Logit(\mu_{ij}) &= \beta_0 + T_{ij}\gamma + \eta_i \end{aligned} \tag{2.9}$$

where Y_{ij} is an indicator variable: $Y_{ij} = 1$ for crashes and $Y_{ij} = 0$ for controls; β_0 is the intercept; $\eta_i \sim N(0, \sigma_\eta^2)$ is the driver-specific random effect that incorporates correlation among observations from the same driver; T_{ij} is the exposure status: $T_{ij} = 1$ for cellphone use and $T_{ij} = 0$ for reference level; γ is the regression coefficient of the exposure.

We apply the three different weights (Eq. (2.3-2.5)) to the random-effect logistic regression model (Eq. (2.9)). The models are fitted in SAS 9.4 platform using the GLIMMIX procedure, where the estimated propensity scores are treated as survey observation weights. The treatment effect, in the form of odds ratio (OR), is calculated by taking exponential of the estimated regression coefficient, $e^{\hat{\gamma}}$.

2.4 Case study: the causal effect of cellphone distraction using SHRP2 NDS cohort sample

We apply the above propensity score weighting methods to the SHRP2 NDS database to evaluate the causal effect of cellphone engagement on crash risk. Three cellphone related secondary tasks are considered: overall cellphone distraction, cellphone visual-manual, and cellphone talking. For each task, three treatment effects, ATE, ATT, and ATO are estimated. To avoid extreme weights, 5% of the observations with propensity scores at upper and lower tails for ATE and 5% at right tail for ATT are trimmed. ATO is robust to extreme weights and no trimming is needed.

2.4.1 Assessing the positivity and covariate balance

The covariates distributions of exposed group and reference group are compared. Figure 2.1 displays the histograms of the estimated propensity scores for overall cellphone use events and model-driving events. The estimated propensity scores show a relative wide common range (0 ~ 0.84), implying that the positivity assumption is reasonable. The substantial difference in densities of the propensity scores between two groups suggests that the two groups of events differ significantly in their baseline characteristics, pointing to the need to adjust for the imbalance in these covariates.

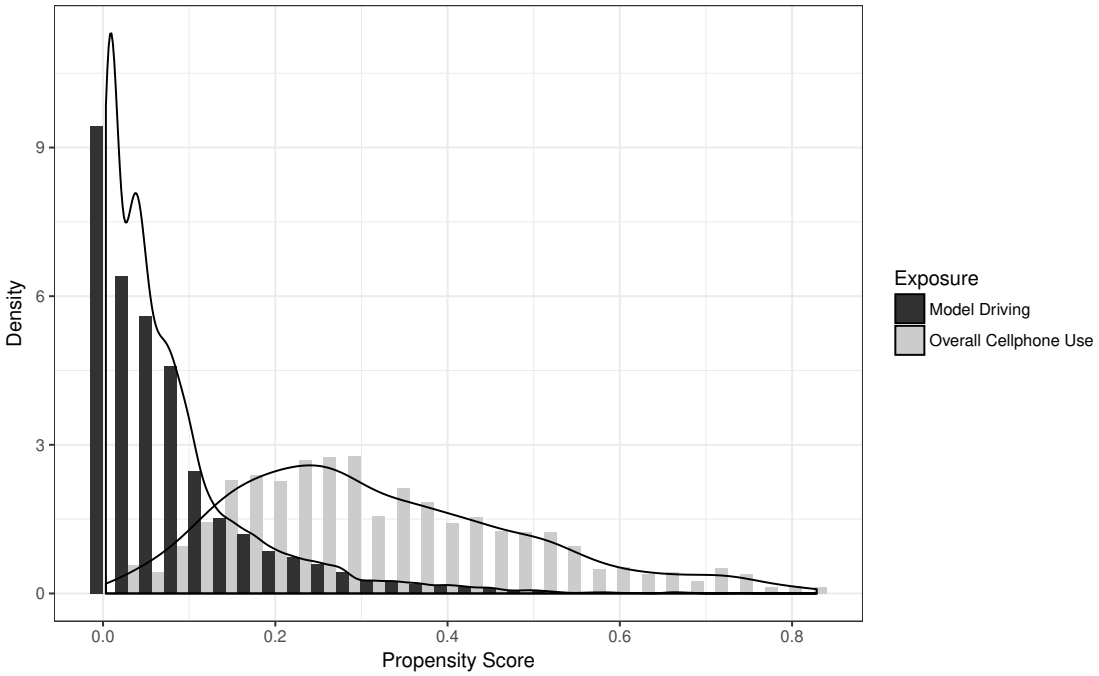

Figure 2.1: The distribution of propensity score by exposure

The SMDs (Eq. (2.7-2.8)) for cellphone use before and after propensity score weighting are visualized using the Love plot in Figure 2.2 (Ahmed et al., 2006). Positive SMD indicates

that the corresponding covariate level is overrepresented in cellphone use group compared with model-driving group. Smaller absolute value indicates a better balance. As can be seen, the cellphone exposure varies by covariates for the unadjusted case-cohort data. Low income and young drivers tend to use cellphone more often while driving. In addition, the cellphone engagement depends on driving environment,*e.g.*, drivers tend to use cellphone more during the night compared to daytime.

The three propensity score methods consistently improve covariate balance compared to unadjusted samples. Majority of adjusted SMDs fall within the 0.1 threshold, which indicates satisfactory balance has been achieved. As age is highly unbalanced and the SMD is larger than 0.1 for ATE, a stratified treatment estimation is conducted with repsct to the age group.

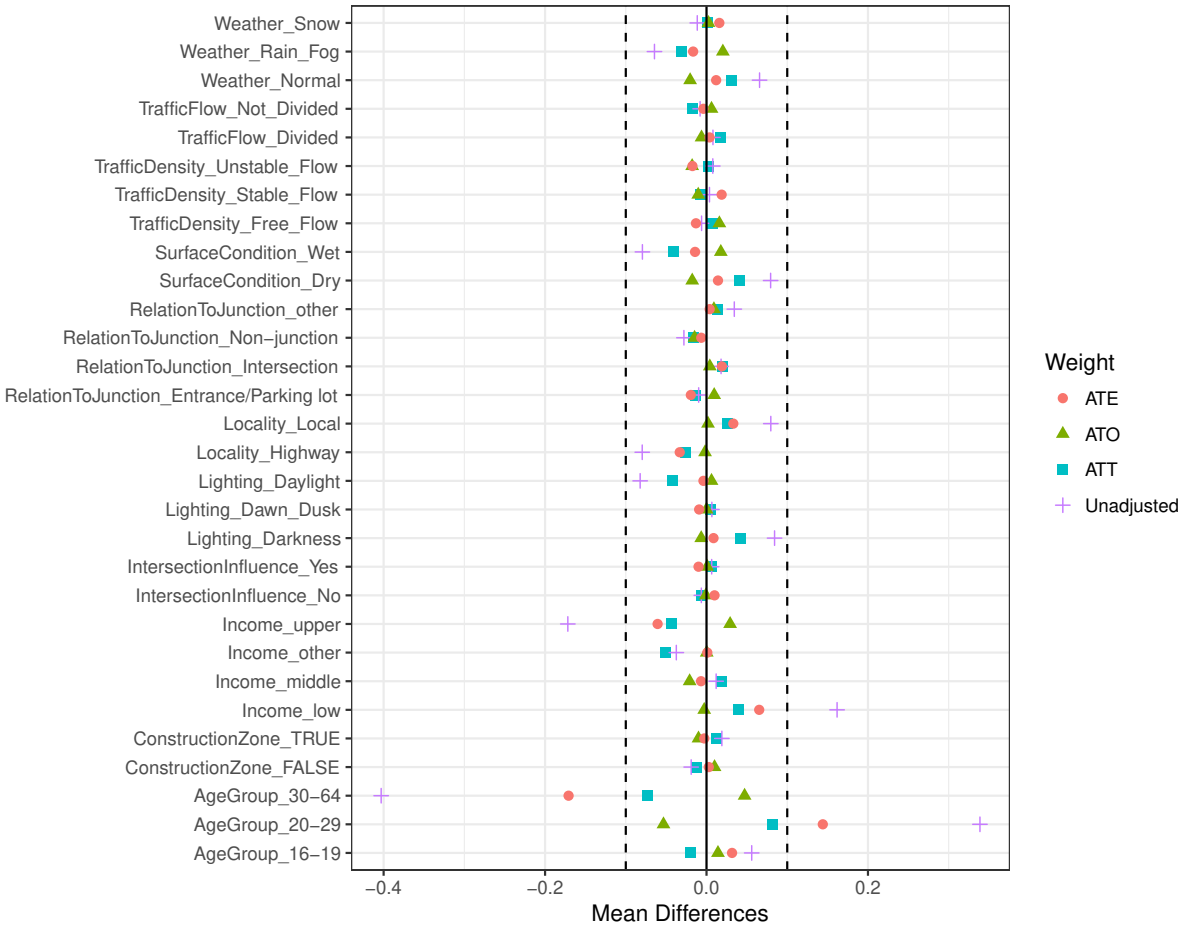

Figure 2.2: SMDs for cellphone use

2.4.2 The crash OR associated with cellphone distraction

The causal ORs by propensity score methods for the overall cellphone use, cellphone visual-manual and cellphone talking are presented in Table 2.4. The outputs include both overall ORs and ORs stratified by age group. Three types of ORs corresponding to ATE, ATT, and ATO as well as the unweighted ORs are reported for comparison.

Table 2.4: Crash OR of cellphone engagement vs model-driving

	Age Group	Weight	OR	Percentage Difference(%)	95% CI	Ave. |SMD|	Max. |SMD|
Overall cellphone use	Overall	Unweighted	4.02		(3.09, 5.25)	0.07	0.22
		ATE	4.12	2.49	(3.15, 5.40)	0.04	0.13
		ATT	4.21	4.73	(3.24, 5.48)	0.03	0.1
		ATO	3.75	-6.72	(2.81, 5.02)	0.01	0.03
	Teenage (16-19)	Unweighted	3.73		(2.33, 5.97)	0.13	0.27
		ATE	4.67	25.2	(2.81, 7.75)	0.05	0.11
		ATT	4.3	15.28	(2.63, 7.03)	0.07	0.15
		ATO	4.09	9.65	(2.51, 6.69)	0.01	0.03
	Young Adult (20-29)	Unweighted	4.21		(2.83, 6.27)	0.06	0.21
		ATE	3.87	-8.08	(2.62, 5.72)	0.03	0.08
		ATT	4.21	0	(2.83, 6.25)	0.03	0.07
		ATO	4.23	0.48	(2.87, 6.24)	0.01	0.01
	Middle-Age (30-64)	Unweighted	2.27		(1.05, 4.91)	0.07	0.2
		ATE	1.76	-22.47	(0.90, 3.45)	0.03	0.11
		ATT	1.83	-19.38	(0.86, 3.90)	0.03	0.12
		ATO	2.22	-2.2	(1.03, 4.76)	0.02	0.03
Visual-manual	Overall	Unweighted	6.2		(4.50, 8.54)	0.05	0.19
		ATE	6.19	-0.16	(4.47, 8.57)	0.03	0.08
		ATT	6.63	6.94	(4.72, 9.30)	0.04	0.14
		ATO	5.94	-4.19	(3.86, 9.15)	0.02	0.05
	Teenage (16-19)	Unweighted	4.09		(2.55, 6.53)	0.13	0.12
		ATE	5.25	28.36	(2.99, 9.23)	0.05	0.19
		ATT	4.63	13.2	(2.70, 7.94)	0.06	0.12
		ATO	4.58	11.98	(2.60, 8.06)	0.02	0.05
	Young Adult (20-29)	Unweighted	6.61		(4.14, 10.54)	0.06	0.25
		ATE	5.58	-15.58	(3.51, 8.88)	0.03	0.08
		ATT	6.03	-8.77	(3.82, 9.52)	0.03	0.09
		ATO	5.96	-9.83	(3.82, 9.31)	0.01	0.02
	Middle-Age (30-64)	Unweighted	3.47		(1.29, 9.32)	0.08	0.22
		ATE	3.54	2.02	(1.32, 9.53)	0.07	0.17
		ATT	4.29	23.63	(1.62, 11.35)	0.07	0.24
		ATO	3.59	3.46	(1.22, 10.53)	0.04	0.13
Talking	Overall	Unweighted	2.76		(1.64, 4.63)	0.08	0.29
		ATE	2.71	-1.81	(1.75, 4.20)	0.06	0.2
		ATT	2.62	-5.07	(1.58, 4.32)	0.09	0.27
		ATO	4.15	50.36	(2.12, 8.10)	0.03	0.06
	Teenage (16-19)	Unweighted	2.46		(1.07, 5.66)	0.13	0.34
		ATE	2.82	14.63	(1.05, 7.56)	0.1	0.24
		ATT	2.71	10.16	(1.04, 7.07)	0.09	0.25
		ATO	3.81	54.88	(1.14, 12.75)	0.07	0.14
	Young Adult (20-29)	Unweighted	3.22		(1.54, 6.75)	0.08	0.18
		ATE	2.62	-18.63	(1.42, 4.83)	0.04	0.13
		ATT	3.14	-2.48	(1.71, 5.75)	0.05	0.1
		ATO	3.36	4.35	(1.68, 6.71)	0.01	0.04
	Middle-Age (30-64)	Unweighted	1.13		(0.32, 4.01)	0.09	0.28
		ATE	0.65	-42.48	(0.19, 2.21)	0.05	0.15
		ATT	0.63	-44.25	(0.10, 3.95)	0.06	0.14
		ATO	1.57	38.94	(0.43, 5.70)	0.03	0.08

As shown in Table 2.4, the average |SMD| of propensity score weighting methods are uniformly lower than unweighted method, indicating that propensity score weighting methods give better overall covariate balance than unweighted method. Among the three propensity score weighting methods, the overlap weights achieve the best balance uniformly in terms of both average and maximum |SMD|. This is consistent with the theoretical results in **?**.

In general, the impact of overall cellphone use on crash risk varies by age groups. For teenage and young adult drivers, overall cellphone use significantly increases the crash risk regardless of different weights, with ORs ranging from 3.87 to 4.67. Middle age drivers are less influenced by overall cellphone use, with crash ORs ranging from 1.76 to 2.22. For middle age drivers, the impact of overall cellphone use on crash risk is not significant based on ATE and ATT.

For overall cellphone use, the causal ORs for teenage drivers are greater than unweighted OR and the percentage difference varies from 9.65% to 25.20%. This implies that the crash risk for teenage drivers are underestimated before adjusting for the confounding effects. The pattern is opposite for middle age drivers, for whom the risk is overestimated. For young driver, there are both under-estimation and over-estimation, ranging from -8.08% to 0.48%.

Cellphone visual-manual has a significant impact on drivers of all ages. The crash ORs for visual-manual are uniformly higher than overall cellphone use and cellphone talking, indicating that cellphone visual-manual is the most dangerous cellphone subtasks. The impact of cellphone talking on crash risk is significant only for teenage drivers and young drivers, but not for middle age drivers.

The ORs vary considerably for different propensity score weightings. The substantial

variation indicates heterogeneity in the causal effects across subpopulations. Indeed, in the ideal case where cellphone engagement is randomized, the cellphone engagement and model-driving will be balanced in both observed and unobserved confounders, and all weighting methods, including no weighting, will lead to the same results. Therefore, it is critical to clearly specify the target population for interpreting the causal effects.

The underlying hypothetical population and counterfactual arguments of the three causal effect estimands are different. The ATE focuses on the entire population where the study sample is drawn from and assume all units switched their exposure status; the ATT focuses on the events with cellphone distraction, and assume the distracted units switched to model-driving; the ATO focuses on events that share the similar context described by covariates combinations. For example, the ATE OR of teenage drivers (Table 2.4) is 4.67, indicating that the crash odds if all teenage drivers were engaged with cellphone distraction is 4.67 times the crash odds if all of them were model-driving. The ATT OR, however, indicates that the odds of crash would have been 3.3 (OR = 4.30) times lower if those distracted driving periods were switched to model-driving. Compared with ATE and ATT, the ATO OR is lower, indicating that if we focus on events with similar conditions, the impact of cellphone use on teenage driver is smaller.

2.5 Summary and discussion

While the comprehensive NDS data provide an unprecedented opportunity to evaluate the causal mechanism of crashes in real life driving contexts, the samples, like other observa-

tional studies, is highly unbalanced due to the volitional driving behavior. Both driver's preference and the driving environment influence cellphone engagement decision, and at the same time, are closely associated with driving safety. The relationships among those factors will lead to confounding effects and bias effect estimation. This paper applies propensity score approaches to evaluate the causal effect of cellphone use and emphasizes the necessity of balancing confounders and proper interpretation of results.

The application to the SHRP 2 case-cohort data shows that age, income, and lighting are highly unbalanced and lead to bias in effect estimation. The three propensity score weighting methods, *i.e.*, ATE, ATT, and ATO, effectively improve covariate balance. Stratification by age group is adopted in combination with propensity score weighting to accommodate the substantial difference among drivers.

The choice of estimand depends on the target population from which the research question is raised. Interpretation of the three estimands involve different counterfactual scenarios. In reality, it's unlikely that drivers will use cellphone under all driving scenarios. As ATE estimand is based on the comparison of cellphone use versus absence under any circumstance, it may not be suitable for driving distraction evaluation. The ATT is useful for assessing the impact of cellphone use regulation as it estimates the effects of converting existing cellphone distraction to the selected reference level, *e.g.*, model-driving or all-driving. The ATO estimand focuses on the events with similar driving scenarios and provides an effect estimation for the cellphone use under these "dilemma" conditions. The ATT and ATO estimands are therefore preferred in evaluation and interpretation of the effect of driver behavior on crash risk.

This study provides a rigorous estimation of the causal effect of cellphone use on crash risk under causal inference framework. The results confirm that cellphone use does increase crash risk and the safety impacts vary by age group, implying driving experience and maturity of drivers play an important role in the crash risk management. Teenager drivers tend to be more vulnerable to overall cellphone distraction than young adult and middle age drivers. The results imply potential underestimation of cellphone use effects among for teenage drivers by unadjusted ORs. The results demonstrate that cellphone visual-manual tasks impose much higher risk than cellphone talking across all age groups.

It should be noted that causal effect is only identifiable under unconfoundedness assumption. While SHRP2 NDS data provided a rich information on potential confounders, it is not always guaranteed that unmeasured confounders exist in the estimation of cellphone distraction related crash risk.

This study provides the evidence that there is severe imbalance among SHRP2 case-cohort samples in the evaluation of crash risk associated to cellphone distraction. Propensity score weighting methods improve the covariate balance and the ATT and ATO estimands are suitable for driver behavior risk evaluation. The results provide insights into the causal relationship between cellphone related distractions and crash risk and would contribute to the understanding of the risk of driver behavior.

Chapter 3

Deconfounding through Noisy Proxies: Reconsidering Generative Objectives for Latent Representation[1]

3.1 Introduction

Inferring the individualized treatment effects from observational data is a fundamental challenge shared by many decision-making application domains, including healthcare (Glass et al., 2013), advertising (Bottou et al., 2013), and policy making (LaLonde, 1986), among others. Recent advances in machine learning have motivated new causal inference methodologies inspired by modern learning perspectives, such as representation learning, adversarial training, etc.

[1]This chapter is a published paper: Lu, Danni, Chenyang Tao, Junya Chen, Fan Li, Feng Guo, and Lawrence Carin. "Reconsidering Generative Objectives For Counterfactual Reasoning." *Advances in Neural Information Processing Systems* 33 (2020).

In this work we focus on the problem of causal estimation from observational data, which differs from standard supervised learning in fundamental ways (Pearl, 2009). First, only partial observation of the potential outcomes, the one corresponding to the assigned intervention, can be made. The lack of counterfactual labels prohibits direct validation of the estimated CE. Second, observational studies are susceptible to selection bias due to *confounding*. In particular, some variables obfuscate causation as they affect both treatment assignment and outcome (Vander Weele and Shpitser, 2013), and they may be latent. Without a proper confounder compensation mechanism, causal estimation can face severe bias.

To resolve this difficulty, the classical statistics literature has mainly focused on sample-based adjustment strategies, namely *matching* and *weighting*. Matching pairs units that are similar with respect to particular matching criteria (Schwab et al., 2018b), forming basic elements of synthetic *"randomized trials"*; weighting reassigns importance weights to each sample unit to create a *pseudo* population of better balance (Hainmueller, 2012, Li et al., 2013, 2019). Both approaches typically make the unconfoundedness assumption Rosenbaum (1984), assuming that there are no latent variables that affect both the outcome and the treatment assignment. To guard against model mis-specification-induced failures (Robins et al., 1994), balancing weights are often used in conjunction with outcome regression models to achieve double robustness (Scharfstein et al., 1999). However, these classical solutions are constantly challenged by modern datasets, characterized by features such as high dimensionality (Belloni et al., 2014) and complex interactions (Zhang and Hyvärinen, 2009), and they typically make the unconfoundedness assumption.

More recently, representation learning emerged as a new, promising alternative to ap-

proach covariate balance (Johansson et al., 2020, Li and Fu, 2017). Such schemes explicitly seek an intermediate (low-dimensional) representation that is both (*i*) predictive of the outcome (Veitch et al., 2019); and (*ii*) matched between treatment groups (Johansson et al., 2016). From a learning perspective, these two points serve to promote the generalization performance for counterfactual predictions (Shalit et al., 2017). On the flip side, causal perspectives also motivate invariant feature representation learning under general machine learning setups (Arjovsky et al., 2019).

Recent strides in generative modeling techniques, such as the *variational auto-encoder* (VAE) (Kingma and Welling, 2014) and the *generative adversarial network* (GAN) (Goodfellow et al., 2014a), have equipped causal estimation with new learning principles. Rather than appealing to predictive goals (Veitch et al., 2019), these schemes learn stochastic rules that mimic the data generating procedure, *i.e.*, how to synthesize *realistic* counterfactuals based on observed data (Yoon et al., 2018). Such *generative causal models* typically relax model assumptions posited by standard causal estimation machinery, allowing black-box type inference using flexible learners such as deep networks. Despite their reported strong empirical performance, questions remain: (*i*) Confounding: Do we fully trust the observed confounders? (*ii*) Balancing: What if the covariates are unbalanced? (*iii*) Counterfactual validation: How to avoid over-fitting?

Notably, in-depth discussions on (*iii*), causal validation procedure, has received attention in the literature only recently, despite its paramount importance (Athey and Imbens, 2016, Wager and Athey, 2018b). The promise of a fully automated causal estimation procedure has inspired many (unreliable) heuristic proxies (Schuler et al., 2018) (*e.g.*, plug-in surrogate or predictive loss) and principled evaluation strategies have only appeared quite recently.

While scholarly consensus on best practice is yet to be reached (Dorie et al., 2019), prominent examples from this category include influence function based causal validation (Alaa and Van Der Schaar, 2019) and rank-preserving causal cross-validation (Saito and Yasui, 2019). Of particular interest is the Robinson residual decomposition employed by the R-learner (Nie and Wager, 2017) and generalized causal forests (Athey et al., 2019), which construct a directly learnable objective.

Motivated by the preceding discussions, this work seeks a unified treatment that accommodates (i)-(iii). We revisit the generative perspective of causal modeling, and demonstrate how explicitly accounting for balancing and counterfactual validation helps to improve causal estimation. In particular, we present a variational procedure, termed *Balancing Variational Neural Inference of Causal Effects* (BV-NICE), to address the challenges of generative learning for causal estimation. Our key contributions include: (i) repurposing variational inference as random feature representation learning scheme to facilitate causal estimation; (ii) reformulating the variational objective to better balance confounder representations between comparison groups; (iii) incorporating causal validation targets to scrutinize inferred causal effect. Our approach features direct modeling of causal effects, rather than the difference between the outcome models. It joints strength from distribution matching, representation learning and generative causal estimation, resulting a principled attempt that better addresses the challenges in counterfactual inference. To embrace a more holistic picture, we also cover related issues such as identifiability and establish border connections to the literature on causal discovery with the extended discussions found in our Appendix.

3.2 Preliminaries

3.2.1 Problem setup

We consider the basic setup under the potential outcome framework (Imbens and Rubin, 2015b, Rubin, 1974b). Assume a sample of n units, with unit i associated with a covariate $\boldsymbol{X}_i \in \mathbb{R}^p$, a treatment indicator $T_i \in \{0,1\}$ and potential outcomes $[Y_i(0), Y_i(1)] \in \mathbb{R}^2$. The fundamental problem of causal inference Holland (1986) is that only the outcome associated with the prescribed treatment is observed, *i.e.*, $Y_i \triangleq Y(T_i) = T_i Y_i(1) + (1 - T_i)Y_i(0)$, known as the factual data. The *individualized treatment effect* (ITE) is defined as the expected difference between outcome $\tau(\boldsymbol{x}) \triangleq \mathbb{E}[Y_i(1) - Y_i(0)|\boldsymbol{X}_i = \boldsymbol{x}]$, and our goal is to learn a generalizable model $\tau(\boldsymbol{x})$ that predicts the ITE given observed covariates $\boldsymbol{x}$. We often assume the decomposition $\tau(\boldsymbol{x}) = \mu_1(\boldsymbol{x}) - \mu_0(\boldsymbol{x})$, where $\mu_t(\boldsymbol{x}) \triangleq \mathbb{E}[Y(t)|\boldsymbol{x}], t \in \{0,1\}$ are known as the outcome models. Another key concept in causal estimation is the *propensity score* (PS): $e(\boldsymbol{x}) \triangleq p(T = 1|\boldsymbol{x})$, *i.e.*, the conditional probability of receiving the treatment given $\boldsymbol{x}$. While the identifiability of causal effect can only be established in the average sense for observational studies, under the assumptions of unconfoundedness:$\{Y(0), Y(1)\} \perp\!\!\!\perp T|\boldsymbol{X}$, and positivity: $p(T|\boldsymbol{X}, Y(0), Y(1)) \in (0,1)$ (Rosenbaum and Rubin, 1983a), individualized predictions still hold promise. A typical predictive scheme minimizes the prediction loss for the factual observation, *i.e.*, $\hat{\mu} = \min_\mu \{\sum_i (Y_i - \mu_t(\boldsymbol{X}_i))^2\}$. Alternatively, generative schemes seek to identify a data generation procedure $p_\theta(\boldsymbol{x}, t, y)$ that is consistent with factual observations $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i, t_i)\}_{i=1}^n$.

3.2.2 Robinson residual decomposition

Under unconfoundedness, it is easy to verify $\mathbb{E}[\epsilon(T)|\boldsymbol{X}, T] = 0$, where $\epsilon(T) \triangleq Y(T) - (\mu_0(\boldsymbol{X}) + T\tau(\boldsymbol{X}))$ is known as the Robinson residual (Robinson, 1988). Denoting the conditional mean outcome as $m(\boldsymbol{x}) \triangleq \mathbb{E}[Y|\boldsymbol{x}] = \mu_0(\boldsymbol{x}) + e(\boldsymbol{x})\tau(\boldsymbol{x})$, and we can rewrite Robinson residual as $\epsilon(T) = Y(T) - m(\boldsymbol{X}) - (T - e(\boldsymbol{X}))\tau(\boldsymbol{X})$. Note that this decomposition holds for any outcome distribution, including binary outcomes. This directly motivates the R-learning (Nie and Wager, 2017) objective $\hat{\tau} = \arg\min_\tau \{1/n \sum_i (y_i - \tilde{m}(\boldsymbol{x}_i) - (t_i - \tilde{e}(\boldsymbol{x}_i))\tau(\boldsymbol{x}_i))^2\}$, where $\tilde{m}(\boldsymbol{x})$ and $\tilde{e}(\boldsymbol{x})$ are estimated surrogates for the mean outcome and propensity score model. Recently, many have considered the of direct modeling of causal effect (τ) through the R-decomposition (Athey et al., 2019, Chernozhukov et al., 2018, Powers et al., 2018, Zhao et al., 2017a,b), rather than indirectly through (μ_0, μ_1).

3.2.3 Variational inference

A general learning principle is to maximize the expectation of the log-likelihood wrt observed data, *i.e.*, $\ell(\theta) := \sum_i \log p_\theta(\boldsymbol{x}_i)$, which constitutes *maximum likelihood estimation* (MLE). For a latent variable model $p_\theta(x, z)$, we consider x as an observation (*i.e.*, data) and z as latent variable. The marginal likelihood $p_\theta(x) = \int p_\theta(\boldsymbol{x}, \boldsymbol{z})\, \mathrm{d}z$ typically does not have a closed-form expression, and to avoid direct numerical estimation of $p_\theta(\boldsymbol{x})$, variational inference (VI) instead optimizes a variational bound to the marginal log-likelihood $\log p_\theta(\boldsymbol{x})$ (Blei et al., 2017, Tao et al., 2018). The most popular choice is known as the *Evidence Lower Bound* (ELBO), given by

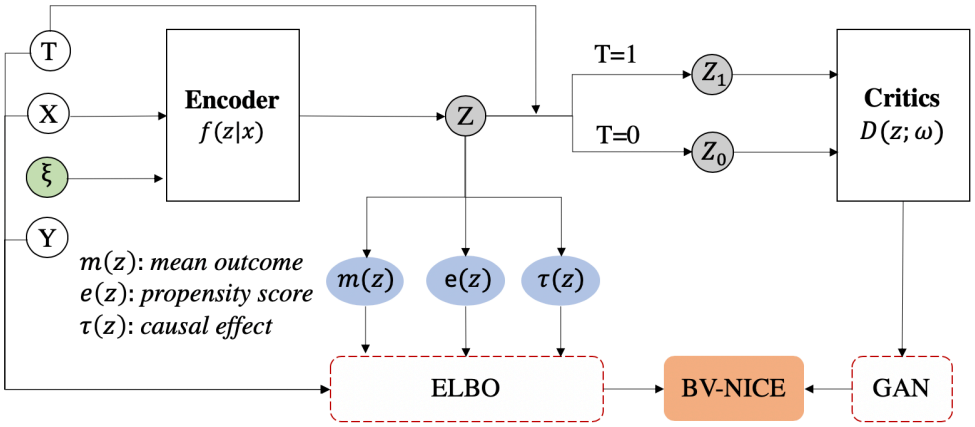

Figure 3.1: BV-NICE model architecture.

$$\text{ELBO} \triangleq \mathbb{E}_{Z\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x},\boldsymbol{Z})}{q_\phi(\boldsymbol{Z}|\boldsymbol{x})}\right] \leq \log p_\theta(\boldsymbol{x}), \tag{3.1}$$

where $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is an approximation to the true posterior $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ and the inequality is a result of Jensen's inequality. This bound tightens as $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ approaches the true posterior $p_\theta(\boldsymbol{z}|\boldsymbol{x})$. For estimation, we seek parameters θ that maximize the ELBO, and the commensurately learned parameters ϕ are often used in a subsequent inference task with new data.

3.2.4 Adversarial distribution matching

Consider the problem of matching a model distribution $p_G(x)$ to some true data distribution $p_d(x)$ presented as empirical samples, wrt some discrepancy measure, $d(p_d, p_G)$. Typically, $p_G(x)$ is given in the form of a stochastic sampler. In the GAN framework, the discrepancy is first estimated by maximizing an auxiliary variational functional $V(p_d, p_G; D) : \mathcal{P}\times\mathcal{P} \to \mathbb{R}$ between distributions $p_d(x)$ and $p_G(x)$ satisfying $d(p_d, p_G) = \max_D\ V(p_d, p_G; D)$, where $\mathcal{P}$ is

the space of probability distributions and $V(p_d, p_G; D)$ is estimated using samples from the two distributions. Function $D(x; \omega)$, parameterized by ω and known as the *critic* function, is intended to maximally discriminate between samples of the two distributions. Subsequently, one seeks to match the generator distribution $p_G(x)$ to the unknown true distribution $p_d(x)$ by minimizing the estimated discrepancy, resulting in a minimax game between the critic and the generator: $\min_G \max_D \; V(p_d, p_G; D)$.

3.3 Balancing VI for causal estimation

Inspired by the above, we present BV-NICE, a model seeks to improve the current practice of generative learning of causal inference from the following perspectives: (a) automated feature representation learning that explicitly accounts covariate balance, (b) a built-in mechanism for automated model selection directly targets causal effect estimation accuracy, (c) acknowledging the uncertainty in the observed confounders by introduction of inferred latent variables.

We frame our construction under variational inference based on the following considerations: (i) We treat covariate $\boldsymbol{x}$ as noisy proxies for the true, unobservable confounders (latent $\boldsymbol{z}$); (ii) We treat covariate $\boldsymbol{x}$ as noisy proxies for the true, unobservable confounders (latent $\boldsymbol{z}$); (iii) Matching for the prior $p(\boldsymbol{z})$ naturally regularizes for model generalization. Consider the following latent variable model $p_\theta(\boldsymbol{x}, y, t, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(y|\boldsymbol{z}, t)p(t|\boldsymbol{z})p(\boldsymbol{z})$ (Figure 3.2), where $(\boldsymbol{x}, y, t)$ are the observables, $\boldsymbol{z}$ is the (continuous) latent variable, and θ denotes the model parameters. In accordance with standard practice, we model discrete variables

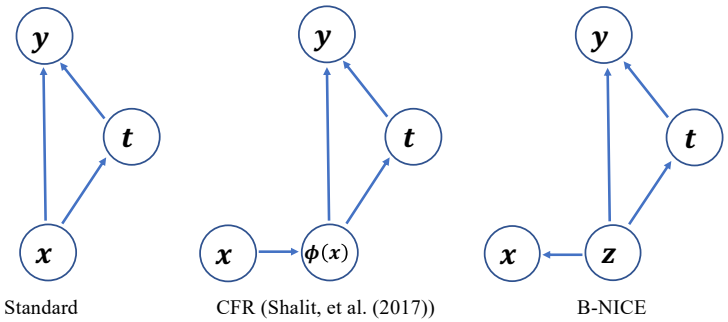

Figure 3.2: Comparison of causal graphs for different models.

with multinomial logistic and continuous variables with Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, where $\boldsymbol{\mu}$ is a function of $\boldsymbol{z}$ and also possibly t depending on the context, with $\boldsymbol{\sigma}^2$ set to some prescribed value to avoid overfitting. We parameterize stochastic encoders $q_\phi(\boldsymbol{z}|\boldsymbol{x}, y, t)$ to infer unobserved confounders $\boldsymbol{z}$. For flexible inference, we model all functions with deep neural nets. Plugging into Eq. (3.1) gives us a tractable objective for stochastic optimization (see Eq. (3.6)). We relegate the specifics of our modeling choices in the subsections that follow, after revealing more causal insights embodied in our reformulation.

Note that our approach relaxes the common (restrictive) Gaussian assumption on the outcome model, namely $\mathcal{N}(y; \mu(\boldsymbol{x}), \sigma^2)$, made in the causal literature (Dorie et al., 2019), which is likely misspecified in realistic settings. In our setup $Y|X$ is modeled as an infinite mixture of Gaussians, such that it can approximate arbitrary distributions. Added randomness is also an effective way to combat overfitting for flexible learners, especially in the low-sample regime. Individual samples are now represented as an infinite number of points in the latent space, where the distribution discrepancy term forces overlap between unit representations. Later we will also see that stochastic representation learning provides

a promising alternative to classical covariate balancing strategies, such as weighting (Figure 3.3). The added randomness dissuades flexible learners from memorizing training examples.

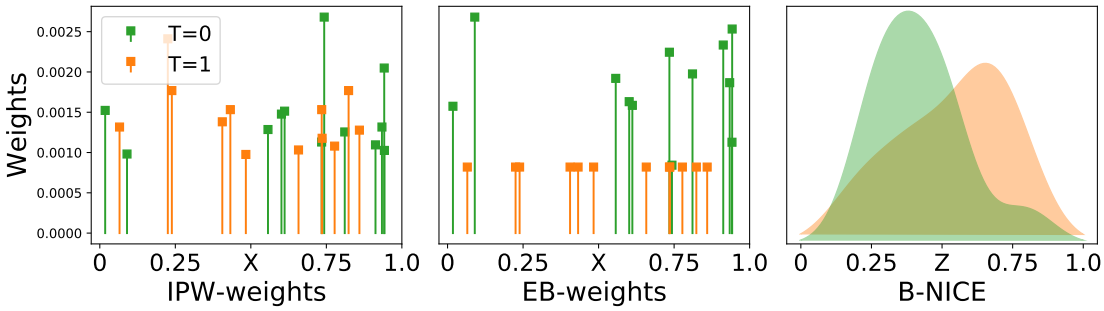

Figure 3.3: Comparison of weighted and stochastic representations.

3.3.1 A unifying view for VI and R-learner

A key feature we seek to incorporate is to automatically favor solutions that more accurately describe causal effect based on the factual observations. Unlike a model-selection procedure, where candidates are screened in an *ad hoc* manner, we want our model to explore the parameter space, to identify the best candidates for causal descriptives as part of training. This precludes options such as *meta-learners* (Künzel et al., 2019) and influence function based estimator (Alaa and Van Der Schaar, 2019), as they function as a causal estimator and cannot be efficiently trained in an end-to-end manner. We choose to work with the Robinson residual decomposition, and show how the resulting R-learner (Nie and Wager, 2017) relates to VI. This implies our variational framework automatically assumes the model selection property.

It is convenient to denote $\mu_t(\boldsymbol{z}) \triangleq \mu_y(\boldsymbol{z}, t)$. Given the Gaussian assumption, it is easy to

construct the causal effect estimator $\tau(\boldsymbol{z})$ as $\tau(\boldsymbol{z}) = \mu_y(\boldsymbol{z}, t=1) - \mu_y(\boldsymbol{z}, t=0)$. Under the R-learning framework, one models the mean outcome $m(\boldsymbol{z})$ and $\tau(\boldsymbol{z})$ rather than $(\mu_0(\boldsymbol{z}), \mu_1(\boldsymbol{z}))$. It is easy to see these two modeling choices are related by

$$\begin{cases} m(\boldsymbol{z}) = e(\boldsymbol{z})\mu_1(\boldsymbol{z}) + (1 - e(\boldsymbol{z}))\mu_0(\boldsymbol{z}) \\ \tau(\boldsymbol{z}) = \mu_1(\boldsymbol{z}) - \mu_0(\boldsymbol{z}) \end{cases} \tag{3.2}$$

allowing us to rewrite (μ_0, μ_1) as

$$\begin{cases} \mu_0(\boldsymbol{z}) = m(\boldsymbol{z}) - e(\boldsymbol{z})\tau(\boldsymbol{z}), \\ \mu_1(\boldsymbol{z}) = m(\boldsymbol{z}) + (1 - e(\boldsymbol{z}))\tau(\boldsymbol{z}). \end{cases} \tag{3.3}$$

A key insight is given by the observation

$$\begin{aligned} \epsilon(\boldsymbol{z}, t, y) &= y - m(\boldsymbol{z}) - (1 - e(\boldsymbol{z}))\tau(\boldsymbol{z}) \\ &= y - \{t\mu_1(\boldsymbol{z}) - (1-t)\mu_0(\boldsymbol{z})\}. \end{aligned} \tag{3.4}$$

Note that the RHS is the residual error for prediction given $(\boldsymbol{z}, t)$. Consequently, under our Gaussian assumption,

$$\ell_R(\boldsymbol{z}, y, t) = \epsilon(\boldsymbol{z}, t, y)^2 = -2\sigma^2 \log p_\theta(y|\boldsymbol{z}, t) \tag{3.5}$$

Plugging this result back into the ELBO, and recalling that $p_\theta(t|\boldsymbol{z})$ is essentially the propen-

sity score model $e(\boldsymbol{z})$, we obtain the following factorization

$$\text{ELBO}(\boldsymbol{x}, y, t|p_\theta, q_\phi) = \mathbb{E}_{\boldsymbol{Z}\sim q_\phi}[\underbrace{\log p_\theta(\boldsymbol{x}|\boldsymbol{Z})}_{\text{Optional}} + \underbrace{\overbrace{\log p_\theta(y|\boldsymbol{Z}, t)}^{\tau\text{-loss}} + \overbrace{\log p_\theta(t|\boldsymbol{Z})}^{\text{PS-loss}}}_{\ell_{\text{V-NICE}}} - \overbrace{KL(q_\phi(\boldsymbol{z}|\boldsymbol{x}, y, t) \parallel p(\boldsymbol{z}))}^{KL\text{-loss}}] \tag{3.6}$$

Since our primary goal is to model the causal effect τ, we discard the first term related to the likelihood of $\boldsymbol{x}$ and treat the rest as our training target, which we term $\ell_{\text{V-NICE}}$. This choice is motivated by the fact that to correctly infer causal effect we only need the part of $\boldsymbol{x}$ that is predictive of (y, t) (Veitch et al., 2019). Modeling $\boldsymbol{x}$ indiscriminately takes away representation capacity of $\boldsymbol{z}$ (Alemi et al., 2018, Higgins et al.), compromising our main objective.

Intuitively, $\ell_{\text{V-NICE}}$, our reformulated ELBO, is a combination of τ-loss and propensity score loss, regularized by KL-divergence on the inferred latent distribution to encourage better generalization. Two differences compared to existing generative causal models are that (a) our model is directly parameterized through causal triplet $(\tau(\boldsymbol{z}), m(\boldsymbol{z}), e(\boldsymbol{z}))$ to emphasize the causal perspective; and (b) it does not seek to capture the generative procedure for $\boldsymbol{x}$. It is easy to see as we set $\sigma^2 \to 0$, V-NICE approximately recovers the R-learner. The R-learner employs a two stage procedure to estimate τ based on pilot estimates of (e, m), while we learn the triplet jointly. A potential benefit is that our triplet share the refined representation learned, and direct modeling of τ allows structural constraints to be imposed (Künzel et al., 2019). Our discussion also bridges R-learning and likelihood-based learning.

Benefits of integrating the R-loss. A major difference in the construction of R-learner objective, relative to the standard two-learner setup, is that the propensity score is

explicitly involved. This allows additional information to be leveraged in many practical settings. For example, a common scenario is that significant lags can be expected between the application of a treatment and the observation of the outcome (*e.g.*, when the target outcome is the patients' recovery in one year time whether or not administrating a drug). In such scenarios, there will be data available with only confounder and treatment to refine propensity score estimate, which in turn improves treatment effect estimation in R-learning, but can not be used for outcome modeling in the two learner setup. A similar argument holds when additional knowledge is known about the treatment assignment (*e.g.*, when the data is a hybrid of observational and randomized trial). In the same spirit, R-learning also allows the use of data where the treatment information is missing, as they can still be used to improve the estimate of average outcome $m(\boldsymbol{x})$.

3.3.2 Balancing VI

Our next goal is to establish a mechanism that enables covariate balance. Further denote $q_t(\boldsymbol{z}) = \int q_\phi(\boldsymbol{z}|\boldsymbol{x}, t) p_d(\boldsymbol{x}|T = t)\,\mathrm{d}\boldsymbol{x}$. To achieve better balance for subsequent causal estimation, one seeks to match the confounder distributions between treatment groups. In other words, q_1 should be close to q_0. To this end, we augment the original ELBO with a distribution discrepancy score $\mathbb{D}(q_0 \parallel q_1)$, resulting in

$$\ell_{\text{BV-NICE}} \triangleq \ell_{\text{V-NICE}}(p_\theta, q_\phi) - \lambda \mathbb{D}(q_0 \parallel q_1) \tag{3.7}$$

as a modified training objective for balancing VI (BV-NICE), where $\lambda > 0$ specifies the

regularization strength.

Choice of discrepancy score. While the marginal densities of q_0 and q_1 are intractable, it is relatively easy to acquire samples from the respective distributions. This motivates leveraging adversarial distribution matching strategies to (indirectly) optimize the discrepancy term through a *mini-max* game. Hence, we indirectly assess $\mathbb{D}(q_0 \parallel q_1)$ via the use of a critic function (the max step), and then update the model accordingly to reduce the discrepancy (the min-step). In this study, we appeal to the KL-divergence as our discrepancy measure, which can be recast in its Fenchel dual form as (Dai et al., 2018, Tao et al., 2019)

$$\mathbb{D}_{KL}(q \parallel p) = \mathbb{E}_q\left[\log q - \log p\right] = \max_{\nu>0}\{\underbrace{\mathbb{E}_q\left[\log \nu\right] - \mathbb{E}_p\left[\nu\right] + 1}_{V_{KL}(p,q;\nu)}\}, \tag{3.8}$$

and note the maximizer ν^* satisfies $\nu^* = \frac{q}{p}$. Our choice for $\mathbb{D}_{KL}$ is motivated by the following considerations: (*i*) easy implementation relative to integral probability metric (IPM)-based schemes; (*ii*) it also bounds generalization performance (Sec 3.3); (*iii*) this approach also encourages parameter sharing as the ELBO involves a KL term.

Note that this choice is not restrictive, as practitioners are free to choose their favorite distribution matching schemes, such as Wasserstein (Arjovsky et al., 2017, Shalit et al., 2017), MMD (Gretton et al., 2012, Li and Fu, 2017) or JSD (Goodfellow et al., 2014a, Yoon et al., 2018), that possess other appealing properties. We provide a more thorough discussion of this in the Appendix.

To implement KL-matching, we model $\log \nu$ as a deep neural network $\vartheta_\psi(\boldsymbol{z})$, as our

Algorithm 1 BV-NICE

Empirical data $\hat{p}_d = \{(\boldsymbol{x}_i, y_i, t_i)\}_{i=1}^n$, imbalance λ
for $k = 1, 2, \cdots$ **do**
$(\boldsymbol{x}, y, t) \sim \hat{p}_d, \boldsymbol{z}' \sim p(\boldsymbol{z}), \boldsymbol{z}_\phi = G_\phi(\boldsymbol{\xi}, \boldsymbol{x}), \boldsymbol{\xi} \sim p(\boldsymbol{\xi})$
$\phi_{k+1} \leftarrow \nabla_\phi \{\log p_\theta(t, y|\boldsymbol{z}_\phi) - \vartheta_\psi(\boldsymbol{x}, \boldsymbol{z}_\phi)$ % Encoder
$- \lambda[\tilde{\vartheta}_{\tilde{\psi}}(\boldsymbol{z}_\phi^{t=0}) - \exp(\tilde{\vartheta}_{\tilde{\psi}}(\boldsymbol{z}_\phi^{t=1}))]\}$ % Balancing
$\theta_{k+1} \leftarrow \nabla_\theta \{\log p_\theta(t, y|\boldsymbol{z}_\phi)\}$ % Model
$\psi_{k+1} \leftarrow \nabla_\psi \{\vartheta_\psi(\boldsymbol{x}, \boldsymbol{z}_\phi) - \exp(\vartheta_\psi(\boldsymbol{x}, \boldsymbol{z}'))\}$ % Critic
$\tilde{\psi}_{k+1} \leftarrow \nabla_{\tilde{\psi}} \{\tilde{\vartheta}_{\tilde{\psi}}(\boldsymbol{z}_\phi^{t=0}) - \exp(\tilde{\vartheta}_{\tilde{\psi}}(\boldsymbol{z}_\phi^{t=1}))\}$ % Critic
end for

critic function, where ψ denotes the network parameters. This gives the following neural estimator for the KL term[2]

$$\hat{\mathbb{D}}_{KL}(q \parallel p) = \max_\psi \{\mathbb{E}_{Z\sim q}[\vartheta_\psi(Z)] - \mathbb{E}_{Z'\sim p}[\exp(\vartheta_\psi(Z'))]\} \tag{3.9}$$

In our case, the distributions are characterized by a neural sampler via the reparameterization trick, *e.g.*, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ as $G_\phi(\boldsymbol{\xi}, \boldsymbol{x})$, $\boldsymbol{\xi} \sim p(\boldsymbol{\xi})$. Gradients of the sampler can be easily obtained by directly differentiating $\hat{\mathbb{D}}_{KL}$ with respect to ϕ.

3.3.3 Practical implementation

Random feature encoder. To enable flexible encoding of latent features, we employ an implicit neural sampler $r_{\phi'}(\boldsymbol{z}|\boldsymbol{x})$. Here $r_{\phi'}(\boldsymbol{z}|\boldsymbol{x})$ can either be an explicit sampler that has a tractable likelihood (*i.e.*, reparametrized Gaussian (Kingma and Welling, 2014), autoregressive flow (Kingma et al., 2016)) , or more expressive implicit sampler that directly

[2]Note that we have dropped the constant term for clarity.

maps $\boldsymbol{x}$ and a noise vector to a latent sample, *i.e.*, $\boldsymbol{z} = G_{\phi'}(\boldsymbol{\xi}, \boldsymbol{x}), \boldsymbol{\xi} \sim \mathcal{U}([-1, 1]^k)$. For the implicit sampler, the KL divergence can be estimated, and optimized over, through the Fenchel mini-max estimator introduced above. We have experimented with both strategies and find the implicit feature encoder produces better results.

Another empirical choice we need to make is whether to include treatment and outcome during the encoding phase. Both choices induces a valid lower bound, with the one includes (t, y) being sharper as both variables contain information about the latent according to our model assumption. While the inclusion is practiced in Louizos et al. (2017), we argue otherwise. First, it complicates inference procedure, as auxiliary models must be introduced for efficient sampling of the latent, which may further introduce approximation error. Secondly, the identification of casual effect is plausible when the assignment scheme is independent with potential outcome conditional on a potnetially large set of covariates. The inclusion of outcome in the encoder will, on the contrary, potentially introduce bias and violates this unconfoundedness assumption.

Practical variants. Modifications to the original VI procedures are often considered by the practioners, as it is often observed that strictly conforming to VI setup typically leads to suboptimal results. These "hacks" can sometimes be understood as some compensation mechanism at play to correct for the model mis-specification, which biases the VI estimation. We consider in this work two variants that are more principally derived: β-VAE (Higgins et al.) and *adversarial auto-encoder* AAE (Makhzani et al., 2015). The former seeks to address the vanishing KL often observed in training, while the later explicitly targets the mismatch between the aggregated posterior and prior. Both strategies diminishes the role of KL term in ELBO, which often compromises empirical performance via synthesizing

uninformative latents to reduce the mismatch to the prior. Limited by space, we describe implementation details in the Appendix.

Inferring causal effects. Given a new observation $\boldsymbol{x}$, we wish to infer the expected effect $\tau(\boldsymbol{x})$ for a given intervention under the learned model. Since under BV-NICE causal effect $\tau(\boldsymbol{z})$ is defined based on the latent variable $\boldsymbol{z}$ rather than the observed $\boldsymbol{x}$, the estimation of the causal effect becomes a two-stage process. In the first stage we infer the hidden $\boldsymbol{z}$ given $\boldsymbol{x}$, and in the second stage we average over the latent variables to estimate the causal effect for $\boldsymbol{x}$. An estimate of the causal effect is given by

$$\tau(\boldsymbol{x}) \approx \frac{1}{m} \sum_j \tau(\boldsymbol{z}'_j), \boldsymbol{z}'_j \sim q_\phi(\boldsymbol{z}|\boldsymbol{x}). \tag{3.10}$$

Counterfactual cross-validation with R-residual. A major obstacle in counterfactual reasoning is that due to the absence of counterfactual observations, models can not be validated directly. In our setting, we applied R-loss to hold out factual observations to cross-validation our model. Although it may seem similar to the CV applied in standard machine learning, a key distinction should be noted: that our CV target is explicitly defined wrt the counterfactual estimates. As noted in Nie and Wager (2017), Schuler et al. (2018), factual residual does not effectively assess counterfactual performance, resulting biased or unreliable estimation.

Comparison to related models. The most closely related models are CE-VAE (Louizos et al., 2017) and R-learner (Nie and Wager, 2017). Compared with CE-VAE, our model: (a) accounts for covariate balance; (b) exploits alternative parameterization, $(m(\boldsymbol{z}), \tau(\boldsymbol{z}))$ instead of $(\mu_0(\boldsymbol{z}), \mu_1(\boldsymbol{z}))$, a direct model of the causal effect; (c) does not

necessarily need to learn the covariate model $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, which can dominate representation learning for rich $\boldsymbol{x}$, compromising causal estimation; and (d) employs flexible representation learning (flexible approximate posterior), parameterized by a neural sampler. Compared to R-learner, we jointly train the outcome $m(\boldsymbol{z})$ and propensity $e(\boldsymbol{z})$ models together with the causal effect model $\tau(\boldsymbol{z})$, regularized by the latent prior $p(\boldsymbol{z})$, and assign a probabilistic interpretation.

3.3.4 Generalization bounds for BV-NICE

We provide theoretical justification for the use of KL balancing. In particular, we show that the counterfactual generalization error can be bounded by the factual error plus a KL-term of the representation distributions between the treatment groups, adjusted by the variance of the conditional outcome model. We also provide additional discussions on other theoretical aspects in the Appendix.

Definition 3.1. The expected loss for the unit and treatment pair $(\boldsymbol{z}, t)$ is

$$\ell_h(\boldsymbol{z}, t) = \int_{\mathcal{Y}} L(Y_t, h(\boldsymbol{z}, t)))p(Y_t|\boldsymbol{z})\,\mathrm{d}Y_t, \tag{3.11}$$

where $L(y; h)$ denotes some loss with respect to observation y and hypothesis h, and $\boldsymbol{z}$ is parameterized via the stochastic encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$. The expected factual and counter factual

losses of h and ϕ are:

$$\begin{aligned}\epsilon_{\mathrm{F}}(h,\phi) &\triangleq \textstyle\int_{\mathcal{Z}\times\{0,1\}} \ell_h(\boldsymbol{z},t)p_\phi(\boldsymbol{z},t)\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}t,\\ \epsilon_{\mathrm{CF}}(h,\phi) &\triangleq \textstyle\int_{\mathcal{Z}\times\{0,1\}} \ell_h(\boldsymbol{z},t)p_\phi(\boldsymbol{z},1-t)\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}t,\end{aligned} \tag{3.12}$$

where $p_\phi(\boldsymbol{z},t) = \int q_\phi(\boldsymbol{z}|\boldsymbol{x},y,t)p_d(\boldsymbol{x},y,t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}y$. The expected factual treated ($t=1$) and control ($t=0$) losses are

$$\begin{aligned}\epsilon_{\mathrm{F}}^{t=1}(h,\phi)_1 &\triangleq \textstyle\int_{\mathcal{X}} \ell_{h,\phi}(\boldsymbol{z},1)q_1(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}t\\ \epsilon_{\mathrm{F}}^{t=0}(h,\phi)_1 &\triangleq \textstyle\int_{\mathcal{X}} \ell_{h,\phi}(\boldsymbol{z},0)q_0(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}t\end{aligned} \tag{3.13}$$

where $q_t(\boldsymbol{z})$ are the posterior of $\boldsymbol{z}$ given t.

Definition 3.2. *Precision of estimating heterogeneous effects* (PEHE) for a causal effect estimator $\hat{\tau}$ is defined as

$$\epsilon_{PEHE}(\hat{\tau}) \triangleq \mathbb{E}\|\hat{\tau}-\tau\|^2_{L^2(\mathbb{P})}, \tag{3.14}$$

where $L^2(\mathbb{P})$ denotes the L^2 norm with respect to the feature distribution $\mathbb{P}(\boldsymbol{x})$.

The following statements assert the generalization error for PEHE can be bounded by the factual error plus a KL-discrepancy term, adjusted by the variance of outcome.

Lemma 3.3. *Let $\Phi : \mathcal{X} \to \mathcal{R}$ be an invertible representation with Ψ its inverse. Let $q_t, t \in \{0,1\}$ be defined as Sec 3.3.2. Let $u \triangleq p(T=1)$ be the prevalence of treatment, and $h : \mathcal{R}\times\{0,1\} \to \mathcal{Y}$ is a hypothesis. Assume $\|l_{h,\Phi}(\Psi(r),t)\|_\infty \le M$ for $t=\{0,1\}$. Then we have:*

$$\epsilon_{CF}(h,\Phi) \le (1-u)\cdot\epsilon_F^{t=1}(h,\Phi) + u\cdot\epsilon_F^{t=0}(h,\Phi) + \tfrac{1}{2M}\sqrt{\tfrac{1}{2}\mathbb{D}_{KL}\left(q_0\|q_1\right)} \tag{3.15}$$

Theorem 3.4. *Under the conditions of Lemma 3.3, and assuming the loss L used to define* $l_{h,\Phi}$ *is the squared loss, we have:*

$$\epsilon_{PEHE}(h,\Phi) \leq 2\epsilon_F^{t=0}(h,\Phi) + 2\epsilon_F^{t=1}(h,\Phi) + \frac{1}{M}\sqrt{\frac{1}{2}\mathbb{D}_{KL}(q_0||q_1)} - 4\sigma_Y^2 \tag{3.16}$$

This result bears resemblance to the generalization bound proved in Shalit et al. (2017). The key difference is that we have replaced the IPM bound with a KL bound. The original implementation of *counterfactual regression* (CFR) model used the Sinkhorn iterations or MMD computed their IPM, which scales quadratically with respect to mini-batch size. Our Fenchel dual KL estimation scales linearly with respect to sample size, and consequently more scalable. And the new assumption on $\ell_{h,\phi} \in L^\infty$ is generally easily satisfied in practice, while the RKHS assumed by CFR is difficult to verify.

3.4 Related work

Bayesian causal estimation can be classified based on how the uncertainty is accounted for. Classical approaches place uncertainty on the model itself, with prominent examples such as *Bayesian additive regression trees* (BART) (Chipman et al., 2010). To flexibly model the complex causation, Bayesian nonparametric (BNP) schemes have become popular (Hill, 2011). Alaa and van der Schaar (2018) investigated the fundamental limit of information rate for BNP causal models. Closest to this paper is the work of causal estimation VAE (CE-VAE) (Louizos et al., 2017), where latent variables are introduced to account for the uncertainty, with the model learned through variational Bayesian analysis.

Our work enhances CE-VAE by infusing additional causal perspectives into its construction: we explicitly address the covariate balancing issue and elaborate how VI connects to R-learning, based on which a reformulated ELBO is derived. Also highly relevant are the works of Bayesian *counterfactual risk minimization* (CRM) (London and Sandler, 2019, Wu and Wang, 2018), where KL-divergence on the policy (model) distributions is regularized to upper bound excess risk. Our BV-NICE differs in promoting representation balance to reduce generalization risk.

Representation learning has drawn considerable attention in counterfactual inference. Early work explored the use of shrinkage estimators, such as LASSO (Belloni et al., 2014) and elastic-net (Athey et al., 2018). Recently, nonlinear representation learning has gained considerable momentum in recognition of growing data complexity (Li and Fu, 2017). Popular strategies include kernelization (Li and Fu, 2017), neural encoding (Kallus, 2018b), and representation embedding (Veitch et al., 2019). While most approaches adopt a deterministic design (Alaa and Schaar, 2018, Li and Fu, 2017, Shalit et al., 2017), stochastic variants are considered in the works of CE-VAE (Louizos et al., 2017), CE-GAN (Lee et al., 2018) and CE-IB (Parbhoo et al., 2018), which enable additional flexibility and better matching, and consequently improved generalization (Kingma and Welling, 2014). Distinct from prior arts, BV-NICE directly targets representations for causal estimation and balancing rather than focusing on predictive performance (Nie and Wager, 2017). See the Appendix for further discussions and causal perspectives on invariant representation learning (Arjovsky et al., 2019, Zhang et al., 2013, 2015).

Generative causal learning is an emerging subject in causal inference. The burgeoning field of generative modeling provides ample new tools and inspiration for causal

modeling. GAN-based variants have been most successful in finding direct applications for counterfactual practice (Abdullahi et al., 2017, Averitt et al., 2020, Kocaoglu et al., 2017, Ozery-Flato et al., 2018, Yoon et al., 2018), and to a lesser extent with variational schemes (Louizos et al., 2017, Parbhoo et al., 2018, Puli and Ranganath, 2019). Indirectly, the counterfactual literature has also greatly benefited from borrowing tools originally developed for generative modeling (Shalit et al., 2017), such as distribution matching schemes (Arjovsky et al., 2017, Gretton et al., 2012). Our work presents a principled attempt to integrate generative and causal views, by bringing together counterfactual reasoning, variational learning and adversarial matching.

Covariate balancing is challenged by the fragility of conventional schemes applied to modern datasets. As discussed previously, matching criteria often fail in the presence of nuisance noise (Luo and Zhu, 2019), while the use of weighting strategies are limited by their restrictive linear assumptions (Athey et al., 2018), unreliable propensity estimates (Kallus, 2018b), or unscalable numerical schemes (Hainmueller, 2012). This motivates a variety of work exploring representation learning with direct regularization of imbalance metrics, such as Mahalanobis, Wasserstein, and MMD measures (Athey et al., 2018, Bertsimas et al., 2015, Zubizarreta, 2012, 2015), to learn a proper representation, and possibly in conjunction with a (learned) weighting strategy (Johansson et al., 2018, Kallus, 2018b), to mitigate the representation mismatch. A generalization argument was provided by Shalit et al. (2017) to support such practice. While some works demonstrate the gains from adopting a sophisticated balancing criteria (Yao et al., 2018), here we advocate the use of a simple, flexible KL-balancing under a generative framework.

Hidden confounding is detrimental to many representation learning and covariate

balancing methods that posit the *ignorability or unconfoundedness* assumption Rosenbaum (1984). The residual confounding due to noisy measurement and unobserved confounders remains as major challenges in practice, threatening the validity of causal estimation (Groenwold et al., 2009). Sensitivity analysis is advised to assess the potential effect of unmeasured confounders on causal estimates Franks et al. (2019), Rosenbaum and Rubin (1983b). Extensive investigations have been done on robust recovery of (equivalent) causal graphs with the presence of unobserved latents (Maathuis et al., 2009, Spirtes et al., 2013, Witte et al., 2020), and potential synergies can be exploited between recent advances in causal discovery and counterfactual reasoning. Limited by space, we defer an extended discussion on this to the Appendix.

Consistency and identifiability are key concepts of parallel interest to generalizability. Beyond the common assumption of strong ignorability, conditions to ensure identifiability in the presence of latent variables have been adequately discussed in D'Amour (2019), Miao et al. (2018) and the references therein, and we note their settings are drastically simpler than what's assumed by BV-NICE. Most related to this work are the emerging theories on the identifiability for latent variables under the general framework assumed by variational inference (Khemakhem et al., 2020). While a full exposition on the topic in the context of causal inference is beyond the scope of this study, we refer readers to our Appendix for some preliminary discussions.

3.5 Experiments

To compare the performance of the proposed BV-NICE and its predecessors, we consider a wide range of semi-synthetic and real-world tasks to validate our models experimentally. Details of the experimental setup are described the Appendix, and our code is available (upon publication) from `https://github.com/DannieLu/BV-NICE`. Importantly, we want to experimentally unveil aspects that are important for the design of generative causal models. More analyses can be found in the Appendix.

3.5.1 Experimental setups

For all instantiations, we use fully-connected *multi-layer perceptrons* (MLP) as our flexible learner. We randomly sample model architectures (number of layers, hidden units) and other hyper-parameters (learning rate, batch-size, regularization strength, etc.). For practical cross-validation, we use 7/3 split for training and validation respectively, and rely on validation outcome RMSE to set best configuration [3]. Similarly, results are reported when validation RMSE stops improving. We standardize covariate $\boldsymbol{x}$ and outcome y to zero mean and unit variance according to training set statistics.

To extensively validate the proposed procedure in a realistic setup, we consider the following three datasets: (*i*) *IHDP1000* (Hill, 2011): a semi-synthetic dataset with $1,000$ simulations of different treatment and outcomes mechanism. (*ii*) *ACIC2016* (Dorie et al., 2019): a benchmark dataset released by *Atlantic Causal Inference Competition*, which in-

[3] Note this is equivalent to the Robbinson residual validation.

volves 77 semi-synthetic datasets with 100 replications each. (*iii*) *JOBS* (LaLonde, 1986): a real-world dataset with binary outcomes, a small portion of the data comes from randomized trials. (*iv*) SHRP2 (Guo, 2019b): a 3-year case-cohort study of driver behavior and environmental factors at the onset of crashes and under normal driving conditions, derived from over 1 million hours of continuous video recordings. Detailed descriptions of these datasets can be found in the Appendix.

Evaluation metrics. To quantitatively assess the performance of competing causal inference procedures, we consider the following performance metrics from the literature: (*i*), ITE accuracy as quantified by ϵ_{PEHE}; (*ii*) *policy risk* $R_{pol} \triangleq 1 - \pi_f \cdot \mathbb{E}[Y(1)|f(\boldsymbol{X}) = 1)] - (1 - \pi_f) \cdot \mathbb{E}[Y(1)|f(\boldsymbol{X}) = 0)]$ (Shalit et al., 2017), where $f(\boldsymbol{x}) : \mathcal{X} \to \{0, 1\}$ denotes a decision rule whether to apply the treatment and π_f denotes the portion of population receives the treatment under $f(\boldsymbol{x})$. Note that policy risk only applies to datasets with RCT.

Baseline solutions To compare, the following strong or popular causal estimation baselines are considered: linear regression (OLS, with the T-learner setup); *Bayesian Additive Regression Trees* (BART) (Chipman et al., 2010), *Causal Random Forests* (Causal RF) (Wager and Athey, 2018b), and *Counterfactual Regression* (CFR) (Shalit et al., 2017).

3.5.2 Dissecting VI for counterfactual reasoning

We first investigate which factors greatly impact the performance to support decision choices for the construction of generative causal models. In particular, we seek answers to the following points through the lens of empirical experiments: (*a*) level of uncertainty in feature representation; (*b*) degree of balancing (overlapping); (*c*) sorts of distributional regulariza-

tions.

To see how representation uncertainty affect performance, we introduce a randomness parameter $\eta \geq 0$, that scales the noise input to the our stochastic feature encoder, *i.e.*, $\boldsymbol{z} = G(\eta \cdot \boldsymbol{\xi}, \boldsymbol{x})$. We carried out grid search for configuration of (λ, η) on both IHDP and ACIC. In Figure 3.4, we plot the response curves for imbalance parameter and randomness parameter, with their respective counterpart fixed at optimal. Normalized $\sqrt{\epsilon_{PEHE}}$ on IHDP and ACIC test sample are reported, the lower is better. The upper panel shows sensitivity to imbalance regularization λ (left) and sensitivity to randomness parameter η (right). The lower panel shows respective projections of the under-balanced, proper-balanced, and over-balanced feature distributions. Optimal results, as measured by ϵ_{PEHE}, appear at some moderate level of imbalance and representation randomness. This is consistent with theoretical predictions, because perfectly balanced representation (large λ, Figure 3.4C), compromise the discriminative power of latent representation, while under balanced representation (small λ, Figure 3.4A), are subject to the selection bias.

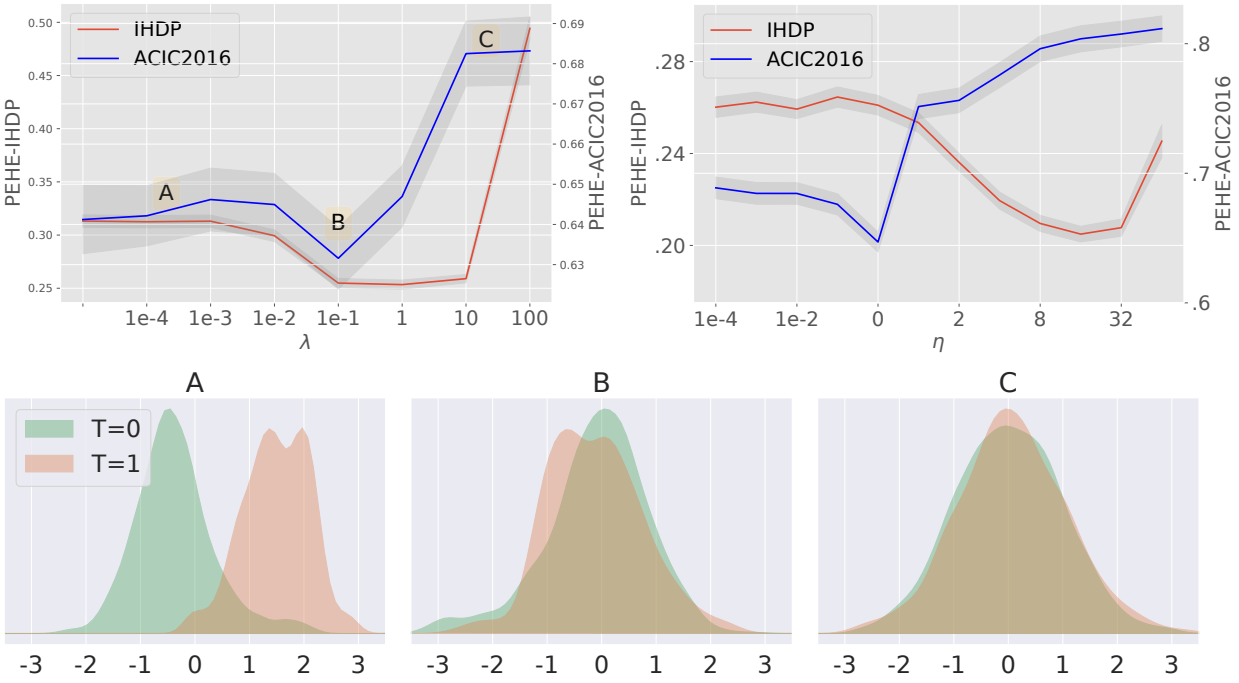

Figure 3.4: Sensitivity analysis of imbalance and randomness in feature representation.

3.5.3 Evaluation on semi-synthetic and real datasets

Table 3.1 summarizes the performance of BV-NICE along with its competing solutions. For both datasets, the proposed BV-NICE performs strongly, giving best results both in terms of in-sample and out-of-sample performance. In Figure 3.5, we plot the mean $\sqrt{\epsilon_{PEHE}}$ computed on ACIC2016 for each simulation type. The dataset index is sorted based on out-of-sample PEHE of BV-NICE. We can see that, with very few exceptions, BV-NICE consistently outperforms its counterparts being compared. These results underscore the importance of modeling representation uncertainty in causal effect estimation. Additionally, we applied BV-NICE to the JOBS dataset, and show the policy risk curve in Figure 3.6. In

the inclusion rate regime $[0.5, 0.9]$, BV-NICE gives significantly lower risks.

Table 3.1: Comparison of performance on semi-synthetic datasets

	IHDP1000		ACIC2016	
$\sqrt{\epsilon_{PEHE}}$	With-in-sample	Out-of-sample	With-in-sample	Out-of-sample
OLS	$0.29 \pm .09$	$0.30 \pm .11$	$0.52 \pm .13$	$0.65 \pm .16$
CFR	$1.47 \pm .35$	$1.46 \pm .36$	$0.52 \pm .14$	$0.90 \pm .26$
BART	$0.30 \pm .08$	$0.33 \pm .11$	$0.58 \pm .12$	$0.70 \pm .17$
Causal RF	$0.63 \pm .01$	$0.63 \pm .16$	$0.68 \pm .01$	$0.81 \pm .02$
BV-NICE	$\mathbf{0.20 \pm .04}$	$\mathbf{0.20 \pm .06}$	$\mathbf{0.50 \pm .13}$	$\mathbf{0.62 \pm .17}$

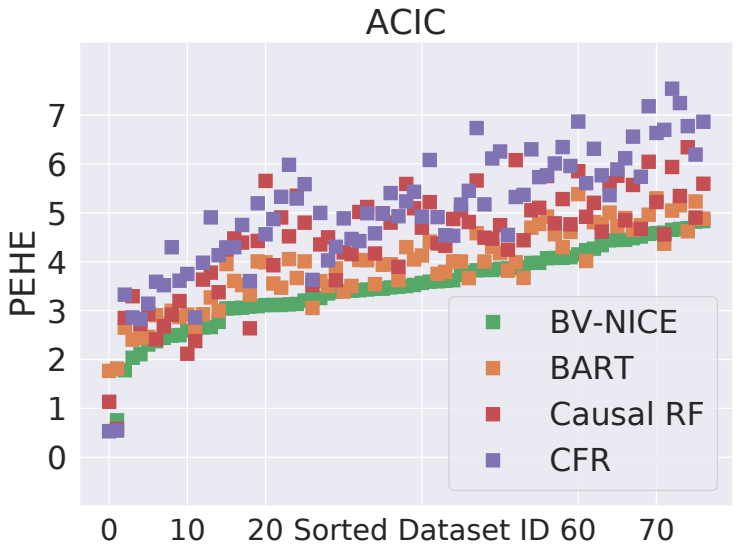

Figure 3.5: BV-NICE result visualization on ACIC2016.

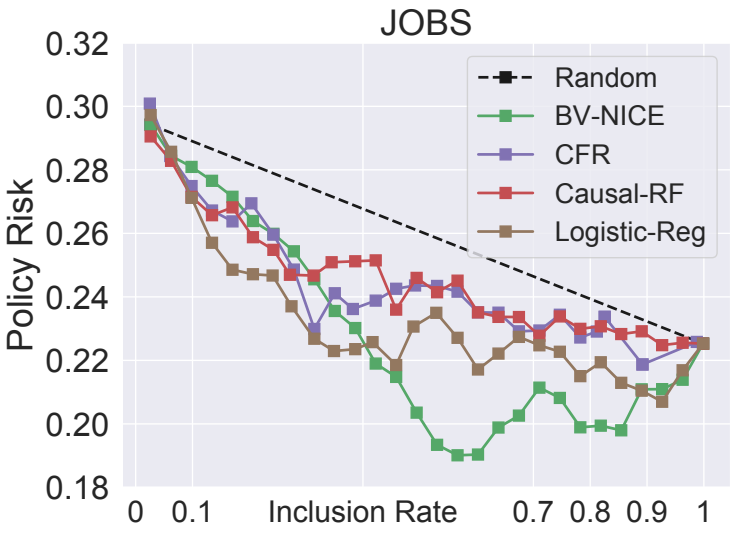

Figure 3.6: BV-NICE result visualization on JOBS.

3.5.4 Traffic safety risk analysis with naturalistic driving data

In our last experiment, we apply the proposed BV-NICE to analyze the risk factors in traffic safety (Dingus et al., 2016a, Klauer et al., 2014b), in the hope that a fine-grained picture of intervention effectiveness can better inform driving safety regulations to reduce the number of tragic events. Note this study is further characterized by the challenge of rare-event modeling, due to the exceptionally low incidence rates of traffic accidents (Guo and Fang, 2013). Only $1k$ crashes were flagged and annotated by trained analysts to represent the potential risk factors, along with $20k$ normal driving baselines for control. Given the prevalence of smart phone usage in modern life, our analysis concentrates on the risk analysis of cellphone use during driving. Following Lu et al. (2020a), 11 variables are included as confounders out of 84 variables originally recorded by the study, with the inclusion criteria

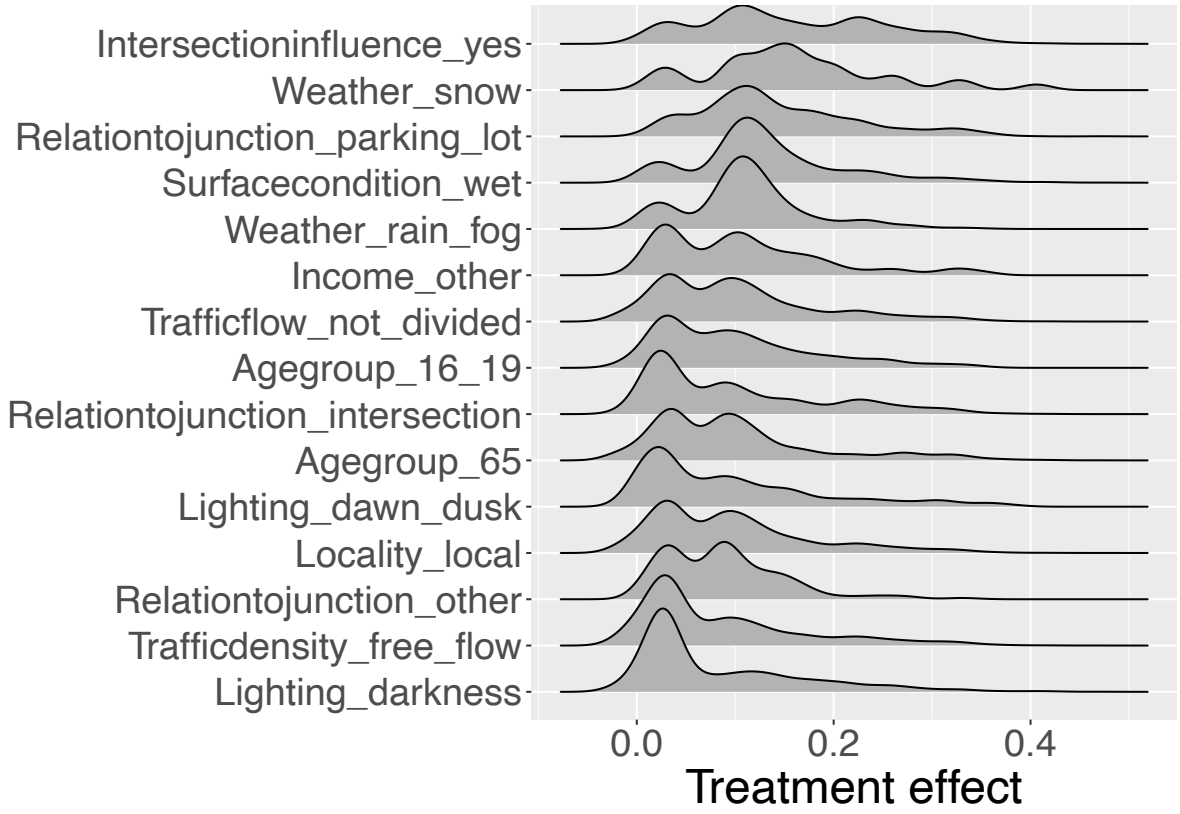

Figure 3.7: Cellphone risk modulation by exogenous factors, larger values imply stronger risk reduction.

derived based on both domain knowledge and statistical independence tests. In Figure 3.7, we visualize how exogenous factors modulates the heterogenous risk distribution of cellphone use, in terms of expected reduction in incidence rate. We see restricting cellphone use is most effective in reducing collisions in bad road conditions (*e.g.*, snowy, wet, rainy, foggy), followed by complex environments (*e.g.*, parking lot crossing, intersections). More statistical summaries and comparisons to alternative causal effect estimators can be found in the Appendix.

3.6 Conclusion

This study revisits design principles for training objective of generative causal models. In particular, we highlight the significance of covariate balancing and uncertainty of representation, which is largely missing from prior investigations. We further present a strong causal inference procedure, called BV-NICE, which bridges R-learning and variational inference. We extensively test our model on realistic datasets, and our results reveal the intricate nature of practical causal estimation procedures. While the empirical performance largely conforms to guiding principles, caution needs to be exercised to avoid the pitfalls, which do not appear in violation of theoretical predictions, yet can severely degrade performance. Further scrutiny is warranted for the study of robust causal estimation with flexible learners, that ameliorates the burden of exhaustive search of parameters.

Chapter 4

Deconfounding through Data Heterogeneity: Invariant Causal Estimation under Population Shifts

In previous chapters we have explored the use of covariate balancing techniques to derive valid causal estimates. In this chapter we take novel perspectives to address issues arising from population shifts, also known as data heterogeneity in many contexts, which can not be satisfactorily resolved using the balancing techniques developed above.

Our development is motivated by the following observations: (i) observational data is often collected from different environments or can be divided into clear subgroups that show distinctive patterns and distributions, as a result, the causal estimates do not generalize; (ii) for cost-effectiveness and statistical efficiency reasons, some study designs (*e.g.*, case-crossover studies) preferentially target populations with specific characteristics (*e.g.*, high-risk groups), so the resulting causal estimand is subjective to sampling bias and does not

extend to the general population. A shared challenge can be readily identified from the two scenarios described above: *under what conditions, and how, can we build causal models that generalize well to distributions that are different from the training data.*

In this chapter, we seek new strategies to enhance such *out of distribution* (OOD) generalizability from a causal invariance perspective, via leveraging the heterogeneity that the data naturally presents. The significance of our work is underscored by the ubiquitousness of such data heterogeneity issues in real-world problems. Importantly, our research applies not only to the observational studies, but also more generally to the gold standard *randomized controlled trials* (RCT). As such, we advocate the view that ascertaining invariance (*i.e.*, causal stability) and establishing balance are equally important prerequisites to claim causal treatment effects. Details follow.

4.1 Challenges of data heterogeneity

A major challenge in machine learning is how to build models that transfer, in the face of data heterogeneity. This is to say, when a model is optimized and calibrated well on a set of training instances, performance optimality does not necessarily carry over to the testing data that shows different characteristics distributionally. Typical scenarios include cases where testing data is collected from another environment, a different experimental setting, or simply a shifted time frame.

In this work, we focus on the goal of making reliable predictions, and more generally robust counterfactual reasonings, when training and testing sets show apparent disparities.

We take particular interest in the settings where training instances can be divided into disparate groups based on criteria as mentioned above, hereafter referred to as environments for convenience. We hope to turn the curse of data heterogeneity into a blessing by introducing causal perspectives to maximally exploit useful information that underlies the overt difference.

Not surprisingly, this problem has been extensively studied within the classical framework of machine learning, under different assumptions (Pan and Yang, 2009). Most researches have focused on the (simplest) case of *covariate shift* (Quionero-Candela et al., 2009), where the discrepancies across environments arise from the shifts in covariate distribution $p_e(x)$, while the conditional $p(y|\boldsymbol{x})$ is assumed to be stable. A widely popular strategy to resolve this problem is sample-based weight adjustment (Shimodaira, 2000, Sugiyama et al., 2007), such that training instances are properly weighted so the modified training distribution $\hat{p}_w(\boldsymbol{x}) = \sum_i w_i \boldsymbol{x}_i$ resembles that for the test case $p_{te}(\boldsymbol{x})$. Standard learning theory dictates the generalization error can be bounded by an additional term relating to the discrepancy between re-weighted training and the test distributions (Ben-David et al., 2010).

A few limitations can be readily identified: (i) estimation of adjustment weights is non-trivial for high dimensional complex distributions (Gretton et al., 2009, Sugiyama et al., 2008); (ii) such solutions can not efficiently handle novel distributions as the test distribution must be known *a priori* (Hendrycks et al., 2020, Volpi et al., 2018); (iii) sensitivity to the violations of the stability assumption for the conditional $p(y|\boldsymbol{x})$ is a major concern (Zhang et al., 2013). The first point motivates the development of representation matching schemes, where instead of employing weight adjustments the learning objective seeks to minimize the

difference between intermediate representation used by the solution (Ganin et al., 2016, Johansson et al., 2019).

To address the last two concerns, *distribution robust optimization* (DRO) has been proposed (Gao et al., 2017, Goh and Sim, 2010, Staib and Jegelka, 2019). Absent the knowledge of test distribution, DRO seeks a mini-max formulation: it tries to find a solution that minimize the worst risk under perturbations to the training distribution $p_{tr}(\boldsymbol{x}, y)$ up to some strength γ, which presumably contains the testing distribution $p_{te}(\boldsymbol{x}, y)$. Many popular solutions fall within this category, such as the adversarially robust learning (Goodfellow et al., 2014b, Madry et al., 2018, Sinha et al., 2018). The major difficulties involved are the determination of type of perturbations and perturbation strength (Cubuk et al., 2019), as DRO is expected to land on overly conservative solutions if very generic choices are made, and consequently yields limited utility. Efforts are still needed to better understand the failure of standard treatment under data heterogeneity.

One theory that has received growing attention in recent years is that machine learning models have over exploited the correlations found in data, without accounting for the peculiarities arising from the data collection procedure (Arjovsky et al., 2019). Short of rigorous experimental control procedures, observational data might be marred by selection biases (Chernozhukov et al., 2018, Schnabel et al., 2016) and (hidden) confoundings (Bellot and van der Schaar, 2020). Such overt or implicit data biases are manifested by apparent shifts in data representation across different data sources as well as correlations that do not generalize well (Ahuja et al., 2020, Wang et al., 2019). While the first point can be partly remedied by the transfer procedures discussed above, the second point undermines the validity of employing standard *empirical risk minimization* (ERM) schemes to identify

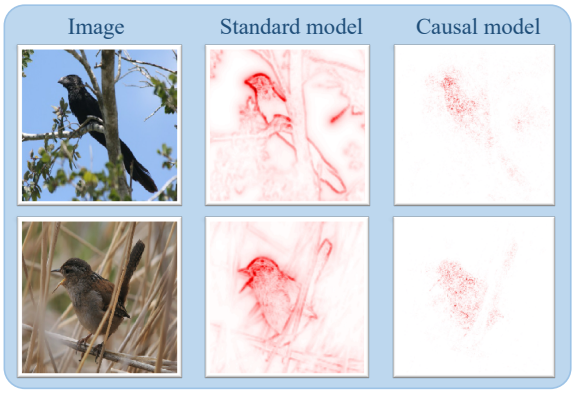

Figure 4.1: Motivating example of a bird classification task, where we have visualized models' decision supporting pixels using salience mapping techniques (Wang et al., 2020). Standard models heavily rely on salient background cues (*e.g.*, trees, reeds) to support their reasoning, which failed to generalize when the subjects are moved out of their natural habitats (*i.e.*, change of environments, such as to an animal shelter). On the other hand, models based on causal invariance principles only attend to features that are causally relevant, as the bird itself.[1]

useful features and relations from data (Vapnik, 1992).

To make things more concrete, let us look at a motivating example (Figure 4.1). Consider bird classification in the wild, where images of birds are taken in their natural habitats, and the goal is to correctly identify the species. In addition to the bird's anatomy that defines its taxonomy, a strong correlation is present between the bird label and background cues (*e.g.*, A-birds are commonly found among branches while B-birds are mostly seen in wetlands). A powerful prediction model, trained naively, will exploit all correlations found in the training data to reason. And it is easier to associate the labels with the more salient visual cues in the background, as is opposed to the subtle details on the bird itself, per principles of Occam's razor and standard learning theory (*i.e.*, a simpler model over similar

[1]Image courtesy of D. Wang, et al.

performance). However, such model will cease to provide reliable predictions in a foreign environment, say a caged bird in captive, as the background has changed and supporting cues are no longer present. On the other hand, a knowledgeable human naturalist is still able to classify accurately based on the bird's anatomy. In this example, we identify the background cues as spurious as they changes with the environment, and the bird's anatomy as stable, invariant causal features that defines the label.

Such inspired, causal invariance emerged as a new, powerful learning principle in recent years, to approach generalization even under apparent distribution shifts (Arjovsky et al., 2019, Peters et al., 2016, Rojas-Carulla et al., 2018, Rothenhäusler et al., 2018, Teshima et al., 2020). Contrast to standard empirical risk minimization based learning, where the generalization to similar scenarios is considered, causally-inspired learning instead embraces robustness against potential perturbations via establishing causation in the model (Schölkopf, 2019). Different from DRO, these models only attend to causally relevant features and associations postulated to be invariant under different settings, informed by the data disparities found in different environments. The reasoning is that unstable relations, also known as *spurious correlations*, are unrelated to the causal explanation (Wang et al., 2019). Only those invariant correlations are expected to be causal, thus enabling generalization beyond training (Muandet et al., 2013). In other words, for better transfer, contributions from spurious features should be effectively blocked.

This study explores the use of invariance principles to identify robust causations that transfer. Our work is motivated by the observation that observational causal conclusions differ considerably for data collected under different experimental designs, and conjectured such sensitivity is inherited from the model's incompetence of dealing with heterogenous in-

puts. This prompts us to reformulate observational causal inference procedures to explicitly focus on invariant causations, which are expected to transcend the population differences due to variations in experimental design and data collection protocols.

The rest of this chapter is organized as follows. In Section 4.2 we introduce the framework of *invariant risk minimization* (IRM) to formalize the idea of learning with invariant features and relations. Based on this, Section 4.3 and Section 4.4 developed invariant causal estimation procedures that are robust to data heterogeneity and population shifts, thereby improving the generalization of causal conclusions. Finally, in Section 4.5, we re-examine the causality in the NDS dataset using these new tools.

4.2 Invariant risk minimization

Problem setup. To motivate our development, we describe a realistic scenario where spurious correlation and data heterogeneity might mislead our analysis, and then formalize the problem mathematically. Consider this in the settings of traffic safety analysis: in the hope to reduce the social-economical cost caused by the tragic events of traffic accidents, we want to build models that (i) quantitatively evaluate a driver's risk of crashing the vehicle under different circumstances; and in particular, (ii) disentangle the contributions from each factor. These results are significant in a number of ways: for (semi-)autonomous driving, it will inform preventive interventions in a timely manner to avert the risk; from a government's perspective, this will help policy makers to revise regularizations and better allocate their investments in infrastructure and driving safety programs.

For that goal, we collect data to support the analysis. We use a set of *covariates* $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^p$ to collectively describe the context (*e.g.*, weather, road condition, traffic, speed, *etc.*) a driver is situated in seconds prior to an potential event, along with the personal characteristics (*e.g.*, age, gender, *etc.*). A binary label $y \in \{0, 1\}$ is also provided indicating whether this data record involves a collision event, which we normally refer to as the *outcome*. In addition, in some situations we take particular interest in a treatment variable $t \in \mathcal{T}$, here assumed binary $\mathcal{T} = \{0, 1\}$ for simplicity, and wonder how the change of the treatment variable alters the risk given the covariates (individual treatment effect $\tau(\boldsymbol{x})$, added to the baseline effect $\mu_0(\boldsymbol{x})$ that without treatment). For instance, the treatment can be whether the subject is currently distracted by a cellphone or not.

To see where the data heterogeneity comes in and how it breeds spurious correlations, we consider a simplified scenario in driving safety analysis, in which the covariate only contains the weather (rain or not), road condition (friction, *i.e.*, how slippery is the road) and the driver's emotional state. Note each participating subjects can be naturally assigned to a label of environment $\mathfrak{e} \in \mathcal{E}$, per which state the subject lives in. Data heterogeneity is then manifested as the variations in the weather pattern, road infrastructure and composition of driver population across different states. While the road friction is expected to provide a universal, invariant characterization of the crash risk, its counterparts weather and driver emotion does not necessarily register a stable association with the collisions, which confusticates analysis:

- **Causal confusion.** The extent to which a rain modifies the road friction ties to the infrastructure, which varies state to state; also the prevalence of rainy days can

differ considerably. For the emotion part, it may be contaminated by human's ability to make accurate predictions into the near feature: a driver may foresee a potential upcoming collision and become either panicked or stimulated, which are likely the consequence rather than the cause of a collision. For a similar argument, certain maneuvers taken before a crash might actually be an instinctive attempt to avoid the tragic event, as opposed to causing a crash.

- **Population shift.** Driver characteristics and road infrastructure vary from state to state, resulting distinct risk profiles across states. Also, despite efforts to make the participating survey population "representative", discrepancies are expected for all sorts of reasons: willingness to volunteer, study design, budget, *etc.* For example, some study designs, such as case crossover studies (see Section 4.4), may disproportionally favor high-risk populations in the survey population, a collateral for the improved statistical efficiency and cost-effectiveness.

These all prevent statistical conclusions made for one state from trivially generalizing to another, or even for the within-state general population that the analysis is supposed to target. For causal effect estimation, it can be expected that such peculiarities will be inherited, and consequently compromise the validity of causal conclusions.

With these issues in mind, we want to devise novel causal estimation procedures that are resilient to such environmental variations, to accommodate stable predictions for heterogeneous populations. Our solution will be formulated under the framework of *invariant risk minimization* (IRM) Arjovsky et al. (2019), elaborated in the sections below. Specifically, we embrace the idea of promoting statistical invariance across environments in our

solution, by recognizing unstable patterns are from non-causal relations, and therefore do not generalize.

For formal development, we will attribute all heterogeneity to the environment, and use $R^{\mathfrak{e}}(h(\boldsymbol{x}[,t]), y)$ to denote the risk function specific to environment $\mathfrak{e}$. We will sometimes refer to environment-independent model $h(\boldsymbol{x}[,t]) : \mathcal{X}[\times\mathcal{T}] \to \mathcal{Y}$ as the hypothesis. The inclusion of treatment t is optional (*i.e.*, given in square brackets) depending on the task of interest. Compactly, we write $R^{\mathfrak{e}}(h)$ for $h \in \mathcal{H}$, and further use subscripted $\mathcal{E}_{tr}$ and $\mathcal{E}_{te}$ to identify training and testing environments, and in general we assume no overlap between $\mathcal{E}_{tr}$ and $\mathcal{E}_{te}$. Here $R^{\mathfrak{e}}(h)$ is typically defined by the standard loss function wrt data distribution, *i.e.*, $R(h) \triangleq \mathbb{E}_{(\boldsymbol{x}[,t],y)\sim p_{\mathfrak{e}}}[\ell(h(\boldsymbol{x}[,t]), y)]$. Typical examples include square loss $\ell(y', y) = (y' - y)^2$ for continuous outcomes and cross-entropy $\ell(\boldsymbol{y'}, y) = -\log p_{\boldsymbol{y'}}(y)$ for categorical outcomes.

4.2.1 Causation, invariance and generalization

To approach the goal of better generalization, we adopt the view promoted by Arjovsky et al. (2019) and consider statistical invariance as the main feature of causation. To clarify the connections and differences to the more traditional definitions of causality, we compare to the classical framework of causal graphs (Pearl, 2009).

First, this invariance view is consistent with the interventional perspective from causal graphs assumed by many causal models. To see this, we equate environment $\mathfrak{e}$ as natural interventions applied to a causal graph. Once an intervention has been administered to a causal graph, its impact will propagate through the links. This will alter the joint distributions between nodes $p(x_i, x_j) \neq p_{\mathfrak{e}}(x_i, x_j)$, resulting unstable correlations. However,

for pair of nodes (x_i, x_j) where a causal link $x_i \to x_j$ exists, the conditional distribution of $p(x_j|x_i) = p_{\mathfrak{e}}(x_j|x_i)$ remains invariant, as long as the node x_j is not directly subjected to the intervention.

Second, this invariance ideology also bypasses some of the conceptual difficulties posed by the causal graphs (Pearl, 2009). In particular, causal graphs require the determination of the direction of causality such as $x_i \to x_j$. However, in many cases, especially when the state of system is governed by a set of equations, such direction of causality is not clearly defined (*e.g.*, $x_i \leftrightarrow x_j$), although stable relations hold nonetheless. Such invariance can be leveraged to sponsor valid inference across environments, albeit not necessarily causal.

Also, in many empirical setting, the identification of causal graph turn out to be an ill-posed problem, while the reliable generalization to novel distributions is of primary interest. The data observed are likely the stochastic proxies to the true interacting entities that do not explicitly appear in our model, so it makes little sense to establish causation among these observables, unless strongly informed by the subject knowledge. Instead, what's more practical and feasible is to find summaries of data that more reliably predicts. Using the image classification example above, while one can not reasonably expect to learn the causal dependencies between the pixels to improve performance, it promises to disentangle the object embeddings from the context embeddings[2] to lend robustness against novel background cues.

[2]That is to identify latent causal units for the label prediction.

4.2.2 Invariant risk minimization (IRM)

In this section, we briefly review the main technique, namely the *invariant risk minimization* (IRM), used in our development. To formalize the concept of learning with invariant correlations, we first define the concept of invariant representation and associated predictors.

Definition 4.2.1. We say a data representation $\Phi : \mathcal{X} \to \mathcal{Z}$ elicits an invariant predictor $w \circ \Phi$ across environments $\mathcal{E}$ if there is a classifier $w : \mathcal{Z} \to \mathcal{Y}$ that is simultaneously optimal for all environments, *i.e.*,

$$w \in \arg\min_{w':\mathcal{Z}\to\mathcal{Y}} \{R^{\mathfrak{e}}(w' \circ \Phi)\} \text{ for all } \mathfrak{e} \in \mathcal{E}. \tag{4.1}$$

The definition above is generic in the sense one can choose to work with arbitrary risk functions $R^{\mathfrak{e}}(h)$ of interest. It also encapsulates the idea of induction in science: distinct but potentially related phenomena, only when described with the correct variables (*i.e.*, the invariant representations), can be shown to obey the same physical laws (*i.e.*, the invariant relations).

Solving for the nested optimization defined above is however, practically infeasible, for the following reasons: (i) Not every representation Φ induces an invariant predictor w; and (ii) How to search for such invariant Φ in the first place? In the context of modern machine learning, one hopes to find a differentiable objective (wrt the model parameters), such that the two goals of finding a good invariant representation and the associated invariant relation can be simultaneously approached.

So motivated, IRM considers relaxed forms that regularize environmental-wise optimal-

ity, which can be optimized in an end-to-end fashion. The heuristic is simple: we solve for a common solution across environments, wrt a representation that is jointly optimized, which allows certain degree of deviation from the optimal solutions of each environment. A general formulation is given by the following regularized objective

$$\min_{\Phi, w} \left\{ \sum_{e \in \mathcal{E}_{tr}} (R^{\mathfrak{e}}(w \circ \Phi) + \lambda \mathbb{D}(w, w_e^*)) \right\}, \tag{4.2}$$

where $\mathbb{D}(w, w')$ denotes a user specified discrepancy measure defined on the solution space, and w_e^* is the environmental specific optimal solution, and $\lambda \geq 0$ specifies the regularization strength.

Before presenting the final IRM solution, some remarks are in order. Note that the choice of discrepancy measure $\mathbb{D}(w, w')$ matters greatly, albeit appears arbitrary. Take linear least-squares regression as an example, one can either choose the intuitive Euclidean distance $\mathbb{D}_{L^2}(w, w_e) = \|\boldsymbol{w} - \boldsymbol{w}_e\|_2^2$ or the deviance from the normal equations $\mathbb{D}_N(w, w_e) = \|\mathbb{E}_{\mathfrak{e}}[(X^{\mathfrak{e}})^T X^{\mathfrak{e}}]\boldsymbol{w} - \mathbb{E}_{\mathfrak{e}}[(X^{\mathfrak{e}})^T Y^{\mathfrak{e}}]\|_2^2$. Arjovsky et al. (2019) showed characteristically different optimization landscapes between the two choices, where the former showed irregularities near the invariant solution using the toy example presented. For complex models, direct comparison of model parameters also makes little sense as the same model can be parameterized differently. Additionally, the discrepancy needs to be efficiently evaluated, so one should not expect to use measures such as $\mathbb{D}(w, w^*) = \|h_w - h_{w^*}\|_{L^p(X)}$. In many scenarios, the estimation of w_e^* itself may construct a challenging problem on its own, which ideally should be avoided. Also, even for the simplest linear models, w and Φ one may encounter indeterminacies that lead to arbitrary solutions if no additional restrictions on the solution

has been imposed.

To resolve the above issues, IRM developed mathematics showing that for linear model it suffices to consider rank-1 representations and a dummy predictor $w = 1$. The final proposal boils down to the following convenient expression

$$\min_{h:\mathcal{X}\to\mathcal{Y}} \left\{ \sum_{e\in\mathcal{E}_{tr}} (R^{\mathfrak{e}}(h) + \lambda \underbrace{\|\nabla_{w|w=1} R^{\mathfrak{e}}(w\cdot h)\|^2}_{\text{Environmental optimality } \Omega^{\mathfrak{e}}(h)}) \right\}. \tag{4.3}$$

Note here IRM no longer explicitly optimizes for the invariant representation Φ. Instead, it treats the model output $h(\boldsymbol{x})$ as the representation and makes a dummy predictor $w = 1$. In a sense IRM has fixed the predictor to avoid indeterminacies and only optimizes for the representation, which turned out to be the model itself. What's non-trivial here is that IRM rewrites the environmental optimality penalty $\mathbb{D}(h, h^{\mathfrak{e}})$[3] as $\Omega^{\mathfrak{e}}(h) = \|\nabla_{w|w=1} R^{\mathfrak{e}}(w\cdot h)\|^2$, whose mathematics is detailed in Sec 4.2.3. Heuristically, $\Omega^{\mathfrak{e}}$ is a generalization of the normal equation violation metric $\mathbb{D}_N(w, w^{\mathfrak{e}})$ defined for an arbitrary risk function $R^{\mathfrak{e}}(h)$. What makes $\Omega^{\mathfrak{e}}$ especially appealing is that the environmental optimality discrepancy[4] is directly vetted against current solution h itself, and does not involve its environment optimal counterpart $h_e^\star$. See Algorithm 2 for implementation details.

To demonstrate that IRM identifies stable causal pattern and is less prone to the bias from data heterogeneity, we consider a minimal heterogeneous toy data from the model

[3] We adapt the notation according to Eq. (4.3).
[4] Defined wrt current representation.

Algorithm 2 IRM

Input: Feature, label and environment triplet $\{(\boldsymbol{x}_i, y_i, \mathfrak{e}_i)\}_{i=1}^n$,
loss function $\ell(\hat{\boldsymbol{y}}, y)$, regularization $\lambda \in [0, 1)$
Parameter: Model parameter θ for $\hat{\boldsymbol{y}} = h_\theta(\boldsymbol{x})$, pseudo-parameter $\omega \leftarrow 1$
for $l = 1, 2, \cdots$ **do**
  Risk $\mathcal{R} = 0$, environment penalty $\Omega = 0$
  for $\mathfrak{e} = 1, \cdots, E$ **do**
    Sample mini-batch $\{(\boldsymbol{x}_{i_b}^{\mathfrak{e}}, y_{i_b}^{\mathfrak{e}})\}_{b=1}^B$ such that $\mathfrak{e}_{i_b} = \mathfrak{e}$
    $\mathcal{R}^{\mathfrak{e}} = B^{-1} \sum_{i_b} \ell(\omega \cdot h_{\theta^l}(\boldsymbol{x}_{i_b}^{\mathfrak{e}}), y_{i_b}^{\mathfrak{e}})$, $\Omega^{\mathfrak{e}} = \|\nabla_\omega \mathcal{R}^{\mathfrak{e}}\|^2$
    $\mathcal{R}$ += $\mathcal{R}^{\mathfrak{e}}, \Omega$ += $\Omega^{\mathfrak{e}}$
  end for
  $\partial\theta^l = \nabla_\theta\{(1-\lambda)\mathcal{R} + \lambda\Omega\}$, $\theta^{l+1} \leftarrow$ `Optimizer`$(\theta^{1:l}, \partial\theta^{1:l})$
end for

below:

$$\text{(Heterogeneous)} \quad \text{Toy A:} \quad X_1 \overset{\mathfrak{e}}{\to} X_2, \quad X_2 \to Y, \quad Y \overset{\mathfrak{e}}{\to} X_3, \tag{4.4}$$

$$\text{(Homogeneous)} \quad \text{Toy B:} \quad X_1 \to X_2, \quad X_2 \to Y, \quad Y \to X_3. \tag{4.5}$$

In the heterogeneous example Toy A, X_1 is an unstable cause of X_2, with its impact depending on environment $\mathfrak{e}$, X_2 is a stable cause of binary outcome variable Y, and Y is a environment dependent unstable cause of X_3. In this setup, the outcome Y is the result of causal feature X_2 and is spuriously correlated with upstream feature X_1 and downstream feature X_3. Each link in the graph is implemented as a simple linear regression model. The goal is to reliably predict Y from the feature set (X_1, X_2, X_3).

We compare IRM with popular off-the-shelf solutions: Logistic regression (LR), LASSO regression (Tibshirani, 1996), Ridge regression (Hoerl and Kennard, 1970), and *linear mixed effect model* (LME) (Agresti, 2003) with environment $\mathfrak{e}$ as random effect. For comparison,

we also generate toy data with no environment heterogeneity from Eq. 4.5 for comparison. In both toy data, we set the ground truth causal model for Y as $\texttt{Logit}(\mathbb{P}(y = 1)) = 0 \cdot x_1 + 1 \cdot x_2 + 0 \cdot x_3$.

The model performance is evaluated with two metrics: (i) the coefficients bias for each feature and (ii) the *negative* log-*likelihood* (NLL), respectively on in-sample and out-sample data. Figure 4.2 and Table 4.1 show that in case of heterogeneous data, IRM outperforms its counterparts and more accurately identifies the true relations. In this example (Toy A), LR and LME produce compromised estimates for true causal feature X_2 (with bias equals to 0.03 and 0.05 respectively) while consider X_3 as a significant predictor (coefficient estimates ≈ 0.3) with. LASSO and Ridge, which penalize the model complexity using L_1 and L_2 norms, suffer from the so-called regularization bias by wrongly picking unstable feature X_3 while completely discarding the true causal feature X_2, leaving the bias close to 1. In terms of prediction performance, IRM also outperforms other competing baselines with a lower NLL. In the homogeneous case (ToyB), IRM delivered competitive performance .

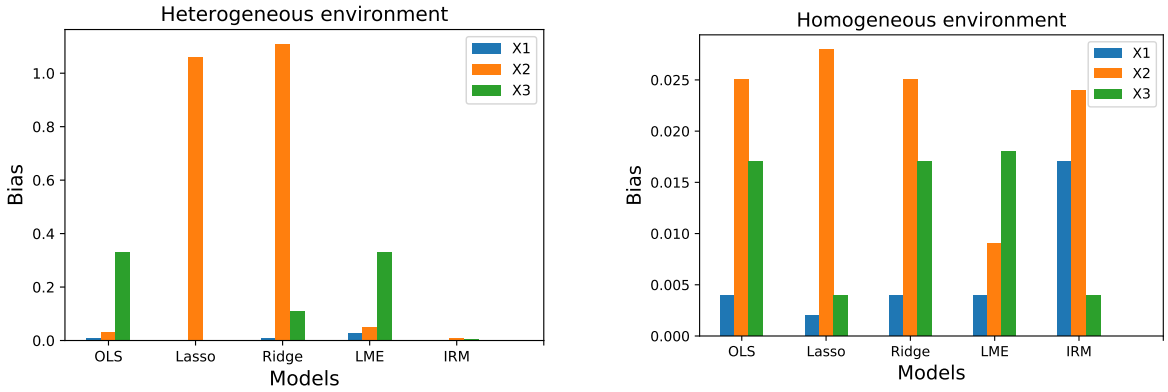

Figure 4.2: Comparison of coefficient bias with and without environment heterogeneity

Table 4.1: Invariant effect estimation

		Bias ↓			NLL ↓	
	Method	X_1	X_2	X_3	in-sample	out-sample
ToyA	LR	0.010	0.031	0.327	**0.648**	0.695
	LASSO	**0.000**	1.062	**0.003**	0.686	0.687
	Ridge	0.008	1.108	0.110	0.669	0.684
	LME	0.027	0.052	0.331	1.010	1.078
	IRM	**0.000**	**0.009**	0.004	0.660	**0.661**
ToyB	LR	0.004	0.025	0.017	**0.648**	0.649
	LASSO	**0.002**	0.028	**0.004**	0.649	0.652
	Ridge	0.004	0.025	0.017	0.648	0.649
	LME	0.004	**0.009**	0.018	1.000	1.044
	IRM	0.017	0.024	**0.004**	0.660	**0.645**

4.2.3 Theory of IRM

This section serves to clarify the deep mathematics behind IRM. Without delving too much into technical details, we try to elucidate (*i*) how does $\Omega^{\mathfrak{e}}$ relate to environment optimality; and (*ii*) under what conditions does IRM learn invariance that generalize. The following theorem explains the environmental invariance penalty term $\Omega^{\mathfrak{e}}$.

Theorem 4.2.2 (Theorem 4, Arjovsky et al. (2019))**.** *For all $\mathfrak{e} \in \mathcal{E}$, let $R^{\mathfrak{e}} : \mathbb{R}^d \to \mathbb{R}$ be convex differentiable cost functions. A vector $\boldsymbol{v} \in \mathbb{R}^d$ can be written $\boldsymbol{v} = \Phi^\top \boldsymbol{w}$ where $\Phi \in \mathbb{R}^{p \times d}$, and where $\boldsymbol{w} \in \mathbb{R}^p$ simultaneously minimize $R^{\mathfrak{e}}(w \circ \Phi)$ for all $\mathfrak{e} \in \mathcal{E}$, if and only if $\boldsymbol{v}^T \nabla R^{\mathfrak{e}}(\boldsymbol{v}) = 0$ for all $\mathfrak{e} \in \mathcal{E}$.*

To understand the above theorem, we consider $\Phi \in \mathbb{R}^{n \times p}$ as the representation of data, where n is the sample dimension and p is the feature dimension, and $\boldsymbol{w} \in \mathbb{R}^p$ is a linear predictor. Note Φ is possibly derived through some nonlinear transformation $\Phi(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}^p$. Then $\boldsymbol{v} = \Phi^T \boldsymbol{w}$ is the predicted value that is fed into the loss function (*e.g.*, in the case of linear least squares regression, $R = \|\Phi^T \boldsymbol{w} - \boldsymbol{y}\|^2$). Theorem 4.2.2 basically says, if Φ is an invariant representation such that the predictor $\boldsymbol{w}$ is simultaneously optimal across all environments $\mathfrak{e} \in \mathcal{E}$, then $\boldsymbol{v}^T \nabla R^{\mathfrak{e}}(\boldsymbol{v}) = 0$ for all $\mathfrak{e} \in \mathcal{E}$. In other words, $\boldsymbol{v}^T \nabla R^{\mathfrak{e}}(\boldsymbol{v}) = 0$ is a necessary condition, hereafter known as the *optimality condition*, for $(\Phi, \boldsymbol{w})$ to be an invariant representation and predictor pair. To connect this to $\mathcal{R}_{\mathcal{E}}$, a key observation is that

$$\nabla_{w|w=1} R^{\mathfrak{e}}(w \cdot \boldsymbol{v}) = \boldsymbol{v}^T \nabla R^{\mathfrak{e}}(\boldsymbol{v}), \tag{4.6}$$

so $\Omega^{\mathfrak{e}} = \|\boldsymbol{v}^T \nabla R^{\mathfrak{e}}(\boldsymbol{v})\|^2$. Then the IRM algorithm translates into finding a solution that

penalizes for the violation of optimality condition $\boldsymbol{v}^T \nabla R^{\mathfrak{e}}(\boldsymbol{v}) = 0$ in each environment.

As a concrete example, consider the least square regression, where the risk is defined by the square loss $\ell(y', y) = (y' - y)^2$, *i.e.*,

$$R^{\mathfrak{e}} = \mathbb{E}_{\mathfrak{e}}[\|\boldsymbol{X}^{\mathfrak{e}}\boldsymbol{w} - Y^{\mathfrak{e}}\|^2]. \tag{4.7}$$

After some standard algebraic manipulations, it is easy to see that

$$\boldsymbol{v}^T \nabla R^{\mathfrak{e}}(\boldsymbol{v}) = \mathbb{E}_{\mathfrak{e}}[(\boldsymbol{X}^{\mathfrak{e}})^T \boldsymbol{X}^{\mathfrak{e}}]\boldsymbol{w} - \mathbb{E}_{\mathfrak{e}}[(\boldsymbol{X}^{\mathfrak{e}})^T Y^{\mathfrak{e}}] \tag{4.8}$$

Recall the optimal least squares predictor for each environment is given by

$$\boldsymbol{w}^{\mathfrak{e}} = \mathbb{E}_{\mathfrak{e}}[(\boldsymbol{X}^{\mathfrak{e}})^T \boldsymbol{X}^{\mathfrak{e}}]^{-1}\mathbb{E}_{\mathfrak{e}}[(\boldsymbol{X}^{\mathfrak{e}})^T Y^{\mathfrak{e}}], \tag{4.9}$$

so the optimality condition $\Omega^{\mathfrak{e}} = 0$ holds if $\boldsymbol{w} = \boldsymbol{w}^{\mathfrak{e}}$ for all $\mathfrak{e} \in \mathcal{E}$.

Another key question is, under what conditions does the invariant result generalize to all environments of interest. Intuitively, to learn useful invariance, sufficient degree of diversity must be encoded in the training environments. On the other hand, arbitrarily splitting based on random choice variables does not work, as such artificial diversity may further introduce spurious association that destroy the invariance of interest. Arjovsky et al. (2019) described one such scenario where generalization is guaranteed for linear models, which we informally summarize below.

For tractability, IRM assumes the following model:

$$Y^{\mathfrak{e}} = Z_1^{\mathfrak{e}} \cdot \boldsymbol{\beta} + \epsilon^{\mathfrak{e}}, \quad Z_1^{\mathfrak{e}} \perp\!\!\!\perp \epsilon^{\mathfrak{e}}, \quad \mathbb{E}[\epsilon^{\mathfrak{e}}] = 0, \quad X^{\mathfrak{e}} = (Z_1^{\mathfrak{e}}, Z_2^{\mathfrak{e}}) \cdot S. \tag{4.10}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{d\times 1}$ and $Z_1 \in \mathbb{R}^{1\times d}, Z_2 \in \mathbb{R}^{1\times q}$. The key concept to ensure the generalization of invariance is the so called *linear general positions*.

Definition 4.2.3 (Linear general position)**.** A set of training environment $\mathcal{E}_{tr}$ lie in a linear general position of degree r if $|\mathcal{E}_{tr}| > d - r + \frac{d}{r}$ for some $r \in \mathbb{N}$, and for all non-zero $\boldsymbol{x} \in \mathbb{R}^{d\times 1}$:

$$\text{Dim}\left\{\text{Span}\left\{\mathbb{E}_{\mathfrak{e}}\left[(X^{\mathfrak{e}})^T X^{\mathfrak{e}}\right]\boldsymbol{x} - \mathbb{E}_{\mathfrak{e}}\left[(X^{\mathfrak{e}})^T \epsilon^{\mathfrak{e}}\right]\right\}\right\} > d - r \tag{4.11}$$

This definition encodes the heuristic notation of co-linearity for the training environments, with each new environment laying in linear general position removes one degree of freedom in the space of invariant solutions.

Theorem 4.2.4 (Generalization of IRM, Theorem 9, Arjovsky et al. (2019))**.** *Assume there exists $\tilde{S}$ such that $(X^{\mathfrak{e}})^T \tilde{S} = Z_1^{\mathfrak{e}}$, for all environments $\mathfrak{e} \in \mathcal{E}$. Let $\Phi \in \mathbb{R}^{d\times d}$ has rank $r > 0$. Then if at least $d - r + \frac{d}{r}$ training environments $\mathcal{E}_{tr} \subseteq \mathcal{E}$ lie in a linear general position of degree r, we have that*

$$\Phi\mathbb{E}_{\mathfrak{e}}\left[(X^{\mathfrak{e}})^T X^{\mathfrak{e}}\right]\Phi^T w = \Phi\mathbb{E}_{\mathfrak{e}}\left[(X^{\mathfrak{e}})^T Y^{\mathfrak{e}}\right] \tag{4.12}$$

holds for all $\mathfrak{e} \in \mathcal{E}_{tr}$ iff Φ elicits the invariant predictor $\Phi^T w$ for all $\mathfrak{e} \in \mathcal{E}$.

The key assumptions involved in the above theorem are linearity, centered noise and the independence between the noise $\epsilon^{\mathfrak{e}}$ and the causal variables $Z_1^{\mathfrak{e}}$. Note spurious correlations

can be expected between $Z_2^{\mathfrak{e}}$ and $\epsilon^{\mathfrak{e}}$. Theorem 4.2.4 is interesting in that it does not require Gaussian assumptions and the causal variables are defined as a latent reprsentation Z_1. It also shows for a high-rank representation Φ it requires less environment to generalize, which is a good thing. Empirically, a few environments with sufficient diversity generally suffices.

4.3 RICE: Risk Invariant Causal Estimation

4.3.1 Treatment effect under heterogeneity: issues and paradoxes

Now we want to extend our discussions connect the main theme of this thesis: the estimation of (individualized) causal effects in an observational setup with the presence of (severe) data heterogeneity. Recall in causal effect estimation, the overarching objective is to de-bias the confounding effects from those variables $\boldsymbol{x}$ that simultaneously affect the assignment of treatment t and outcome y, in order to accurately characterize the treatment effect $\tau(\boldsymbol{x}) = \mathbb{E}[Y_1(\boldsymbol{x}) - Y_0(\boldsymbol{x})]$. For an extensive coverage on the observational causal inference in standard settings, we refer readers to Chapter 3. Here we focus on the new challenges arising from data heterogeneity and seek novel causal estimators that generalize better.

To understand why data heterogeneity undermines the validity of standard observational causal inference procedures, we first revisit the two key technical assumptions that makes causal effects identifiable from observational data: respectively the assumptions of *unconfoundedness*:$\{Y(0), Y(1)\} \perp\!\!\!\perp T|\boldsymbol{X}$, and *positivity*: $p(T|\boldsymbol{X}, Y(0), Y(1)) \in (0, 1)$ (Rosenbaum and Rubin, 1983a). Realistically, it is hard to ascertain the posited unconfoundedness,

due to noisy observations and potential hidden confounders, resulting residual confounding whose removal is technically infeasible.

Taking an interventional perspective from the last section, environment exerts natural interventions to the causal graph, causing tangible distribution shifts and changes in correlation structure. Since the exact mechanism remains unknown to the observer (*i.e.*, hidden confounding) and it often aggravates the overlap issue, standard strategies adopted by observational causal inference become inadequate. In a sense, treatment effect model is more vulnerable to data heterogeneity compared to prediction models, as such negative impacts can be amplified by adjustment procedures.

Specifically, recall that standard approaches build causal estimators based on the outcome models $\hat{p}(y|\boldsymbol{x}[,t])$, possibly adjusted by the propensity score estimator $\hat{p}(t|\boldsymbol{x})$, which we typically referred to as the *nuisance models*. For the reasons discussed in previous sections, a naive causal estimator built in such a way will inherit the susceptibility to spurious correlations and all other data peculiarities absorbed by the nuisance components. Such issues raise concerns for the credibility of causal estimations, and consequently prohibit reliable extrapolation to novel environments and populations.

As a concrete example, we provide a case where unprotected causal effect estimation under data heterogeneity leads to paradoxical conclusions. Details of this numerical example is provided in Section 4.3.4. In Figure 4.3 (left), we show a causal estimator based on simply pooling all data from different environments, a standard strategy adopted by many empirical researches that destroys useful information on the environmental heterogeneity. This naive estimator predicts a trend opposite to the ground-truth causal effect size, which gives an

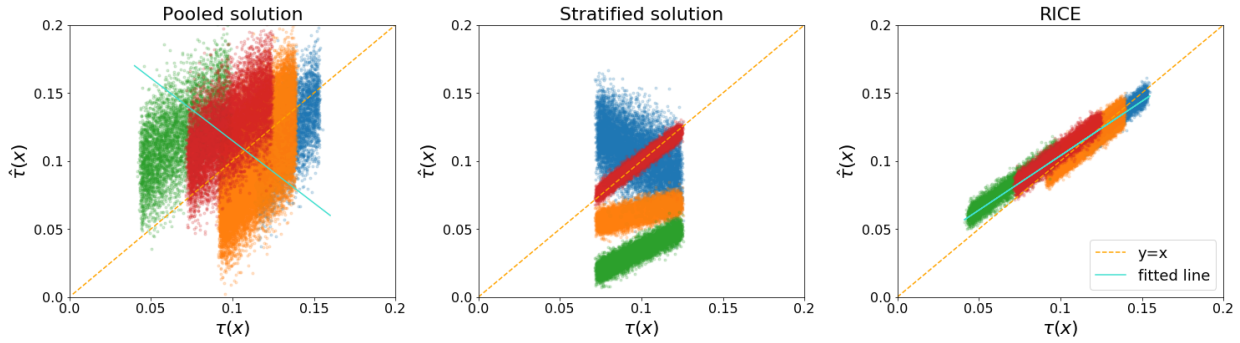

Figure 4.3: Simpson paradox of causal effect estimation under data heterogeneity. We plot the estimated causal effects against the ground-truth causal effects. Due to environmental heterogeneity (color coded), causal estimator based on pooled data from all environments gives a paradoxic opposite trend (left), while the stratified estimator does not generalize to other environments (middle). The only reasonable solution is given by the proposed risk invariant causal estimator (right).

interesting parallel to the classical *Simpson paradox*.

The potential non-randomized treatment assignment mechanism leading to confounding, insufficient overlap, and heterogeneous sub-population, thus estimating treatment effect using such imbalanced data is challenging and needs carefully pretreatment (Rubin, 1974a). Early work on heterogeneous treatment effect estimation mainly focused on predefined sub-groups, *i.e.*, stratified analysis using predefined major feature or summarized propensity score. Figure 4.3 (middle) highlights the failures of such strategy when we want to move beyond the training population.

To overcome such difficulties, our key insight is that data heterogeneity constructs a special manifestation of the unobserved confounding, where partial information has been revealed, so the identification still holds promise. In other words, while an inquiring investigator still have no knowledge of the unobserved confounders to de-bias the solution in standard ways, extra information on when these hidden confounders is expected to vary

becomes available, and this may help to gauge a more reliable estimate for the effect of interest.

We recognize the silvering lining here is that establishing invariance in causal estimation promises a new outlet. Instead of capitalizing on rectifying the unconfoundedness issue (possibly through collecting more variables or carrying out sensitivity analysis), we appeal to the establishment of statistical invariance to project out factors that do not generalize. Our key observation is that standard empirical risk minimization based estimation can not effectively decouple correlation and causation from observational data. As such, strong non-causal correlates often obfuscate the solutions. Such issues can be remedied by cleverly exploiting the environmental heterogeneity under the invariant risk minimization framework.

Comparison can be made to the framework of *instrument variable* (IV) regression, where variables orthogonal to the outcome yet correlated with the treatment are exploited to eliminate confounding effects. Note that IV regression typically involves strong subject knowledge, and does not readily extend to the settings where more generic risk definitions are required.

4.3.2 Invariant treatment effect estimation

To enable counterfactual reasoning under the causal invariance constraints, we propose to bind IRM to the R-learning framework for causal effect estimation that generalize (Figure 4.4). We seek to improve the robustness of causal estimation in the presence of data heterogeneity, as well as improving the generalizability of our causal estimator to novel environments and populations.

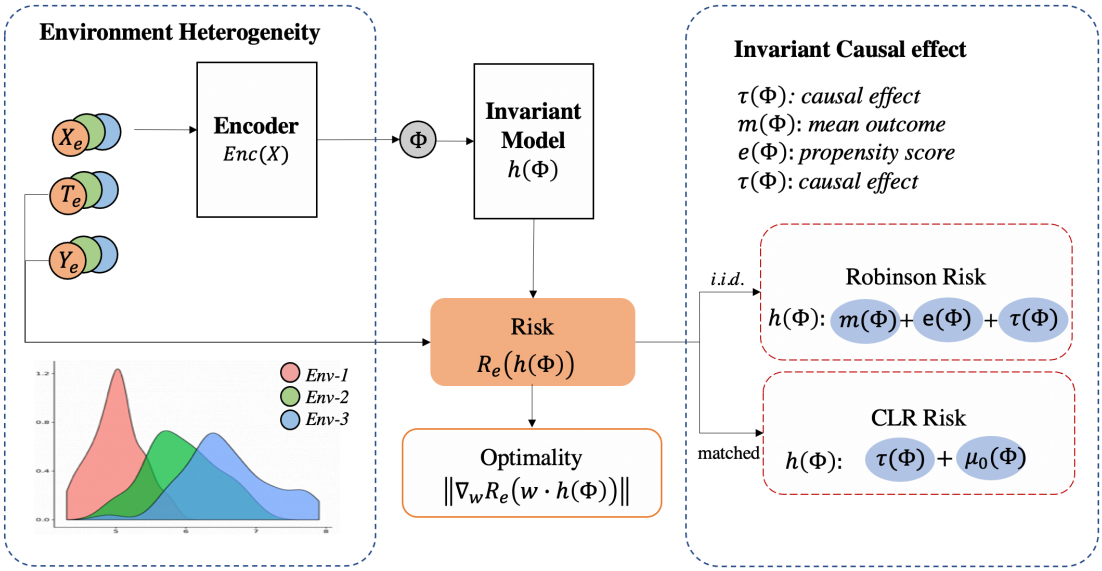

Figure 4.4: RICE model architecture.

The key assumption we have adopted here is that the causal effect of interest is stable across different environments. That is to say, variations in the appeared disparities in environment-specific estimates are likely to be caused by spurious correlations between the covariate and outcome/treatment. Alternatively, even if there exists variations in the causal effect across different environments, we are only interested in the stable effects that can be reliably transferred to different environments.

Our setup follows similarly to that described in Chapter 3, with the newly included environmental label $\mathfrak{e}$ in addition to the covariate, treatment and outcome triplet $(\boldsymbol{x}, t, y)$. Specifically, let $(e(\boldsymbol{x}), m(\boldsymbol{x}), \tau(\boldsymbol{x}))$ be the models for propensity score, mean outcome and causal effect respectively. Then the Robinson residual for data point i subsequently writes

$$\epsilon_i \triangleq y_i - m(\boldsymbol{x}_i) - (t_i - e(\boldsymbol{x}_i))\tau(\boldsymbol{x}_i). \tag{4.13}$$

This allows us to write the risk functions for environment $\mathfrak{e}$ as

$$R_\tau^{\mathfrak{e}}(\tau; m, e) = \sum_{i \in e} \epsilon_i^2, \;\; R_y^{\mathfrak{e}}(m) = \sum_{i \in e} (m(\boldsymbol{x}_i) - y_i)^2, \;\; R_t^{\mathfrak{e}}(e) = \sum_{i \in e} \mathcal{L}_{CE}(e(\boldsymbol{x}_i), t_i), \tag{4.14}$$

where we have used $\mathcal{L}_{CE}$ to denote the cross-entropy loss. Note that here we have made the assumption that the outcome, propensity score and treatment effect models are stable across all environments $\mathcal{E}$. This assumption can be relaxed if subject knowledge informs otherwise.

Since our main goal is to estimate the treatment effect $\tau(x)$, we will consider the outcome and propensity score models $(m(\boldsymbol{x}), e(\boldsymbol{x}))$ as nuisance components and perform a two stage learning for causal effect $\tau(\boldsymbol{x})$. In the first stage, we solve the regularized IRM for the mean outcome and propensity models respectively

$$\hat{m}_\lambda \leftarrow \arg\min_{m: \mathcal{X} \to \mathcal{Y}} \left\{ \sum_{\mathfrak{e} \in \mathcal{E}_{tr}} \left((1-\lambda) R_y^{\mathfrak{e}}(m) + \lambda \| \nabla_{w|w=1} R_y^{\mathfrak{e}}(w \cdot m) \|^2 \right) \right\}, \tag{4.15}$$

$$\hat{e}_\lambda \leftarrow \arg\min_{e: \mathcal{X} \to \mathcal{T}} \left\{ \sum_{\mathfrak{e} \in \mathcal{E}_{tr}} \left((1-\lambda) R_t^{\mathfrak{e}}(e) + \lambda \| \nabla_{w|w=1} R_t^{\mathfrak{e}}(w \cdot e) \|^2 \right) \right\}. \tag{4.16}$$

The optimal regularization parameters λ_m^* and λ_e^*, and the resulting optimal estimates $(\hat{m}^*, \hat{e}^*)$ are determined by the invariance-based environment-wise cross-validation procedure detailed in Sec 4.3.3. In the second stage, we fix the nuisance model estimates and solve IRM for $\tau(\boldsymbol{x})$

$$\hat{\tau}_\lambda \leftarrow \arg\min_{\tau: \mathcal{X} \to \mathcal{Y}} \left\{ \sum_{\mathfrak{e} \in \mathcal{E}_{tr}} \left((1-\lambda) R_\tau^{\mathfrak{e}}(\tau) + \lambda \| \nabla_{w|w=1} R_\tau^{\mathfrak{e}}(w \cdot \tau) \|^2 \right) \right\}, \tag{4.17}$$

and again use the cross-environment validation to determine the best model estimate $\hat{\tau}^*(\boldsymbol{x})$.

4.3.3 Cross Validation

We now describe how to use cross-validation to tune the regularization strength parameter λ. As we are seeking generalization across environments, standard random sample splits would be inappropriate for IRM. Instead, we consider cross environment predictions: for each split, we partition the environments into two non-overlapping sets, with the one set used for training and the other set of hold-out environments retained for validation. To get our final solution, we average over the hold-out risks for all CV splits, determine the optimal regularization strength $\hat{\lambda}^*$ as the one gives the minimal risk score, and use $\hat{\lambda}^*$ to derive an estimate:

$$\hat{g}_{k,\lambda} = \arg\min_g \left\{ \sum_{e \in \mathcal{T}_k} R^{\mathfrak{e}}(g) + \lambda \mathcal{R}^{\mathfrak{e}}(g) \right\}, \quad e(\lambda) = \mathbb{E}_k \left[\sum_{e \in \mathcal{V}_k} R^{\mathfrak{e}}(\hat{g}_{k,\lambda}) \right[, \tag{4.18}$$

$$\hat{\lambda}^* = \arg\min_\lambda \{e(\lambda)\}, \quad \hat{g} = \arg\min_g \left\{ \sum_{e \in \mathcal{E}} R^{\mathfrak{e}}(g) + \hat{\lambda}^* \mathcal{R}^{\mathfrak{e}}(g) \right\}. \tag{4.19}$$

Early stopping and internal validation. While the above procedure helps to gauge the cross-environment generalization performance, we still need an objective criteria to terminate the learning process when appropriate. This is critical especially in the low-sample and complex model regime where over-fitting is of concern. For that purpose, we divide the samples in each training environment ($\mathbf{D}^{\mathfrak{e}}$) into a training set $\mathbf{D}^{\mathfrak{e}}_{tr}$ and a validation set $\mathbf{D}^{\mathfrak{e}}_{val}$, and similarly define the risks associated with these respective sets as $R^{\mathfrak{e}}_{tr}(g)$ and

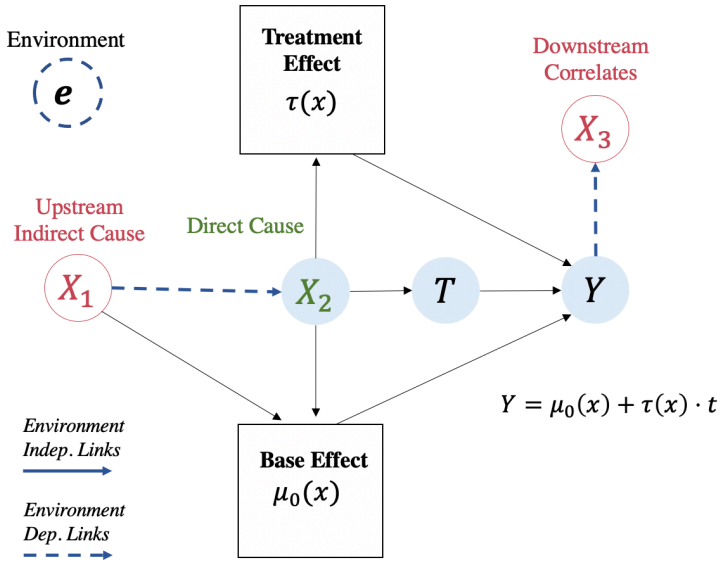

Figure 4.5: RICE toy model

$R^{\mathfrak{e}}_{val}(g)$. Let $\{\hat{g}_t\}$ denote the sequence of solutions found by an arbitrary optimizer to solve Eq. (4.3), where t indexes the iterations. Then the solution is given by

$$\hat{g} = g_{t^*}, \qquad \text{where } t^* = \arg\min_t \left\{ \sum_e R^{\mathfrak{e}}_{val}(g_t) \right\}. \tag{4.20}$$

In practice, we stops training and returns g_t when $R_{val}(g_t)$ cease to improve.

4.3.4 Experiments

Toy model. To realistically simulate challenges from real-world data, we design the following toy model. Without loss of generality, we have $\boldsymbol{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$ as our covariates for the estimation of individualized treatment effect $\tau(\boldsymbol{x})$ wrt binary outcome y. In particular,

the toy data is generated from the following causal graph (see also Figure 4.5)

$$x_1 \overset{\mathfrak{e}}{\to} x_2, \quad x_2 \overset{\mathfrak{e}}{\to} t, \quad (x_1, x_2) \to \mu_0, \quad (\mu_0, \tau) \to y, \quad y \overset{\mathfrak{e}}{\to} x_3, \tag{4.21}$$

where we have used $\overset{\mathfrak{e}}{\to}$ to denote environment-dependent causal links and $\to$ for invariant causal links. In this model, (x_1, x_3) constructs spurious features that are correlated with y. Specifically, x_1 is an upstream non-direct cause of the outcome, while x_3 is a downstream descendent of the outcome. These two construct the most common sources of correlations we see in practice that is non-causal[5].

To simulate environment heterogeneity, we create four artificial environments, each with their unique parameterization of the unstable non-causal links. To lend further authenticity to our simulation, we also add environment imbalance, drawing samples from different environments with skewed probability $p_{\mathcal{E}} = [0.2, 0.1, 0.6, 0.1]$. σ is the Sigmoid function, $\mathcal{U}, \mathcal{B}$ respectively denotes the uniform and Bernoulli random number generator.

$$\mathfrak{e} \sim \mathcal{U}(p_{\mathcal{E}}), \quad x_1 \sim \mathcal{U}([16, 30]), \quad x_2 \sim 5 \cdot \sigma(c_2^{\mathfrak{e}} \cdot x_1) + \xi_2, \quad x_3 \sim c_3^{\mathfrak{e}} \cdot y + \xi_3, \tag{4.22}$$

$$\mu_0 = b_\mu + c_{\mu,1} \cdot x_1 + c_{\mu,2} \cdot x_2, \quad \tau = b_\tau + c_{\tau,2} x_2, \tag{4.23}$$

$$t \sim \mathcal{B}(\sigma(b_t + c_t \cdot x_2)), \quad y \sim \mathcal{B}(\sigma(\kappa_y \cdot (b_y + \mu_0 + t \cdot \tau + \xi_y))), \tag{4.24}$$

where the parameters are set to $\boldsymbol{c}_2 = 10^{-2} \cdot [-5, -3, -4, -7]$, $\boldsymbol{c}_3 = [1, -1, -2, 1.5]$, $b_\mu = 1, c_{\mu,1} = 1, c_{\mu,2} = 1$, $b_\tau = 0, c_{\tau,2} = 8$, $b_y = 20, \kappa_y = 10^{-2}$, $\xi_y \sim \mathcal{N}(0, 0.5^2), \xi_2 = 0, \xi_3 \sim \mathcal{N}(0, 1)$.

Evaluation metrics. To quantitatively assess the performance of competing causal

[5]By causal we mean $\boldsymbol{x} \to y$.

inference procedures, we consider the following performance metrics from the literature: (i) ITE accuracy as quantified by ϵ_{PEHE}; (ii) coefficient error respectively for causal & spurious features $(\epsilon_{\text{CA}}, \epsilon_{\text{SP}})$.

Baseline solutions. To compare, the following strong or popular causal estimation baselines are considered: linear regression (OLS, with the T-learner setup); linear regression with L_1 (LASSO) or L_2 (Ridge) regularizations; *Bayesian Additive Regression Trees* (BART) (Chipman et al., 2010), *Causal Random Forests* (Causal RF) (Wager and Athey, 2018b), and *Counterfactual Regression* (CFR) (Shalit et al., 2017).

To demonstrate the effectiveness of RICE, we use the generated toy data to benchmark its performance. Figure 4.6 visualizes RICE solutions with different environment regularization strength λ, with the predicted treatment effect $\hat{\tau}(\boldsymbol{x})$ plotted against ground $\tau(\boldsymbol{x})$. When $\lambda = 0$, RICE is equivalent to the pooled ERM solution, and this yields a paradoxical solution due to the unresolved confounding from environment heterogeneity: the predicted trend is in the opposite direction to the true causal effect. As we can see, with a proper regularization parameter λ, RICE successfully mitigates the negative impacts from unstable correlations, producing a fairly reasonable estimate at $\lambda = 0.58$.

We further compare RICE with four other benchmark models: Counterfactual Regression (CFR), BART, Causal Random Forest (RF), and BV-NICE. We consider two setup: (i) Toy A: data with environment heterogeneity, generated using Eq. 4.22 - Eq. 4.24 described above; and (ii) Toy B: data without environment heterogeneity, generated by setting $\boldsymbol{c}_2 = -5 \times 10^{-2}, \boldsymbol{c}_3 = 0$. We summarize the results in Table 4.3. We see that in the presence of data heterogeneity, RICE outperforms all other baseline models in both in-sample and

out-sample predictions. In the homogeneous setup where data can be safely pooled, RICE also delivered near-optimal performance. This shows it is generally safe to replace standard inference procedures with RICE when data heterogeneity is suspected: even if the data is indeed homogeneous, the extra price paid is merely more computations rather than the model accuracy. This is usually a reasonable deal when computational resource is abundant.

Table 4.2: Comparison of treatment effect estimation

	Toy-A		Toy-B	
$\sqrt{\epsilon_{PEHE}}$	In-sample	Out-sample	In-sample	Out-sample
CFR	0.0406	0.0484	0.0204	0.0205
BART	0.0489	0.0594	**0.0142**	**0.0144**
Causal RF	0.0364	0.0458	0.0216	0.0217
BV-NICE	0.0744	0.1097	0.0151	0.0152
RICE	**0.0231**	**0.0397**	0.0151	0.0151

Table 4.3: Comparison of treatment effect estimation

	Heterogeneous		Homogeneous	
$\sqrt{\epsilon_{PEHE}}$	In-sample	Out-sample	In-sample	Out-sample
CFR	0.0406	0.0484	0.0204	0.0205
BART	0.0489	0.0594	**0.0142**	**0.0144**
Causal RF	0.0364	0.0458	0.0216	0.0217
BV-NICE	0.0744	0.1097	0.0151	0.0152
RICE	**0.0231**	**0.0397**	0.0151	0.0151

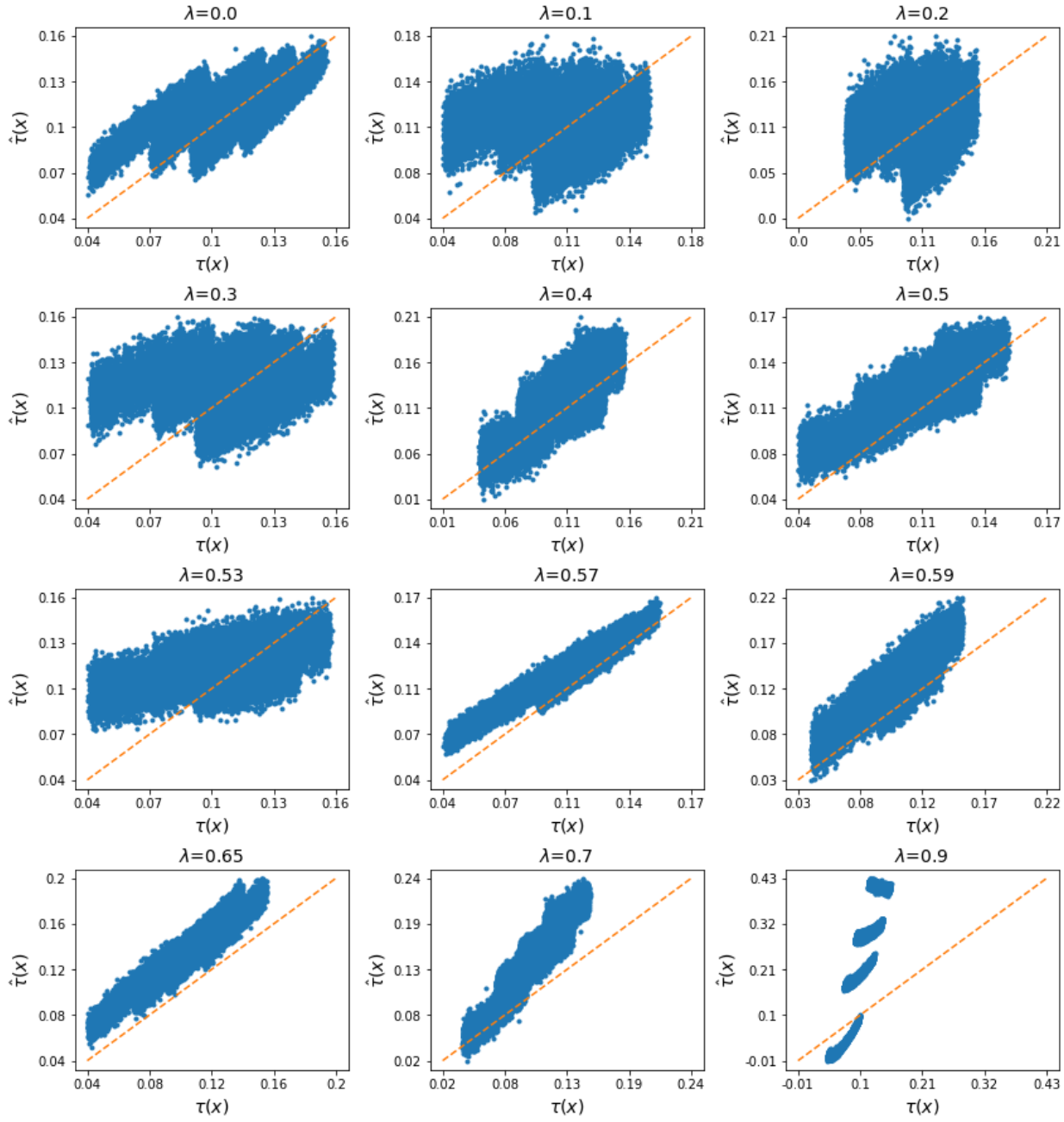

Figure 4.6: RICE solutions along the regularization path.

4.4 Invariant Conditional Logistic Regression

4.4.1 Non-randomized experimental designs and population shift

In this section, we want to discuss the challenges from non-randomized experimental designs, and show how the integration of invariant learning strategies can help improve the generalization of causal conclusions for such setups. While our discussion applies more generally to all experimental designs that employs a biased sampling strategy, we will focus on the representative case-crossover studies as a working example.

Non-randomized experimental designs and challenges for generalization. A major obstacle against large-scale randomized experiments is cost-effectiveness, especially so for scenarios where the target events is of low incidence rates. Non-randomized study (NRS) designs offer a more sample-efficient alternative, typically by preferentially targeting populations that have experienced the target event (*i.e.*, the case population). The practice of NRS is prevalent, with prominent example such as the celebrated *case-control* studies, where the control set is collected *ad hoc* to match the characteristics of the case samples. However, their conclusions are highly controversial for being subjected to the population sampling bias. In simple words, the results are derived for the high-risk populations collected by the study and do not necessarily carry over to the low-risk population, which typically dominates the general population. More generally, the challenge is how to build such statistical estimators that generalize well to the populations different from the data collected by the specific study design.

In this study, we will concentrate on *case-crossover* (CC) study design, a matched case-

control design in which the matching controls come from the subjects themselves. First introduced to to assess the transient effect of an exposure during a brief interval of time (Maclure, 1991), CC seeks for answer to questions like "was this event triggered by the exposure immediate before the event?". Only samples from the same subject matched to conditions at the onset of the event enter the study. The original intention of the CC design was to eliminate the (hidden) confounding from subject heterogeneity and to reduce control-selection-bias (Maclure et al., 2000). This study design is especially helpful in the rare-event situations or when the sampling is costly, where the investigator usually do not have extravagance of accessing ample samples, and it has been widely used in epidemiology.

Operational-wise, CC reaches causal conclusion via assessing the imbalance of the treatment variable between the event and non-event episodes, using statistical tools such as the Mantel-Haenszel test or conditional Logistic regression (Guo et al., 2017a, Klauer et al., 2010, Marshall and Jackson, 1993). While delivering superior sample efficiency and less prone to the confounding effects (Breslow et al., 1980, Breslow, 1980), it has been widely recognized that CC sample is more risk-prone and not representative of the entire population (see Figure B.1, B.2). This is evidenced by a direct comparison of the results respectively derived from the case-crossover and case-cohort designs, where notable statistical discrepancies persist (see Figure B.3 and Table B.4). This cautions extrapolating conclusions from a CC study to a general population without further statistical adjustments. In the sections below we present a reformulated analysis strategy that hopefully overcomes this difficulty.

4.4.2 Invariant conditional Logistic regression

In this section, we describe how to use IRM to enable invariant learning for matching-based designs such as case crossover studies. We focus on reformulating the *conditional Logistic regression* (CLR), a classical stratified analysis procedure widely applied in such contexts.

We use a bi-level indexing system (i, j) to index the data used for CLR, where i denotes the stratum and j indexes the samples within each stratum. In the case of NDS study, each participating driver defines one such stratum. For each stratum i, we have n_i data points $\boldsymbol{s}_i = \{(\boldsymbol{x}_{ij}, t_{ij}, y_{ij}, \boldsymbol{u}_i)\}_{j=1}^{n_i}$, where $\boldsymbol{x}_{ij}$ are the associated/observed covariates (exogenous factors), t_{ij} are the binary treatment whose effect we want to quantify, y_{ij} are the binary label denoting the event of interest, and $\boldsymbol{u}_i$ denotes the stratum specific characteristics, possibly unobserved. Note that $\boldsymbol{u}_i$ is typically regarded as the hidden confounding whose effect an investigator wants to remove. While its exact value is unknown, the matching operation employed in the study design tries to make sure this $\boldsymbol{u}$ is the same for all instances within the same stratum.

Without loss of generality, we always assume there are at least two samples in each stratum, and the event of interest only happens once, denoted by $y_{i1} = 1$ and $y_{ij} = 0$ for $j > 1$. In this study, we assume the following additive form of effect decomposition in logits

$$g(\boldsymbol{x}, t, \boldsymbol{u}) = t \cdot \tau(\boldsymbol{x}, \boldsymbol{u}) + \mu_0(\boldsymbol{x}, \boldsymbol{u}), \tag{4.25}$$

where $\tau(\boldsymbol{x}, \boldsymbol{u})$ and $\mu_0(\boldsymbol{x}, \boldsymbol{u})$ are respectively the individualized treatment effect and baseline

effect. For simplicity, we only concern models that are linear, *i.e.*,

$$\tau(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{x}^T \boldsymbol{\phi} + \boldsymbol{u}^T \boldsymbol{\phi}_u, \quad \mu_0(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{x}^T \boldsymbol{\beta} + \boldsymbol{u}^T \boldsymbol{\beta}_u, \tag{4.26}$$

with model parameters $(\boldsymbol{\phi}, \boldsymbol{\beta})$ explicitly characterizing the effects of interest. The conditional Logistic regression model seeks to maximize the probability of getting $y_{i1} = 1$ amongst all n_i observations under the Logistic model

$$\mathbb{P}(y_{i1} = 1 | \{t_{i,j}, \boldsymbol{x}_{i,j}\}_{j=1}^m, \sum_j y_{ij} = 1, \boldsymbol{u}_i) = \frac{\exp(g(\boldsymbol{x}_{i1}, t_{ij}, \boldsymbol{u}_i))}{\sum_j \exp(g(\boldsymbol{x}_{ij}, t_{ij}, \boldsymbol{u}_i))}. \tag{4.27}$$

Plugging Eq. (4.25-4.26) into Eq. (4.27) gives

$$\mathbb{P}(y_{i1} = 1 | \boldsymbol{s}_i) = \frac{\exp(\boldsymbol{x}_{i1}^T (t_{i1} \cdot \boldsymbol{\phi} + \boldsymbol{\beta}))}{\sum_j \exp(\boldsymbol{x}_{ij}^T (t_{ij} \cdot \boldsymbol{\phi} + \boldsymbol{\beta}))}. \tag{4.28}$$

This allows us to define the stratum loss and the corresponding risk for CLR

$$\ell_{\text{CLR}}(g; \boldsymbol{s}_i) = \log \mathbb{P}(y_{i1} = 1 | \boldsymbol{s}_i), \quad R(g) \triangleq \mathbb{E}_i[\ell_{\text{CLR}}(g; \boldsymbol{s}_i)]. \tag{4.29}$$

To estimate stable effects across environments, we define the environment risk as

$$R^{\mathfrak{e}}_{\text{CLR}}(g) = \mathbb{E}_{i \in e}[\ell_{\text{CLR}}(g; \boldsymbol{s}_i)]. \tag{4.30}$$

and then solve for

$$\mathcal{L}_{\text{CLR}}(g; \lambda) = \sum_e \{(1 - \lambda) R^{\mathfrak{e}}_{\text{CLR}}(g) + \lambda \| \nabla_{w|w=1} R^{\mathfrak{e}}_{\text{CLR}}(w \cdot g) \|^2\}, \tag{4.31}$$

where $\lambda \in [0, 1)$ characterizes the regularization path.

Note that our ICLR model differs from standard CLR models not only in the invariance regularization, but also in the modeling of treatment effect heterogeneity. Standard CLR models only models the constant population average treatment effect $\tau = c_\tau$, while in ICLR we model the conditional treatment effect $\tau(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\phi}$. We also include the matching factors $\boldsymbol{x}_m$ in our ICLR model to better model effect heterogeneity. To avoid redundant parameterization, we freeze the baseline weights on matching factors ($\boldsymbol{\beta}_m$) to zero.

4.4.3 Experiments

Toy model. We simulate data from the following toy model to benchmark our model. Similar to the toy model from last section, we use three covariates $\boldsymbol{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$ respectively denoting upstream non-direct cause, direct cause, and downstream correlates (through the outcome y) of the treatment $\tau(\boldsymbol{x})$.

$$x_1 \xrightarrow{\mathfrak{e}} x_2, \quad x_2 \xrightarrow{\mathfrak{e}} t, \quad (x_1, x_2) \to \mu_0, \quad (\mu_0, \tau) \to y, \quad y \xrightarrow{\mathfrak{e}} x_3, \tag{4.32}$$

where $\xrightarrow{\mathfrak{e}}$ and $\to$ respectively denote environment-dependent and independent links.

Different from the RICE setup, we simulate data in Strata. Specifically, each stratum consists of one event ($y = 1$) and a varying number of non-event ($y = 0$). For this, we first sample the number of instances n_i in each stratum, with $n_i \in \{2, 3, 4, 5\}$ with respective probability $[0.1, 0.2, 0.5, 0.2]$. We sample and compute $(\boldsymbol{x}_{ij}, t_{ij}, \tau_{ij}, \mu_{ij})$ from the model above

Eq. (4.22-4.24), and denote $\boldsymbol{l}_i = [l_{ij} = t \cdot \tau_{ij} + \mu_{ij}]$ the result logits for each instance. We then sample $\boldsymbol{y}_i = [y_{ij}]$ from the multinomial distribution $\mathcal{M}(\boldsymbol{p}_i)$ to ensure only one of y_{ij} equals 1. Note the parameters are set to: $\boldsymbol{c}_2 = 10^{-2} \cdot [-5, -3, -4, -7]$, $\boldsymbol{c}_3 = [3, -2, -1, 1]$, $b_\mu = 1, c_{\mu,1} = 1, c_{\mu,2} = 1$, $b_\tau = -1.8, c_{\tau,2} = 8$, $b_y = 1, \kappa_y = 10^{-2}$, $\xi_y = 0, \xi_2 = 0, \xi_3 \sim \mathcal{N}(0, 1)$.

In Figure 4.7, we visualize ICLR estimates for the causal effect $\tau(\boldsymbol{x})$ along the regularization path $\lambda \in [0, 1]$, where the predicted causal effect $\hat{\tau}$ is plotted against the ground-truth value. We see that in this example, although the ERM solution ($\lambda = 0$) captures the general trend, its causal estimates are heavily contaminated by the unstable correlations which results in strong biases. As we gradually enforce more invariance, the solutions become less affected by the unstable correlations, reaching optimal predictions at $\lambda = 0.95$. Further stressing the invariance leads to overly conservative estimates that fail to encode the effect variance. In Figure 4.8, we plot the individual leave-one-out cross-validation curves (left) and their average (right). We see the LOO-CV curve aligns well with the accuracy of treatment effect estimation (Figure 4.7), which justifies the use of LOO-CV as model selection criteria.

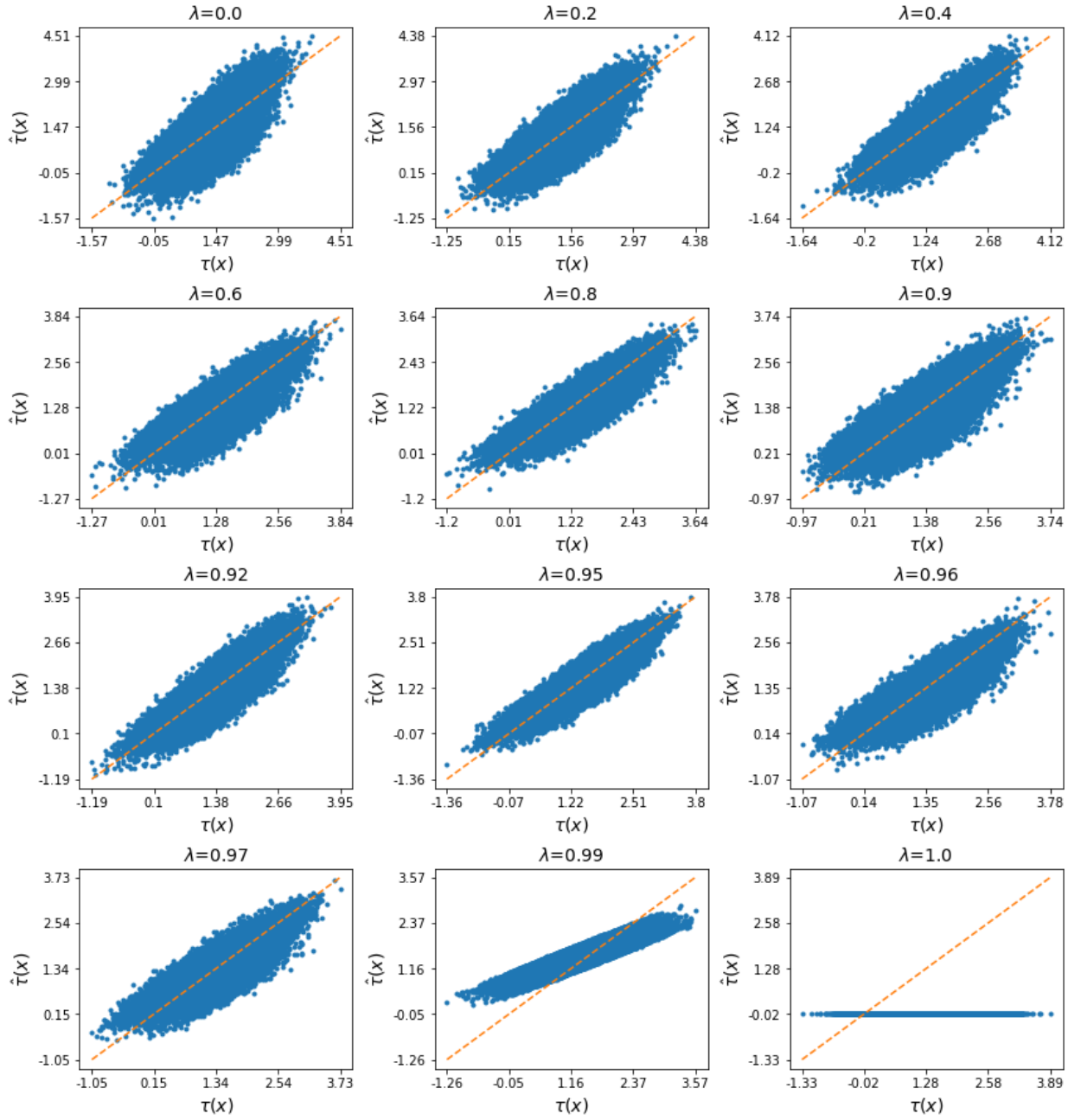

Figure 4.7: ICLR estimates with different environment penalties.

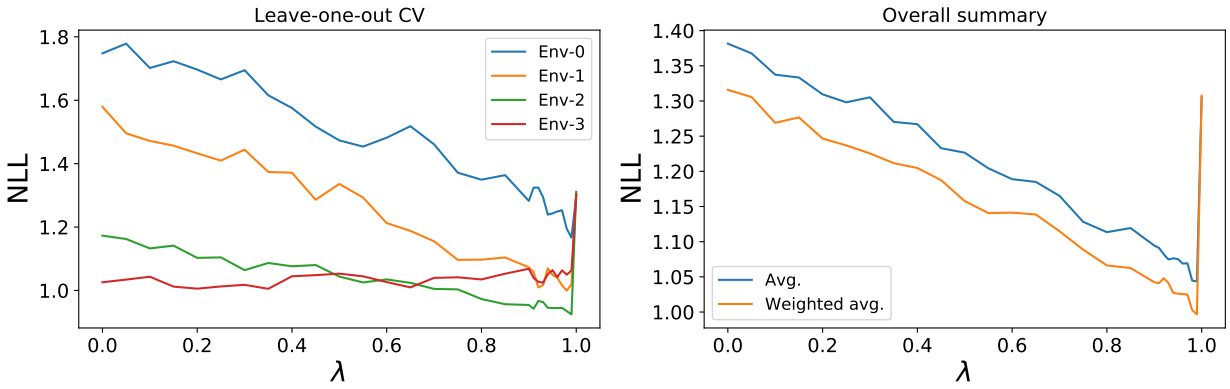

Figure 4.8: ICLR cross validation of NLL

4.5 Applications to the NDS data

4.5.1 NDS study designs and issues

In this section, we apply the proposed RICE models to analyze the causal effect of cellphone distraction on crash risk using SHRP2 NDS data. Distraction has been considered one of the most critical factors in driving safety, with more than 9% of fatal accidents were reported as distraction-related (NHTSA, 2019a). In 2017, more than 3,000 people were killed in motor vehicle crashes in United States alone due to distracted driving, and electronic devices such as cellphones construct a major source of such distractions(NHTSA, 2019b). Consequently, much attention has been focused on the safety impact of cellphone distractions during driving. We provide a brief summary of the SHRP2 NDS data below and point to Chapter 2.2 for a more detailed description.

SHRP2 NDS is an innovative observational study to evaluate the safety impact of driver behaviors by continuous recording of driving information through a comprehensive data acquisition system installed on participants' vehicle (Hankey et al., 2016). Unlike well-controlled randomized trials, the engagement of cellphone distractions is influenced by the constant changing driving contexts and the confounding pattern is expected to change from state to state. While such data heterogeneity complicates analysis in traditional ways, it also permits our RICE framework to evaluate the stable treatment effect, *i.e.*, the safety impact of cellphone distractions that is consistent across different environments and driver populations.

While the NDS data is abundant, the crash cases are still rare. As such, case-referent designs are often used to sample proper baselines for the evaluation of relative crash risks (Guo, 2019b, Miettinen, 1976). The case-cohort design and CC design are two major post hoc approaches employed. The former relies on a random sample to represent the population of interest, and the latter utilize a matching mechanism to eliminate potential confounding (Guo, 2019b, McEvoy et al., 2005, Redelmeier and Tibshirani, 1997b). These two designs generate different data heterogeneity patterns.

One major hurdle of generalizing CC results to more general driver population is the heterogeneity in the data. There're two sources of heterogeneity in the data, treatment effect heterogeneity and environmental heterogeneity. Treatment effect heterogeneity refers to the fact that the safety impact of distraction depends on the driver and driving contexts. A population level estimation fails to generalize to a shifted distribution, *e.g.* CC sample favors high risk drivers and dangerous driving contexts (Figure B.1, B.2) and therefore yields inconsistent estimation compared with case-cohort design (Figure B.3, Table B.4). In

addition, the potential presence of sub-groups with varying confounding pattern exacerbates the estimation bias. For example, the features affecting young drivers' cellphone distraction status may be different from adult drivers. The heterogeneous sub-groups in this example is aforementioned environment heterogeneity. We seek to learn the invariant distraction impact that generalizes across environments. To be noticed, the population in this NDS study refers to driving segments, *i.e.*, a combination of driver and driving contexts.

4.5.2 Experiment setup

Data description. To explore the effectiveness of proposed RICE on the NDS data, we consider the following setup. We use the `State` as the environment variable, which partitions the data into six different environments (`FL, NC, PA, IN, NY, WA`). We identify demographic factors `Age`, `Gender`, `Income`, and driving contexts `Weather`, `Lighting`, `Surface Condition`, `Relation to Junction`, `Highway`, `Intersection`, `Construction Zone`, `Traffic Density` as potential confounders (denoted as $\boldsymbol{x}$). This study is characterized by the challenge of rare-event modeling, due to the exceptionally low incidence rates of traffic accidents. Only $1k$ crashes were flagged and annotated by trained analysts to represent the potential risk factors, along with about $20k$ normal driving baselines for control.

Our goal is to evaluate the heterogeneous treatment effect $\tau(\boldsymbol{x})$ of cellphone distraction t, with $t = 1$ denoting being distracted by cellphone. Only crashes of *level* 1 to *level* 3 are included in the study. For the RICE model, 892 crashes and 23,343 baselines are included. For the CC design, `Weather`, `Traffic Density`, and `Highway` are used as matching factors, with each crash case matched to $1 \sim 4$ baselines from the same driver, with a total number of

~ 530 matched strata. Note that in this study, we compare cellphone distraction with non-cellphone distracted driving, which includes model driving and other source of distractions. We summarize the data statistics in Table 4.4. The observations from PA were dropped due to limited cellphone distraction data points.

Table 4.4: Stratified statistics of NDS data

	Case cohort		Case-crossover	
	Baselines	Crashes	Baselines	Crashes
FL	5,527	257	513	152
IN	1,769	72	121	40
NC	3,854	138	263	75
NY	5,173	197	413	132
PA	1,417	45	71	23
WA	5,603	183	368	108
Total	23,343	892	1,749	530

Models. For RICE, we have used the Robinson residual-based model Eq. (4.13-4.14) for analysis. To be consistent with prior analysis, we focus on linear model for the mean outcome $m(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\phi}_m$, propensity score $e(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\phi}_e$ and treatment effect $\tau(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\phi}_\tau$. We adopt the two-stage learning approach for RICE and first estimate the nuisance models $m(\boldsymbol{x})$ and $e(\boldsymbol{x})$ using the IRM variant of Logistic regression. We do not find $m(\boldsymbol{x})$ and $e(\boldsymbol{x})$ are heavily influenced by the data heterogeneity so we fix $m(\boldsymbol{x})$ and $e(\boldsymbol{x})$ to their ERM estimates in subsequent analyses for simplicity.

Model evaluation. To objectively evaluate the out-of-distribution (OOD) performance of RICE and ICLR models, we use environment-based cross-validation described in Section 4.3.3. In particular, we adopt the leave-one-out (LOO) approach: each time we keep one of

the environment data for validation purpose and use the rest environments for training. We train models with 30 different λ values, with a step-size of 0.05 between $[0, 0.9)$ and a more refined grid of step-size 0.01 between $[0.9, 1]$, where more interesting dynamics happen for this data set. The risk for each λ is averaged over environments to pick the best performing regularization strength.

For the RICE model, we find two practical challenges due to the extreme low event rate, which render the standard risk evaluation inadequate: (i) the predicted risk $\hat{y}(\boldsymbol{x}, t) = m(\boldsymbol{x}) + (t - e(\boldsymbol{x}))\tau(\boldsymbol{x})$ can fall out of the range $[0, 1]$ due to finite-sample estimation errors; (ii) the L^2 loss is not sensitive to difference for binary outcomes. As a result, we observe very noisy OOD risk profiles that do not register stable patterns.

To overcome this difficulty, we consider the predicted $\hat{y}$ as a risk score for the outcome (the higher the riskier), without explicitly associating probabilistic meanings. For a better risk model, we should expect a larger discrepancy between the risk scores for the cases ($y = 1$) and the baselines ($y = 0$). This motivates us to test for the group difference between the two distributions $\hat{q}_1 = \{\hat{y}_i\}_{y_i=1}$ and $\hat{q}_0 = \{\hat{y}_i\}_{y_i=0}$, with plausible choices such as two-sample t-test and rank correlation.

We use the test score of the group risk difference as the evaluation metric to replace the validation set risk $R^{\mathfrak{e}}$ used in Section 4.3.3, which works for general cases. Since the scale of test scores varies across environments, we use respective ranks in each environment, the environment-averaged relative rank to select the optimal penalty λ.

Feature selection. Due to the low-event rate, we observe different combinations of confounders have a big impact on the model performance. This prompts us to apply a

similar cross-validation procedure to select the optimal combination of features for cross-environment generalization. We appeal to the ICLR model for feature selection as the reported likelihood is more statistically grounded and also less noisy compared with the test statistics reported by the RICE model on the NDS SHRP2 data. Specifically, we outline a set of candidate feature combinations and run ICLR LOO-CV on each of them, with the one reporting best hold-out log-likelihood fed to all subsequent analyses with both RICE and ICLR.

Training specifications. Our models are implemented with `PyTorch`. We use `Adam` as optimizer and randomly draw mini-batches for training. For RICE, we sample $n_b = 256$ observations per mini-batch per environment; and for ICLR, we fix mini-batch size at $n_b = 128$ due to a smaller sample size of CC data. We set learning rate to 10^{-3}.

4.5.3 Results and analyses

Environment heterogeneity. First we would like to confirm the presence of environment heterogeneity in NDS data, which justifies the application of invariance-based analysis. In Figure 4.10, we plot the feature correlation matrices for all environments. We observe apparent disparities in the correlation structure of data, possibly due to the variations in natural environments, road infrastructure, population structure, and cultural differences across states. These dissimilarities carry over to the environment-specific model estimates (Figure 4.10), resulting inconsistent coefficient estimates across states. And consequently, naively trained models generalize poorly to unseen environments.

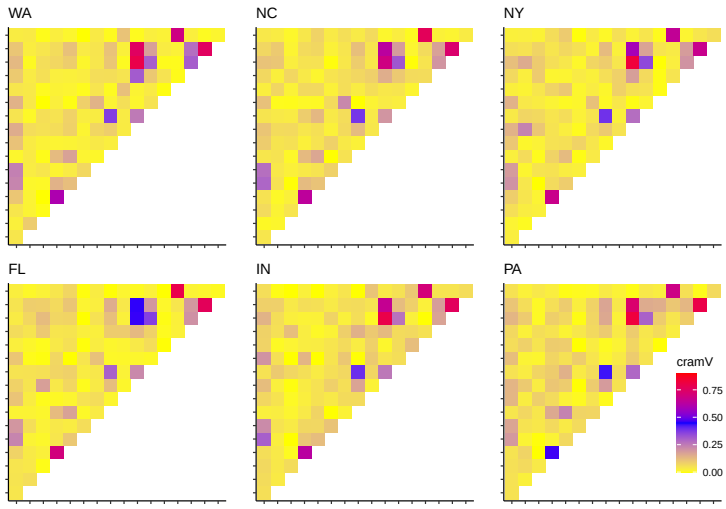

Figure 4.9: Environment heterogeneity of NDS data: feature correlations.

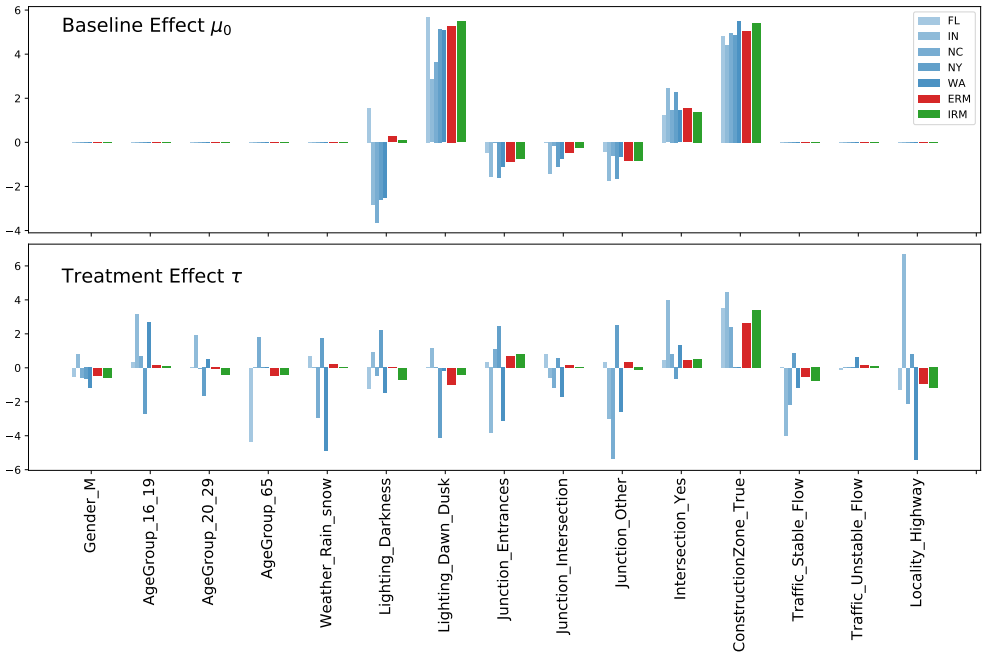

Figure 4.10: Environment heterogeneity of NDS data: model estimates.

Next we want to verify the existence of stable relations that transcend environment

heterogeneity, such that the proposed invariant causal models indeed improve generalization in new environments. To this end, in Figure 4.11 we plot the leave-one-out cross-validation curves for the ICLR model, in which we have varied the regularization parameter λ and used the negative log-likelihood (NLL) for evaluation. The average out-of-distribution NLL shows consistent improvement over the ERM baseline ($\lambda = 0$) as we gradually increase λ, with the optimal value attained at $\lambda = 0.93$. This is further confirmed by visualizing the individual CV curves for each hold-out environment, with most of the environments showing a consistent pattern of better OOD performance under moderate environment penalty. As expected, over capitalizing on the invariance undermines model performance, as the NLL quickly degenerates as we set λ beyond 0.98.

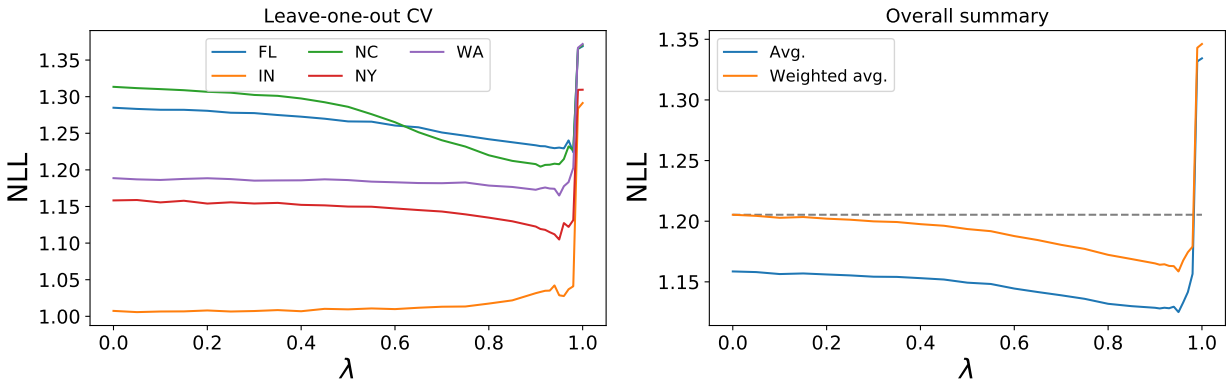

Figure 4.11: ICLR: cross-environment validation results on NDS CC

For the RICE model, while the CV curves are much noisier, we can still recognize a similar trend of improved OOD performance for a relatively large environment regularization (see Figure 4.12, where we have also provided smoothed CV curves for better visualization). Here we have used the two-sample t-test to evaluate the model predicted risk difference between cases and baselines, and similar results can also be confirmed using rank correlation

metrics (*e.g.*, Spearman's r, Kendall's τ). We also compared the LOO parameter estimates between ERM and RICE, and observed that RICE gives lower variance estimates.

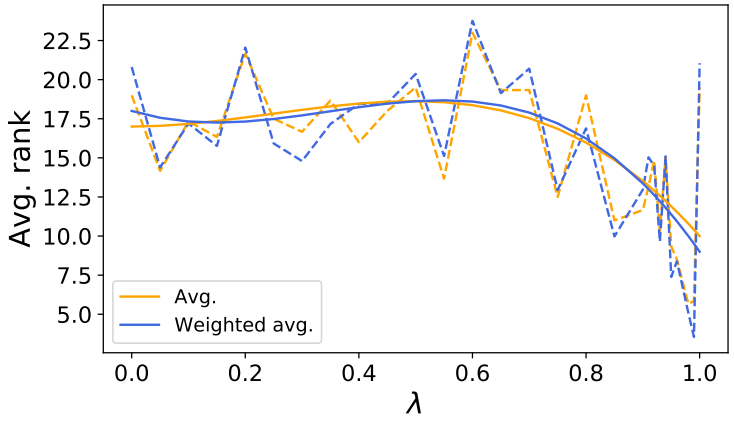

Figure 4.12: RICE: cross-environment validation results on NDS case-cohort

While in general the results lend support for our hypothesized invariant relations and our proposed remedies, there are a few minor departures to our expectation. For example, contrast to its counterparts, environment `IN` does not benefit from the imposed invariance. Such discrepancies imply room for future improvements. For instance, the current uniform environment regularization is a bit crude and by no means optimal, as different environments may have different levels of departure from its optimal, which ideally should be accounted for differently (*i.e.*, different λ for each environment). An efficient optimization procedure that jointly learns model parameters and optimal regularization strength λ would be of great significance. For those environments that the IRM penalty does not help, there are a few possible theories: (*i*) statistical variations due to finite sample size; (*ii*) remaining environments lack the sufficient diversity to resolve the invariant and unstable relations; (*iii*) violation of the invariance assumption, such that not all environments conform to the

same relation. Given the data, we were unable to test these hypotheses.

To use the learned causal model to better understand the impact of cellphone distractions on vehicle collisions, we show the model predicted cellphone distraction effects under different driving conditions in Figure 4.13, based on CC data. Specifically, we plot the distribution of treatment effects for all the samples satisfying particular context condition (*e.g.*, `Constructionzone_true` denotes all driving samples recorded close to construction zones), ordered by corresponding average treatment effect. The plot shows that construction zone, parking lot, and intersection are top three risk factors, followed by female drivers. Notably, cellphone distraction at highway reflects a protective effect, potentially due to other competing distraction sources.

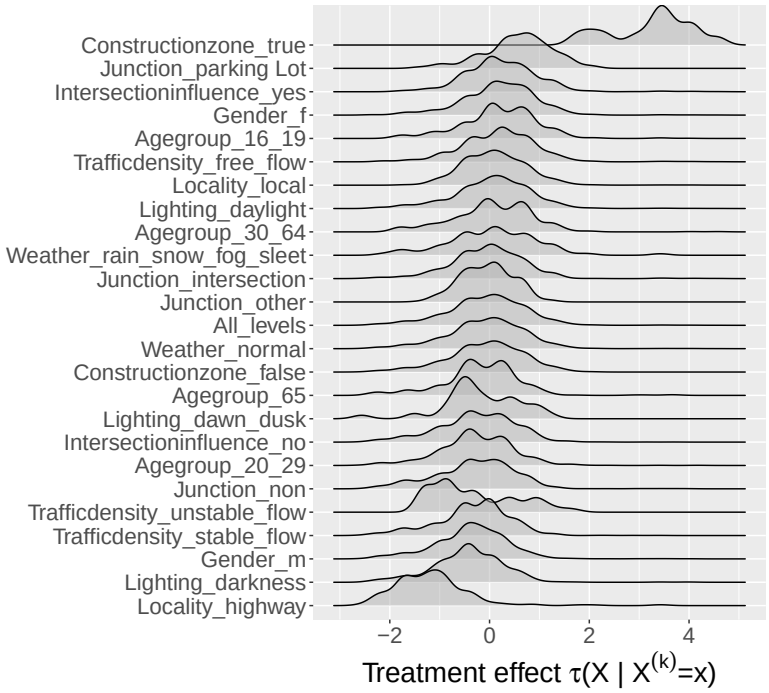

Figure 4.13: ICLR predicted cellphone distraction effects under different driving contexts using CC data.

In addition to conditional treatment effect for individual factors, the individualized treatment effect enables flexible analysis of different risk combinations and driving contexts. Figure 4.14 shows the distribution of cellphone risk for teenage drivers (age 16-19) at construction zones. The risk for teenage drivers is slightly higher than overall average while the risk is dramatic higher when teenage driver uses cellphone at construction zones. Similar pattern is observed at parking lot. The ITE evaluation allows flexible evaluation of risk profile, and enables generalization to out-of-distribution samples. For example, the invariance property of RICE allows generalization to other states outside of the six states we collected sample from as the information of extreme weather, age composition, and landscape information can be incorporated and adjusted in the model.

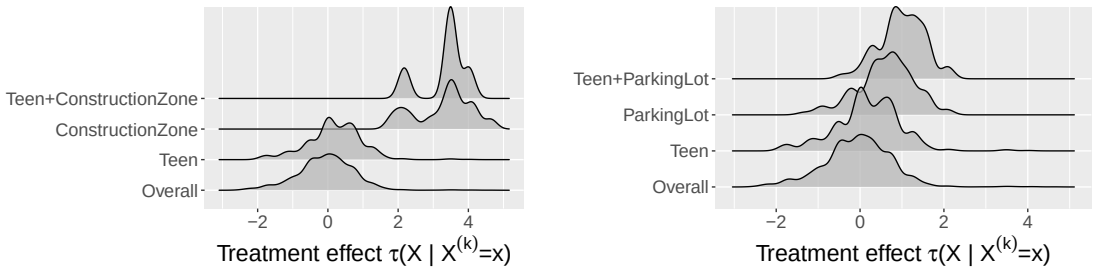

Figure 4.14: Treatment effect distribution under different risk factor combinations

4.6 Discussion and conclusion

This work explores novel strategies to establish causality based on principles of statistical invariance, and shows that promoting stability helps to overcome the challenges from spurious correlations and population shifts common to empirical investigations. Built on the IRM framework, we present new observational causal inference procedures that exploit

data heterogeneity to identify stable causations to sponsor counterfactual reasoning. Our research identifies new settings where the effects from unobserved confounding can be mitigated. Empirical evaluations confirm the effectiveness of our proposal and point to new research opportunities. Finally, we conclude this chapter with some remarks and discussions on related work and future directions below.

Causal invariance and representation learning. Stipulation of invariant relations across different environments constructs a major assumption for robust machine learning (Schölkopf et al., 2012), where the learned models are expected to safely extrapolate beyond training distributions (Bühlmann, 2018). We broadly classify these efforts into two categories: predictive causal models and generative causal models. Models in the first class concentrate on the identification of invariant representations and causal relations via penalizing environmental heterogeneities, which includes prominent examples such as *invariant causal prediction* (ICP) (Heinze-Deml et al., 2018, Peters et al., 2016), IRM (Arjovsky et al., 2019) and *risk extrapolation* (REx) (Krueger et al., 2020). For the second category, the observed cross-environment distribution disparities are tethered by an invariant generation procedure (Hyvarinen et al., 2019, Khemakhem et al., 2020), which leads to a causally invariant source representation that can be exploited to mitigate issues such as sample scarcity (Teshima et al., 2020) and data imbalance (Chen et al., 2020). Our work pertains to the first category, with the main objective focused on the robust identification of stable counterfactuals that generalize to novel distributions.

Two facets of causality. In stark contrast to predictive models, the hallmark feature of causal conclusions is that they are expected to generalize in the presence of external interventions. As such, we promote the view that covariate balancing and statistical in-

variance are both important ingredients to the establishment of causality. In essence, both perspectives seek to remove the confounding effects from unwanted correlations, in order to reach a de-biased causal interpretation from the observational data. This objective is approached with different means under their respective heuristics: covariate balancing schemes adjust the data distributions to make *"apple-to-apple"* comparisons such that the interfering correlations are effectively suppressed, while the invariance-based schemes base their inferences on representations and relations that are stable across different scenarios (*i.e.*, under different interventions).

Balancing and invariance. Although not explicitly explored in this work, exploiting the synergies between the representation balancing and causal invariance constructs an interesting topic and is expected to bear fruit. The IRM framework adopted by our RICE and ICLR does not directly regularize for the postulated invariant representation, and instead penalizes for the (local) environment-optimality condition as a surrogate criteria for invariance. We surmise the solutions can be further strengthened via seeking a balanced invariant representation through organically integrating the representation matching and stability criteria.

RCT and heterogeneity. Our research carries great practical significance, in that the key ideas apply more broadly to other empirical scenarios of interest. This even holds true for the *randomized controlled trial* (RCT), which is considered the gold standard for probing and validating causality. As a critical remark, RCT is also troubled by the population heterogeneity, which often necessitates stratified sampling and the corresponding analytics. Our work exposes the inadequacy of stratified strategies, as they are sample inefficient, still susceptible to spurious associations and do not necessarily generalize to novel populations.

Our invariance-based causal estimation procedures offer a powerful remedy to address these limitations.

This work opens a new door for observational causal inference, and it also generates a number of new questions for future investigations. Of great interest is how to balance the utility - invariance trade-off (Rothenhäusler et al., 2018). Not all spurious correlations are bad, some are useful. The notation of invariance can be relaxed for related environments to better capture the predictive power. Also, automated identification of heterogenous populations can greatly facilitate the operation and deployment of such invariance-based solutions (Sagawa et al., 2019). More generally, the extra ability to handle *Missing-Not-At-Ramdom* (MNAR) data will greatly extend the scope of applicability for the proposed framework (Steck, 2010).

Chapter 5

Conclusion and Future Work

In this chapter, we conclude this dissertation by summarizing our contributions and discussing directions for future work.

5.1 Summary

Establishing causality in observational studies has been a challenging topic for decades. Over the years, considerable efforts has been made to extend causal conclusion from randomized experiments to non-randomized study through modeling treatment assignment mechanism (Rubin, 1977), balancing covariate distributions (Dahabreh et al., 2019, Rosenbaum and Rubin, 1983a), and predicting missing counterfactuals (Künzel et al., 2019, Wager and Athey, 2018a). In general, causal estimation relies heavily on strong assumptions like unconfoundedness and homogeneous units, which are often improbable in practice, hindering more robust conclusions and broader applications. This dissertation investigates novel statistical

approaches for causal effect estimation in observational studies from different perspectives. Apart from introducing classical propensity score based balancing schemes to a new application field NDS safety study, we present novel frameworks for individualized treatment effect estimation to handle data uncertainty and heterogeneity by deriving new learning objectives and embracing trending representation learning techniques.

The safety assessment of driving behavior in observational setup is challenging due to driver's volitional engagement and constant changing driving contexts (Lu et al., 2020a). Different matching schemes were adopted to eliminate confounding due to driver demographics and driving contexts (Klauer et al., 2010, Redelmeier and Tibshirani, 1997a). However, the matching approaches are usually sample inefficient and focuses on a reduced population that might not represent the target population. Chapter 2 introduces classical propensity score based adjustment schemes in restoring covariate balance in novel NDS study, with respect to three different estimands and corresponding target populations. It highlights the importance to adopt causal perspective in analyzing driver behavior associaing ed risk in order to rule out complex confounding impact from multiple risk factors.

In the wake of big data era, traditional causal approaches become inadequate: they tend to suffer from the lack of balance efficiency, individualization, and robustness against high-dimensional data & complex interactions (Lu et al., 2020b). Chapter 3 presents a novel generative Bayesian causal framework BV-NICE for individualized treatment effect estimation. BV-NICE extracts causal information from the noisy and high-dimensional proxies by incorporating stochastic feature encoder that fully acknowledges representation uncertainties. Compared with traditional propensity score method, BV-NICE achieves better covariate balance by directly minimizing representation discrepancy between the balancing

arms. In addition, the integration of Robinson decomposition enables reliable counterfactual validation and yields highly competitive causal estimation in terms of accuracy and robustness compared to state of the art solutions.

Taking a novel perspective paralleling with representation balancing, Chapter 4 emphasizes the importance of ascertaining causal stability and vouches for its key role in out-of-distribution generalization. Specifically, a RICE framework is proposed to extract invariant causal effects across heterogeneous populations that generalize better to novel distributions. By leveraging the observed data disparities and penalizing the un-generalizable effects, RICE establishes invariant causal relationships and rules out spurious confounding patterns aliased with the data disparities. Importantly, RICE delivers the promise of causal identification in broader settings with unobserved confounding and sampling bias.

In summary, this dissertation presents a flexible causal inference framework that acknowledges data uncertainty and heterogeneity. It enjoys three merits: improved balance for complex covariate interactions, enhanced robustness to unobservable latent confounders, and better generalizability to novel populations.

5.2 Future work

Researches stemming from this dissertation can be extended in a number of directions. Below is a brief discussion on some of the most promising and significant headings.

Beyond binary treatments. While in this dissertation we have concentrated on the most basic case of binary treatments to simplify our discussions, we recognize in real-world

settings things are more complicated. A practitioner is often faced with situations of (*i*) multiple, possibly non-exclusive treatments to choose from (Bica et al., 2020, Bottou et al., 2013), (*ii*) treatment with multiple levels, (*iii*) or even continuous treatment. The extra complexity in treatment effect estimation in those scenarios challenges existing causal estimation procedures, including the proposed framework in this dissertation, as the positivity condition is almost surely violated in practice.

Representation learning and causality As we have mentioned in earlier chapters, causality critically depends on the assumptions made, and there are alternative assumptions we can exploit to further explore the link between representation learning and causality. A straightforward attempt would be imposing the representation balancing in addition to the causal invariance, which are respectively the main proposals from the BV-NICE model and RICE model presented in this dissertation. While the IRM framework is based on the notation of invariant representations, in actual implementation, no explicit regularization has been enforced. While the treatment arm balancing is expected to improve counterfactual reasoning, we may also consider more concrete domain shifting mechanisms, such as label shift (Zhang et al., 2013), which has been confirmed to hold for many empirical datasets. This can also be achieved by the distribution matching algorithm from BV-NICE. Alternatively, for specific datasets, making parametric (Xiu et al., 2021) or geometric (Nickel and Kiela, 2017) assumptions on the representation can improve learning efficiency. Combining weighting with representation can also promote better balance (Assaad et al., 2020, Johansson et al., 2018).

Unobserved confounding. Another key question we are keenly interested to explore further is how to approach a good causal estimation under unobserved confounding (Kallus

et al., 2018b). While this dissertation makes some advances in mitigating the negative impacts from latent confounders, more generic solutions are still desired. One promising outlet is to draw power from the sensitivity analysis literature, absorbing the intangibles into some sensitivity parameters to accompany the causal conclusions (Franks et al., 2019, VanderWeele and Ding, 2017).

Experimental designs and uncertainty. Finally, we want to exploit the synergies and create healthy feedback loops between experimental and observational causal inference. Experimental and observational studies are complementary: the former is costly but accurate, with the latter much cheaper but potentially biased. Using small RCT to calibrate and validate observational causal conclusions is an emerging research topic that attracts considerable practical interest (Rosenfeld et al., 2017, Zeng et al., 2020). Also, the line between observational and experimental studies is often blurred in realistic settings: subjects dropout as selection bias and observational data might subject to prior systematic treatments. The key ideas developed in this dissertation apply more broadly to the experimental settings as well, so the algorithms from this dissertation can be repurposed to facilitate integration with minimal modifications.

Appendices

Appendix A

BV-NICE

A.1 Discussion on Causal Inference with Unobserved Confounders

Standard literature on causal inference with observational data often posits "strong ignorability", *i.e.*, all confounders are observed. This assumption is, however, untestable and oftentimes unrealistic for practical considerations. In general, estimating causal effects with unobserved confounding is infeasible (Pearl, 2009). Without additional assumptions, the observed data distribution can be compatible with many potentially contradictory causal explanations, that are indistinguishable based on the data in the eye of an investigator (D'Amour, 2019).

One promising outlet, as adopted by many studies, is to approach a solution using *proxy*

variables (Maddala and Lahiri, 1992, Montgomery et al., 2000). For example, one can not directly measure the socio-economic status of a subject, and yet this might be indirectly assessed through alternative variables such as job types and shopping behavior, which are down stream of the unobserved confounding [1]. Caution that a common mispractice is to treat the proxies as if they are ordinary confounders, as bias can be expected (Fuller, 2009, Griliches and Hausman, 1986, Pearl, 2012, Rothman et al., 2008). It has been established that causal identification is feasible under certain conditions with proxy variables are met (Cai and Kuroki, 2008, Kuroki and Pearl, 2014, Miao et al., 2018, Shi et al., 2018).

While the aforementioned work on proxy variables typically relies on strong assumptions to ensure identifiability under unobserved confounding, a recent trend in literature is to relax the technicality involved to enable universal algorithms for identification. These attempts often induce a latent variable setup (see Figure 3.2), prominent examples include Louizos et al. (2017), Ranganath and Perotte (2018), Tran and Blei (2017), Wang et al. (2018) [2]. Of particular popularity is the CE-VAE model (Louizos et al., 2017), which is based on variational setup. Its generality and capability for modeling complex interactions with minimal modeling effort is especially significant to and well-received by emerging applications such as counterfactual reinforcement learning (Buesing et al., 2018, Lu et al., 2018). As discussed in the main text, our BV-NICE extends and improves upon CE-VAE via inducing more causal and representation learning considerations.

[1] A related approach is to leverage instrumental variables to identify causal effects (Singh et al., 2019).

[2] Note that some of the practice, such as Wang and Blei (2019), has been extensively debated in literature (D'Amour, 2019), for the existence of a non-trivial ignorance region, where conflicting causal estimates can be indifferent to the investigator based on data. This renders the conclusions uncheckable and heavily dependent on modeling choices. For such cases, sensitivity analysis is recommended to reason about the causal implications (Franks et al., 2019).

Since variational inference stands as one of the most active research topics in modern machine learning (Zhang et al., 2018), there exists ample opportunities to further improve VI-based causal inference procedures' performance and applicability. For example, more general variational objectives (Chen et al., 2018, Li and Turner, 2016), non-continuous unobserved confounders (Cong et al., 2019, Jang et al., 2017, Yin and Zhou, 2019), and flexible priors that are data-driven (Tomczak and Welling, 2018) or encode geometric structures (Nagano et al., 2019). Our observations also echos with the recent discussion that whether tightening the bounds (Burda et al., 2016, Rezende and Mohamed, 2015) in general improves performance (Rainforth et al., 2017). And from a causal perspective, proper overlapping appear to be more appropriate compared with exact balancing (Zhang et al., 2020). We will leave these questions for future investigations.

A.2 Discussion on Model Generalization and Identifiability/Consistency

Model generalization and identification of BV-NICE. Generalization and identification are two complementary concepts, and they are equally important, as nobody wants to generalize over unidentifiable predictions. Discussions on model identifiability in the presence of latent in the context of counterfactual reasoning have been adequately covered in the works we discussed in Section A.1 above. Note that their setups are quite different from that of BV-NICE, which deals with more complex scenarios. Making strong, possibly unverifiable assumptions to ensure identifiability, conflicts our original intention to build an off-the-shelf causal inference framework that can deliver strong empirical performance

with minimal technical assumptions. In the main text, we have presented the generalization theory for the ITE estimation of BV-NICE. For identifiability, we make the distinction of (i) identification of population/policy average causal effect; and (ii) identification of the latent variable model. The former has been addressed by the works from N. Kallus and his colleagues in the causal context Bennett and Kallus (2019). For the latter, identifiability of variational inference has only been recently studied. We give a brief review on those two streams of work in the following paragraphs respectively.

Optimal balancing and policy evaluation. A recent stream of work, pioneered by N. Kallus and his colleagues, focused on the robustness to unidentifiability due to unobserved confounding Kallus and Zhou (2018), Kallus et al. (2018b, 2019), framed under the evaluation of policies using observational data. Unlike the setup adopted by BV-NICE, these work seeks to construct weights that are optimal in the sense they minimize an upper bound of *conditional mean square error* (CMSE) of the average causal effect Kallus (2017, 2018a,c). Via solving a *linear constraint quadratic program* (LCQP) in the reproducing kernel Hilbert space (which models the potential outcomes), such schemes optimize the worst case bound in an adversarial fashion, thereby ensuring consistency of the estimator Bennett and Kallus (2019). Note that such strategies are complementary to the development of causal latent variable models such as CE-VAE, BV-NICE, or matrix completion Kallus et al. (2018a): the primary goal is the identification of the average causal effect (policy evaluation), rather than the identification of the latent variable model. In other words, such models are built on the assumption that the corresponding latent variable model can be correctly identified. There are emerging discussions on the extra assumptions needed for the identification of latent variables under the general framework assumed by variational inference (see our discussion

below). And another key difference is that such policy evaluation studies focused on the ATE while BV-NICE targets the ITE.

Identifiable variational inference and implications for causal inference. There are emerging theories on the identifiability for the latent variables under the general framework assumed by variational inference, provided additional assumptions on weak supervision and applies to broader contexts Hyvarinen et al. (2018), Khemakhem et al. (2019). The identifiability is established in the sense that, one can recover the model parameters or the latent variables up to trivial transformations. The key difference to a standard (unidentifiable) variational auto-encoder (VAE) is that identifiable assumes additional supervision provided by observable $\boldsymbol{u}$ for the latent decomposition. More formally, it seeks to learn the following conditional generative model:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{u}) = p_f(\boldsymbol{x}|\boldsymbol{z}) p_{\mathbf{T},\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{u}), \tag{A.1}$$

where $\boldsymbol{\theta} = (f, \mathbf{T}, \boldsymbol{\lambda})$ are the parameters, and the prior on the latent variables $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{u})$ is assumed to be conditionally factorial given in the following form

$$p_{\mathbf{T},\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{u}) = \prod_i \frac{Q_i(\boldsymbol{z}_i)}{Z_i(\boldsymbol{u})} \exp\left[\sum_{j=1}^{k} \mathbf{T}_{ij}(\boldsymbol{z}_i) \boldsymbol{\lambda}_{ij}(\boldsymbol{u})\right[. \tag{A.2}$$

And the estimation similarly optimizes the conditional ELBO. Examples of such auxiliary $\boldsymbol{u}$ include temporal indices for sequence data and labels for supervised learning. The identification under equivalence class can be guaranteed under some fairly involved technical conditions, which is beyond the scope of current presentation, see Hyvarinen et al. (2018) for details. This technique has been applied to successfully resolve the causal graph with

observational data (Hyvarinen et al., 2018), and we can similarly construct identifiable BV-NICE via introducing auxiliary supervisions $\boldsymbol{u}$ based on domain knowledge. We leave the exploration for future work.

A.3 Discussion on The Synergies with Causal Discovery Literature

Synergies between recent advances in causal discovery and counterfactual reasoning should, and can be exploited. For many important applications, such as *reinforcement learning* (RL) (D'Amour, 2019, Lu et al., 2018), those two go hand in hand. Extensive studies have considered causal discovery in the presence of latent variables Ogarrio et al. (2016), Spirtes et al. (2013). This work focuses on the downstream of causal discovery: quantifying the causal effect under the assumed causal model in Figure **??**. It provides a compact and flexible way to model the joint distribution of observed proxies and latent confounders, allowing further likelihood decomposition into utility driven features. To the best of our knowledge, while variational schemes have been considered in the causal discovery literature Hyvarinen et al. (2018), our integration of balancing & the direct modeling of causal effect via Robinson factorization are novel, which have never been explored in causal discovery. Also, extension to the more general causal discovery applications will typically need non-trivial generalization of feature balancing techniques wrt non-binary interventions, a direction we seek to pursue in future work.

Another possible extension of this work that is inspired by causal discovery literature

is to explicitly account for the causal sufficient set Witte et al. (2020). The concepts such as identification of causal sufficient adjustment sets are rarely considered in the context of counterfactual reasoning. Such techniques can help our BV-NICE framework, which practiced an explicit trial of reformulating variational inference objective to achieve feature balance, to structurally model the important covariate X for engineering balanced features, and to trim sufficient features .

A.4 Discussion on Distribution Matching Schemes

We consider three major types of distribution matching schemes: information-theoretic divergence, integral probability metric and moment matching. Below we present a brief overview, and provide discussions on their relevance to causal estimation.

Information-theoretic divergence is one of the most popular metrics for evaluating distribution discrepancy. Formally, one can consider the general f-divergence Csiszár (1963), $\mathbb{D}_f(p_d \parallel p_G) \triangleq \int f\left[\frac{p_d(x)}{p_G(x)}\right[p_G(x)\,\mathrm{d}x$, where $f(\cdot):\mathbb{R}\to\mathbb{R}$ is a convex function satisfying $f(1)=0$, that summarizes the local discrepancy between $p_d(x)$ and $p_G(x)$. Recently, Goodfellow et al. (2014a), Nowozin et al. (2016) showed that distribution matching can be achieved via playing a mini-max game wrt the f-divergence. The most popular model from this category is the vanilla GAN, which optimizes Jensen Shannon divergence (JSD) and has been employed for almost all "adversarial counterfactual reasoning" procedures to date. The KL-divergence used in the proposed BV-NICE also belongs to the family, and we have embraced a more numerically stable estimation procedure through Fenchel mini-max

learning (Dai et al., 2018, Tao et al., 2019).

Integral probability metric (IPM) is a popular alternative to the divergence-based objectives, as their parameter updates can be either uninformative or numerically unstable, and divergence-based objectives may not be continuous wrt the generator parameters (Arjovsky and Bottou, 2017). IPM models seek to optimize an objective of the form $V_{\text{IPM}}(p_d, p_G; D) = \mathbb{E}_{X\sim p_d}[D(X;\omega)] - \mathbb{E}_{X'\sim p_G}[D(X';\omega)]$. When the critic are restricted to to Lip-1 family, IPM recovers the famous Wasserstein-1 distance, also known as the *earth mover's distance* (Rubner et al., 2000). Restricting optimization to a the Lip-1 family is a challenging task (Arjovsky et al., 2017, Gulrajani et al., 2017, Miyato et al., 2017), as such many instead solve the primal form of Wasserstein using Sinkhorn-type iterations (Cuturi, 2013). In CFR the authors have used this strategy to promote balancing (Shalit et al., 2017). One major issue with this solution is computational scalability, as quadratic computations are needed. Similar theoretical results have been derived for the benefit of IPM matching and the proposed KL-matching in the context of counterfactual reasoning. These bounds are not directly comparable and can only be assessed through experiments.

Moment matching is a classical strategy employed in statistical literature to resolve the balance issues. These practice, such as entropy balancing weights and empirical likelihood, often requires solving specialized optimization problems. A modern treatment of moment matching is to leverage the reproducing kernel Hilbert space embeddings, and compute the distance btw respective embeddings $\nu_p(x) \triangleq \mathbb{E}_{X\sim p}[\kappa(x, X)]$, commonly known as the maximal mean discrepancy (MMD) $\text{MMD}(p_d, p_G) \triangleq \|\nu_{p_d} - \nu_{p_G}\|_{\mathcal{H}}$, where $\mathcal{H}$ is the defining RKHS. Note that MMD is a special case of IPM and can be computed with a closed form using the kernel trick, but it also suffers the quadratic scaling like primal Wasserstein. It

can be translated into an algorithm that does not require the adversarial game for generative modeling Dziugaite et al. (2015), Li et al. (2015). Notably, MMD also belongs to the IPM family. Its use in counterfactual reasoning has also been discussed in Shalit et al. (2017). We note that off-the-shelf MMD does not work well for high-dimensional complex data (Bińkowski et al., 2018). In practice, good performance can only be achieved with careful hyperparameter tuning and by introducing auxiliary loss terms to the objective Li et al. (2017a), Zhang et al. (2017). And such insights can be borrowed for counterfactual applications.

A.5 Technical Proofs

A.5.1 Proof of Lemma 1

Proof. We have

$$
\begin{aligned}
& \epsilon_{CF}(h,\Phi) - [(1-u)\cdot\epsilon_F^{t=1}(h,\Phi) - u\cdot\epsilon_F^{t=0}(h,\Phi)] \\
= \; & [(1-u)\cdot\epsilon_{CF}^{t=1}(h,\Phi) + u\cdot\epsilon_{CF}^{t=0}(h,\Phi)] - [(1-u)\cdot\epsilon_F^{t=1}(h,\Phi) + u\cdot\epsilon_F^{t=0}(h,\Phi)] \\
= \; & (1-u)\cdot[\epsilon_{CF}^{t=1}(h,\Phi) - \epsilon_F^{t=1}(h,\Phi)] + u\cdot[\epsilon_{CF}^{t=0}(h,\Phi) - \epsilon_F^{t=0}(h,\Phi)] \\
= \; & (1-u)\int_{\mathcal{X}} l_{h,\Phi}(x,1)(p^{t=0}(x) - p^{t=1}(x))dx + u\int_{\mathcal{X}} l_{h,\Phi}(x,0)(p^{t=1}(x) - p^{t=0}(x))dx \\
= \; & (1-u)\int_{\mathcal{R}} l_{h,\Phi}(\Psi(r),1)(q_0(r) - q_1(r))dr + u\int_{\mathcal{R}} l_{h,\Phi}(\Psi(r),0)(q_1(r) - q_0(r))dr \\
\leq \; & \tfrac{1-u}{2M}TV(q_0,q_1) + \tfrac{u}{2M}TV(q_1,q_0) \\
= \; & \tfrac{1}{2M}TV(q_1,q_0) \leq \tfrac{1}{2M}\sqrt{\tfrac{1}{2}D_{KL}(q_0||q_1)}
\end{aligned}
$$

The last inequality is a result of the Pinsker's inequality. □

A.5.2 Proof of Theorem 2

Proof. This result is immediate by Lemma 1 and the following inequality

$$\epsilon_{PEHE}(h,\Phi) \leq 2\epsilon_{CF}(h,\Phi) + 2\epsilon_F(h,\Phi) - 2\sigma_Y^2, \tag{A.3}$$

which is a consequence of Theorem 1 from Shalit et al. (2017). □

A.6 Practical Variants of BV-NICE

β-VAE (Higgins et al.) reweighs the KL term in the ELBO term, resulting

$$\ell_{\beta\text{-VAE}} = \frac{1}{n}\sum_i \mathbb{E}_{\boldsymbol{Z}\sim q}[\log p(\boldsymbol{x}_i|\boldsymbol{Z})] - \beta \mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x}_i) \parallel p(\boldsymbol{z})), \tag{A.4}$$

where $\beta \geq 0$. When $\beta = 0$, it recovers the classical auto-encoder (Hinton and Salakhutdinov, 2006, Hinton and Zemel, 1994).

AAE (Makhzani et al., 2015) instead seeks to match the aggregated posterior $q(\boldsymbol{z}) \triangleq \int q(\boldsymbol{z}|\boldsymbol{x})p_d(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ to the prior. Different from the original AAE, our feature encoder is stochastic, accounting for feature uncertainties. Particularly,

$$\ell_{\text{AAE}} = \frac{1}{n}\sum_i \mathbb{E}_{\boldsymbol{Z}\sim q}[\log p(\boldsymbol{x}_i|\boldsymbol{Z})] - \beta \mathrm{KL}(q(\boldsymbol{z}) \parallel p(\boldsymbol{z})) \tag{A.5}$$

where $\beta \geq 0$. A similar model is investigated in Fang et al. (2019) under the name *implicit VAE* (iVAE) for text generation. The general observation is that enforcing the ELBO KL term ($\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z}))$) seems to harm the performance, possibly due to over-generalization.

See Table A.2 for comparisons of the above variants and other alternatives.

A.7 Sampling Latent Variable

An estimate of the causal effect is given by

$$\tau(\boldsymbol{x}) \approx \frac{1}{m} \sum_{j} \tau(\boldsymbol{z}'_j), \tag{A.6}$$

with $\{\boldsymbol{z}'_j\}_{j=1}^{m}$ sampled from the posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$. This section primary focus on the case when the approximate posterior is approached in the form of $q_\phi(\boldsymbol{z}|\boldsymbol{x}, y, t)$, such that direct sampling from the (marginal) approximate posterior is infeasible given only covariate $\boldsymbol{x}$ observations [3]. The following strategies to marginalize over (t, y) have been considered in this work:

- **MCMC** When $\boldsymbol{x}$ is modeled, directly sample from $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ using standard MCMC. The primary concern for this strategy is computational overhead. Both gradient-based or rejection-based sampling schemes will impose considerable overhead, plus there is training effort for modeling $p_\theta(\boldsymbol{x}|\boldsymbol{z})$.

- **Auxiliary outcome and propensity models** This strategy is considered by Louizos et al. (2017), where to sample $\boldsymbol{z}$ given $\boldsymbol{x}$, one first sample (t', y') based on learned auxiliary models, then feed the synthetic pairs $(\boldsymbol{x}, t', y')$ to q_ϕ to sample $\boldsymbol{z}'$. More specifically, we partition y into multiple bins, and treat it as categorical variable. We then learn $p_\phi(y|\boldsymbol{x})$ and $p_\phi(t|\boldsymbol{x})$ via standard regression. Our major concerns with this strategy is that, (a) learning auxiliary model might introduce additional bias; (b) sampling through the auxiliary models

[3] Note that in the main text we advocate the use of approximate posterior of the form $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, which trades bound sharpness for sampling flexibility and reduced modeling bias, which turns out to be more favorable. For completeness we present discussions on sampling the infeasible posterior approximation above.

induces additional computation overhead.

• **Direct neural sampler** To more efficiently sample $\boldsymbol{z}$ given $\boldsymbol{x}$, we similarly use a neural sampler $r_{\phi'}(\boldsymbol{z}|\boldsymbol{x})$, defined by $\boldsymbol{z} = G_{\phi'}(\boldsymbol{\xi}, \boldsymbol{x}), \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and train it to match the marginal distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$. We again match conditional distributions $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and $r_\phi(\boldsymbol{z}|\boldsymbol{x})$ wrt the KL, through a conditional critic $\vartheta_{\psi'}(\boldsymbol{z}, \boldsymbol{x})$. In particular, we solve

$$\min_{\phi'} \{ \max_{\psi'} \{ \sum_i (\mathbb{E}_{Z \sim r_{\phi'}(\boldsymbol{z}|\boldsymbol{x})} [\vartheta_{\psi'}(\boldsymbol{Z}, \boldsymbol{x}_i)] - \mathbb{E}_{\boldsymbol{Z}' \sim q_\phi(\boldsymbol{z}|\boldsymbol{x}_i, y_i, t_i)} [\exp(\vartheta_{\psi'}(\boldsymbol{Z}'_i, \boldsymbol{x}_i))]) \} \}. \tag{A.7}$$

A.8 Experimental Setups

Neural network setup We use Xavier initializer for network initialization, and apply *relu* nonlinearity except for the output layer. No batch-normalization or drop-out is used, which may further improve performance. For all critic functions, we apply an additional scaled tanh activation at the output, which greatly stabilizes training dynamics. Unless otherwise specified, we use the AAE variant of BV-NICE and fix the latent dimensions to two throughout our experiments.

The grid search parameters for the dissecting BV-NICE experiment is listed below.

- Noise level: $\eta \in \{2^k\}_{k=1}^6 + \{10^{-k}\}_{k=0}^4 + \{0\}$
- Imbalance level: $\lambda \in \{10^{-k}\}_{k=-2}^4 + \{0\}$

We show the response surface of BV-NICE wrt (η, λ) in Figure A.1. We found that BV-

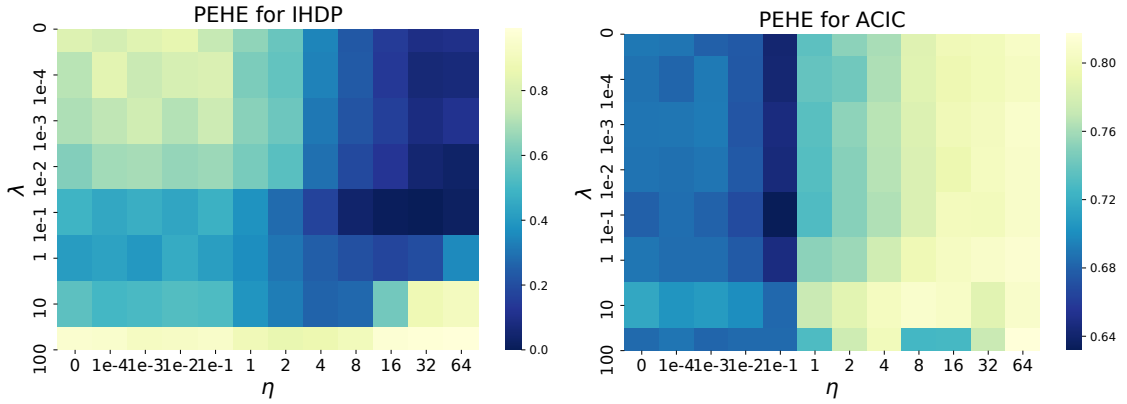

Figure A.1: BV-NICE performance sensitivity to noise and imbalance parameters.

NICE is more sensitive to the injected noise intensity relative to the imbalance [4], which supports our view that properly accounting for the uncertainty merits causal inference. We remark that although in theory a neural net should be able to rescale the injected noise to its optimum regardless of η, in practice η does regulate the level of uncertainty for the output.

Data normalization We found that results can be sensitive to the dynamical range of original inputs, especially for the outcome. Additionally, original PEHE can be misleading as results may be dominated by a few outliers. To ameliorate these issue, we report the normalized PEHE by rescaling the observed outcome to zero mean and unit variance based on the empirical moments of the training set. We also apply this normalization to the observed confounders as a standard pre-processing step. While this causes some discrepancies with existing results in literature, we believe it establishes a more fair comparison btw models.

[4] This is understandable as noisy representation promotes overlapping on its own.

A.9 Categorization of Causal Effect Learners

- T**-learner: Two** separate learners $(\mu_0(\boldsymbol{x}), \mu_1(\boldsymbol{x}))$ are used for the outcome prediction under different interventions.
- S**-learner:** A **single** learner $\mu(\boldsymbol{x}, t)$ is used for the prediction for treatment outcomes given the confounders.
- R**-learner:** A learner that directly models the causal effect $\tau(\boldsymbol{x})$ rather than resorting to the difference between conditional outcomes $\mu_t(\boldsymbol{x}), t \in \{0, 1\}$. Typically the Robinson decomposition is used (Nie and Wager, 2017).
- X**-learner:** Meta-learner for the causal effects via regressing on the pilot estimates produced by other causal estimation models (Künzel et al., 2019). We do not investigate the use of X-learner in this study.

A.10 Baseline Implementations

- **(Regularized) Ordinary Least Squares (OLS):** We implemented a regularized version of OLS (*a.k.a* ridge regression) using the python *scikit-learn* package based on the T-learner setup. We vary the regularization strength from 10^{-5} to 10^{3} in log scale, at a step-size of 10 for each. The best model is picked based on cross-validation. **We found this strategy yields very reasonable performance, significantly out-performing vanilla (un-regularized) OLS and even some of the state-of-the-art alternatives.**

- **Counterfactual Regression (CFR):** Our CFR codebase is derived from `https://github.com/oddrose/cfrnet`, which includes implementations for *treatment agnostic representation network* (TARNet), *balancing neural network* (BNN) (Johansson et al., 2016) and two variants of CFR (Shalit et al., 2017) (Wasserstein & MMD regularizer, rvsp). In our exploratory experiments, we find the CFR with Wasserstein regularizer usually delivers more favorable performance compared with its counterparts. As such, we choose CFR-Wasserstein as our representative from the CFR model family.

- **Causal Random Forest (Causal RF):** In Wager and Athey (2018b) the authors proposed two types of Causal RF, namely *Double Sample Forest* (DSF) and *Propensity Forest* (PF). Both variants are built on the concept of "honesty", with the former leveraging non-overlapping samples for tree-construction and causal effect estimation rsvp, the latter builds the tree solely based on propensity prediction. We adopted the python implementation for both types of Causal RF from the following version of python *scikit-learn* package (`https://github.com/kjung/scikit-learn`). The PF results are reported in the main text, with the DSF results summarized in the SM here.

- **Bayesian Additive Regression Tree (BART):** We use the *bartpy* package for python BART implementation (`https://github.com/JakeColtman/bartpy`). More explicitly, the *ReadOneTrees* branch was compiled on *Jan 30, 2020* for our experiments, as we found stability issues with the *master* branch. We set the number of BART trees so that the cost of computations is comparable with other baselines.

- ***R*-learner:** Our main implementation for the *R*-learner is based on both *scikit-learn*

and *tensorflow* packages. Pilot estimates of the mean outcome and propensity score models, *i.e.*, the nuisance models, are obtained by Random Forest regressor and classifier from the *scikit-learn* ensemble learner package. We use 50 trees for both cases. The causal effect $\tau(\boldsymbol{x})$ is modeled using MLP implementation in *tensorflow*, to be consistent with CFR and BV-NICE implementations. We also experimented with MLP estimator with cross-validation for the nuisance models, and found RF implementation to be more reliable.

- **GANITE:** To complement our comparison of competing solutions, we also include *GANITE*, a prominent example from the category of adversarial causal effect estimation procedures. We extracted the GANITE implementation from the python *perfect_match* package (`https://github.com/d909b/perfect_match`, (Schwab et al., 2018b)), and tuned its parameters for optimal performance. We report the results with unit number set to 30 and α, β set to 0.1.

- **Balancing Weights (BW):** We also experimented with explicit weighting schemes for covariate balancing. In particular, we consider propensity score based weights Li et al. (2018) (*e.g.*, inverse propensity weighting (IPW), matching weights (MW), overlapping weights (OW)), and entropy based (EB) weights [5] (Hainmueller, 2012). For propensity score based models, we use random forest classifier to model the propensity score. To avoid overfitting, we limit leaf size to a minimum of 5 samples. The estimated weights are supplied to a weighted regression forest to estimate $\mu(\boldsymbol{x}, t)$.

[5] Note EB weights ensures that the weighted first moments are matched between treatment groups.

A.11 Summary of Datasets

- **IHDP1000** The IHDP data comes from a study investigating the effect of specialist home visits on the cognitive test scores. This study consists of $1,000$ simulations of different treatment and outcomes mechanism. We use exactly the same data generating process in Shalit et al. (2017). With the semi-synthetic data, we know the counterfactual outcomes $Y_i(1-T_i)$ and the true outcome function $f_t(x)$. The dataset has 747 units with 139 treated and 608 control, 25 covariates. We use the training/test split from Shalit et al. (2017), and further hold out 30% of the training samples for validation purposes.

- **ACIC2016** This is a benchmark dataset released by *Atlantic Causal Inference Competition*. It (Dorie et al., 2019) involved 77 semi-synthetic datasets which shares the same covariates $\boldsymbol{X}$ but with individual outcomes and assignment mechanism $(\boldsymbol{Z}, T)$. Features were extracted from real-world observational study, while the causal mechanism are contributed by domain experts to mimic real data. This dataset include $4,802$ samples and each simulation is replicated 100 times to assess algorithmic stability.

- **JOBS** This dataset comes from a study aiming at assessing whether attending a job training program can help people get employed (LaLonde, 1986).. The data consists of a randomized study and an observational study. The dataset includes 17 covariates and 722 units, with 297 units in the treated group and 425 in the control group, from the study where units receive training randomly and 2490 control units from the PSID (Panel Study of Income Dynamics) study.

- **SHRP2** This data is derived from the Second Strategic Highway Research Plan Nat-

uralistic Driving Study to evaluate the risk factors of car accidents. Volunteer drivers are recruited across the country, with continuous recording devices (video & telemetrics) installed to their vehicles for monitoring driving behavior in natural status. The videos of 6-second around the onset of precipitation event (5 seconds prior, 1 second after) were visually examined by trained data coders to record drivers' behavior and driving environmental variables. The normal driving segments are 6-second episodes of driving segment randomly selected from the more than 1 million hours of continuous driving videos. For those video segments, the driver demographic information, driving behavior and driving environment information were extracted.

A.12 Additional Experimental Results and Analysis

Importance of propensity score We have also observed that, even without the propensity likelihood term $p_\theta(t|\boldsymbol{z})$ in the objective, BV-NICE can still learn CE well. Using the Robinson residual alone also learns a fairly good approximation to the true propensity, and the built-in uncertainty helps to guard against over-fitting: BV-NICE based estimator remains valid approximation throughout training, while typical under-regularized learner quickly fits to the training labels. We also vary the weight of the propensity likelihood term in BV-NICE, and find that in general a small yet none-zero (*e.g.*, $\kappa = 0.01$) weight gives the best performing model.

Learning propensity score To see the effectiveness of Robinson decomposition in

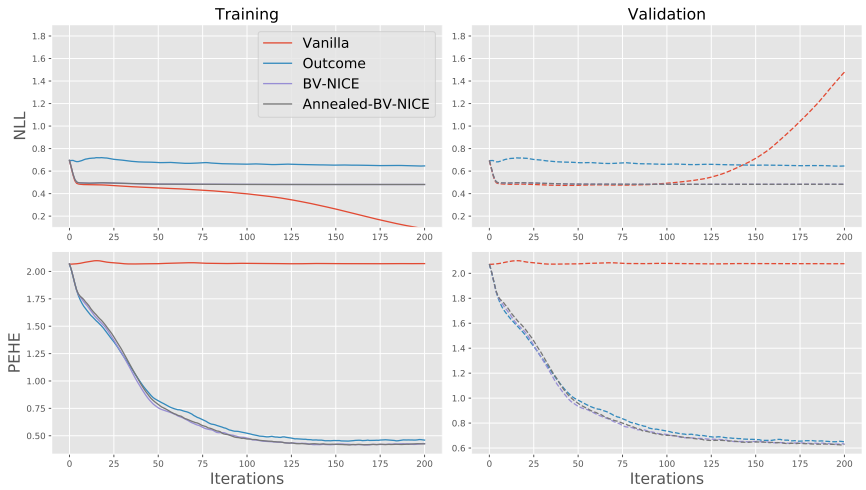

Figure A.2: Ablation study on propensity score for BV-NICE. It is necessary to model both propensity and outcome for accurate causal estimation.

learning propensity score, we experimented with the following setups:

$$\mathcal{L} \quad = \quad \mathbb{E}_{\boldsymbol{Z}\sim q(\boldsymbol{z}|\boldsymbol{x})}[\rho \log p(y|\boldsymbol{Z}) + \gamma \log p(t|\boldsymbol{Z}) + \mathcal{L}_{\text{remaining}}] \tag{A.8}$$

- Vanilla PS: $\mathcal{L} = \log p(t|\boldsymbol{x})$
  - Direct learning of propensity score using cross-entropy with an MLP
- Outcome only: $\rho = 1, \gamma = 0$
- BV-NICE: $\rho = 1, \gamma = 1$
- Annealed BV-NICE: $\rho = 1, \gamma = 10^{-2}$

We use a flexible MLP to model the propensity score and plot the average NLL together with PEHE on both training and test set against training iterations in Figure A.2 using

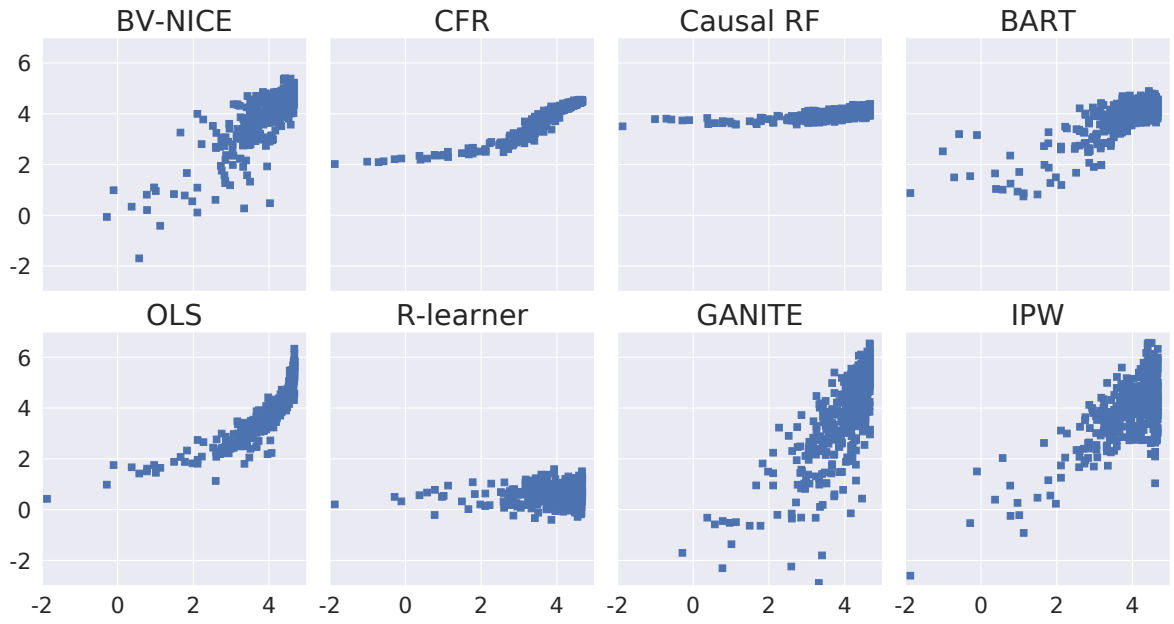

Figure A.3: IHDP1000: estimated against ground-truth causal effect for different models. IHDP1000. Note that while vanilla PS overfits as training progresses, BV-NICE variants showed inherent robustness against overfitting. We have also noticed that, on its own, the Robinson decomposition does not capture the propensity score well (*i.e.*, the outcome only model), and its ITE estimate is also less accurate. As such, the propensity score loss is necessary, and properly weighting it seems appropriate.

Comparison of learned CE In Figure A.3 we compared the learned individual causal effects from different methods. Ideally the estimated CE should align along the diagonal. Performance varies, and a smaller PEHE does not always imply a good estimate based on visual inspection. In this case, our BV-NICE performs reasonably well. While some of the alternatives (CFR, Causal RF, OLS) yields estimates with tighter spread, their distributions tend to be skewed and deviate from the expected diagonal. To note, the success of R-learner critically relies the accuracy of (separately) learned propensity. In this case, a total failure

Table A.1: Extended comparisons on semi-synthetic datasets

	IHDP1000		ACIC2016	
$\sqrt{\epsilon_{PEHE}}$	With-in-sample	Out-of-sample	With-in-sample	Out-of-sample
OLS	0.29 ± .09	0.30 ± .11	0.52 ± .13	0.65 ± .16
CFR	1.47 ± .35	1.46 ± .36	0.52 ± .14	0.90 ± .26
BART	0.30 ± .08	0.33 ± .11	0.58 ± .12	0.70 ± .17
Causal RF	0.63 ± .01	0.63 ± .16	0.68 ± .01	0.81 ± .02
EB	0.37 ± .09	0.42 ± .16	**0.49 ± .15**	0.63 ± .26
IPW	0.40 ± .12	0.43 ± .17	**0.50 ± .15**	0.64 ± .26
DSF	0.74 ± .28	0.73 ± .30	0.71 ± .32	0.84 ± .40
GANITE	1.04 ± 1.04	1.04 ± 1.06	0.72 ± .33	0.83 ± .40
R-learner	1.27 ± .29	1.26 ± .31	0.83 ± .12	0.91 ± .15
BV-NICE	**0.20 ± .04**	**0.20 ± .06**	**0.50 ± .13**	**0.62 ± .17**

is observed. While for some of the test cases R-learner worked pretty well (not shown).

Latent representations We visualize the learned latent representation of BV-NICE and compare it with those produced from competing solutions. In Figure A.4 we compare the learned distributions from TARNet, CFR and BV-NICE. The representation learned via BV-NICE shows more regularity, as instructed by the prior. To highlight BV-NICE better encourages overlapping via adopting the stochastic representation, we plot the distributions of inferred latent confounders for two units coming from opposite treatment groups in Figure A.5. This confounder overlap prevents flexible learners to overfit particular units.

Additional results on semi-synthetic datasets Here we provide additional comparisons with baselines not summarized in the main text. Table A.1 further includes comparisons with the following baselines: (a) GANITE, which is based on adversarial matching for causal estimation; (b) weighted regression, including IPW and EB weights; (c) vanilla

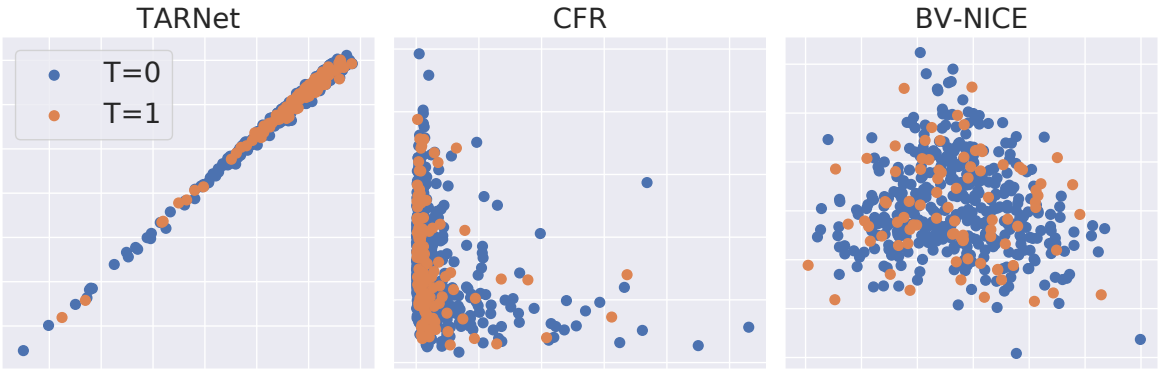

Figure A.4: Comparison of learned representations. Note we sampled representations for BV-NICE.

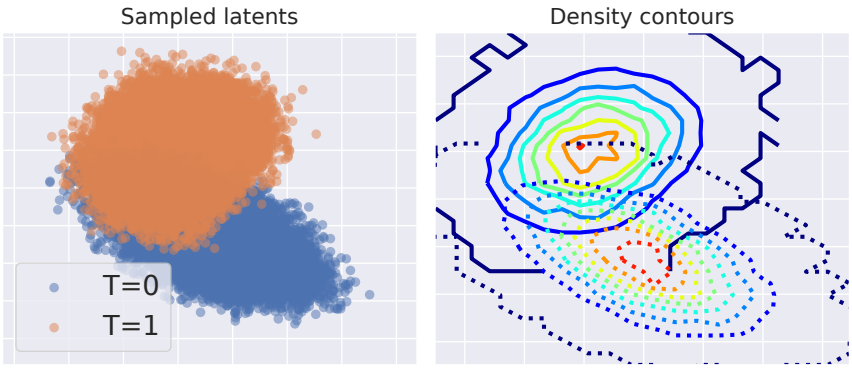

Figure A.5: Overlap of stochastic representation learned by BV-NICE.

R-learner; (d) double sample causal forest. Implementation details are summarized in Section A.10. These baselines comprehensively covers representative examples of alternative generative causal models, weighting schemes and standard R-learning, which complements our comparison. In Figure A.6, we further compare the performance on individual test problems from the ACIC dataset, as we did in the main text.

Overall BV-NICE still performs favorably compared with the extended baselines. The only exception is that EB worked slightly better on the ACIC. We found GANITE suffered

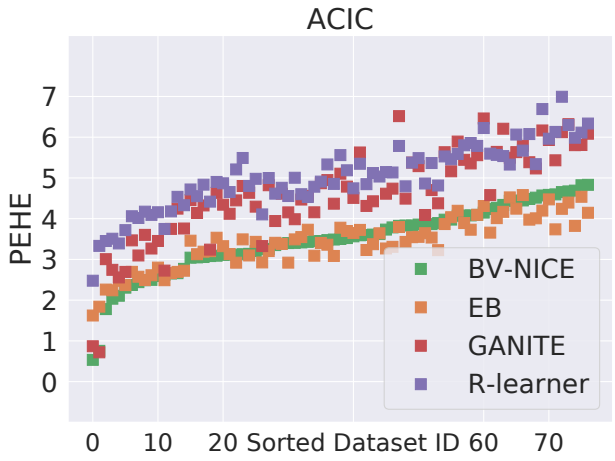

Figure A.6: Extended results visualization on ACIC2016. Lower is better ↓. Index sorted for ACIC based on BV-NICE to facilitate visualization.

severely from stability issues, manifested by the extreme variance observed. Note that stability is a major criticism against adversarial schemes, while the Fenchel mini-max treatment adopted by BV-NICE successfully overcomes this difficulty. One possible explanation for the performance gap observed in here is that most of the causal estimation procedures we experimented do not have a proper stopping rule to guard against overfitting. For example, we experimented with the built-in stopping criteria for GANITE, and observed the results can be highly volatile. For CFR, it performed very well on the sub-problems we sampled for hyper-parameter tuning, yet showed far less satisfactory results on the full sample. We remark that the performance for some of the schemes might be limited by the particular implementation we adopted, gains might be expected with extensive recoding. However, this requirement might be too demanding for practitioners seeking off-the-shelf causal estimation solutions. Our BV-NICE highlights relatively easy implementation and robust performance.

Table A.2: Comparing variants of BV-NICE.

$\sqrt{\epsilon_{PEHE}}$	IHDP100	
	With-in-sample	Out-of-sample
Full-VAE	$1.02 \pm .38$	$1.00 \pm .39$
β-VAE ($\beta = 10^{-2}$)	$1.03 \pm .37$	$1.01 \pm .38$
β-VAE ($\beta = 10^{-1}$)	$1.05 \pm .34$	$1.03 \pm .35$
β-VAE ($\beta = 1$)	$1.03 \pm .37$	$1.03 \pm .36$
NRD-VAE	$0.26 \pm .06$	$0.28 \pm .07$
BV-NICE ($\kappa = 0$)	$0.25 \pm .07$	$0.28 \pm .08$
BV-NICE ($\kappa = 10^{-2}$)	$\mathbf{0.25 \pm .05}$	$0.28 \pm .06$
BV-NICE ($\kappa = 10^{-1}$)	$0.26 \pm .06$	$0.28 \pm .06$
BV-NICE ($\kappa = 1$)	$0.26 \pm .06$	$\mathbf{0.27 \pm .06}$

Comparison of BV-NICE variants To justify our choice of BV-NICE implementation, we consider the following variants for comparison.

- Full-VAE: $q(\boldsymbol{z}|\boldsymbol{x}, y, t), \beta = 1$
- β-VAE: $q(\boldsymbol{z}|\boldsymbol{x}), \beta \in \{10^{-2}, 10^{-1}, 1\}$
- NRD-VAE: $q(\boldsymbol{z}|\boldsymbol{x})$, without Robinson decomposition
- BV-NICE: $q(\boldsymbol{z}|\boldsymbol{x})$+AAE, $\kappa \in \{10^{-2}, 10^{-1}, 1\}$

Table A.2 summarizes results. The original VAE setup is observed to severely harm causal estimation.

Summaries for NDS analysis. In Figure A.7, we compare individual causal effect estimates from CFR, BV-NICE and standard two learner. Note CFR estimates are highly

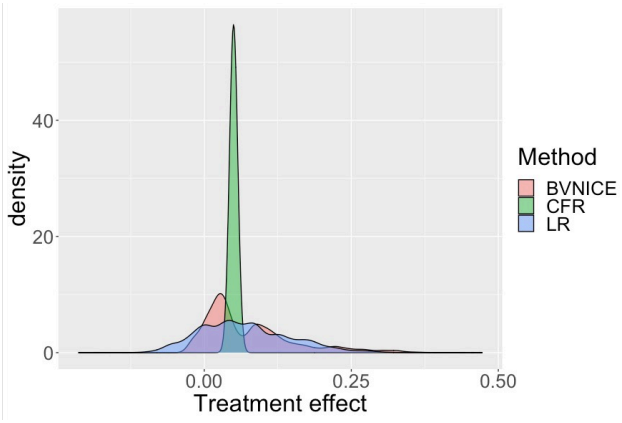

Figure A.7: Comparison on the NDS data.

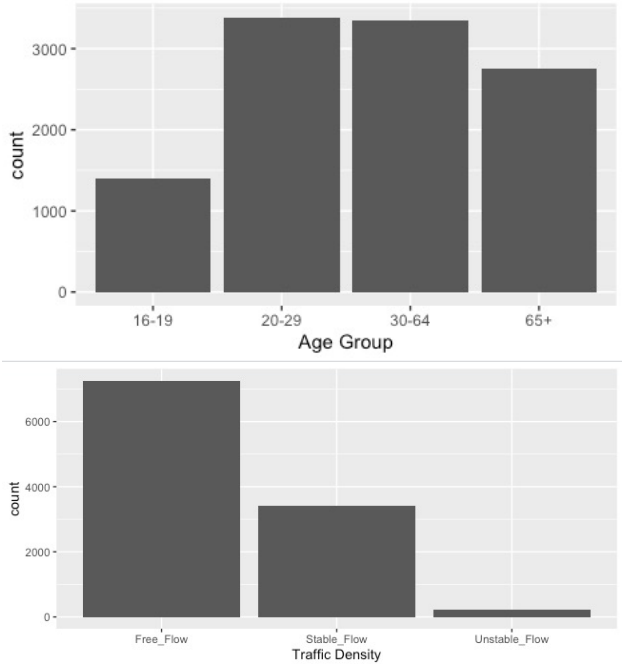

Figure A.8: NDS data summary.

concentrated, which seemed problematic given prior domain knowledge. The two learner estimates appear to be more flat compared BV-NICE estimates. We note more scrutiny is required to make final conclusions, as due to the low prevalence of crashes the validation

might not be reliable, which probably explains the erratic estimates made by CFR.

Challenges from rare-event modeling. Dealing with severe class imbalance poses a major challenge for real-world applications, especially when the accurate classification and generalization of minority classes is of primary interest, such as traffic risk analysis discussed above. Distributions shifts are more likely for under-represented populations, thereby less generalizable with standard learning schemes. This will potentially lead to algorithmic bias: instead of capturing robust, generalizable features, a learning algorithm tend to settle with spurious correlations that is caused by the representation shift. In our setup, the representation matching scheme helps to alleviate this issue, by removing apparent disparities between the distributions. And we show some preliminary success with the demonstration on the NDS driving risk analysis. A more formal, dedicated presentation is being prepared separately. We are also actively exploring alternative variational and causal perspectives to further improve rare-event modeling (Chen et al., 2020, Xiu et al., 2021).

Appendix B

RICE

B.1 RICE Experiments

Table B.1: RICE: cross-validation along the regularization path

	In-sample NLL	Out-sample NLL					
λ	Avg.	CV-0	CV-1	CV-2	CV-3	Avg.	$\sqrt{\epsilon_{PEHE}}$
0.00	0.7022	0.7515	0.6743	0.6687	0.7787	0.7183	0.0262
0.10	0.7034	0.7501	0.6819	0.6671	0.7861	0.7213	0.0260
0.20	0.7015	0.7499	0.6740	0.6651	0.7889	0.7195	0.0317
0.30	0.7028	0.7506	0.6744	0.6682	0.7846	0.7195	0.0231
0.40	0.7005	0.7490	0.6710	0.6664	0.7820	0.7171	0.0214
0.45	0.6999	0.7465	0.6720	0.6659	0.7848	0.7173	0.0265
0.50	0.7016	0.7480	0.6746	0.6665	0.7859	0.7187	0.0187
0.55	0.6999	0.7549	0.6733	0.6627	0.7828	0.7184	0.0182
0.57	0.7007	0.7482	0.6746	0.6659	0.7852	0.7185	0.0140
0.60	0.7017	0.7464	0.6741	0.6682	0.7873	0.7190	0.0324
0.65	0.6995	0.7453	0.6693	0.6663	0.7852	0.7165	0.0310
0.70	0.7003	0.7488	0.6729	0.6647	0.7886	0.7188	0.0473
0.80	0.7015	0.7489	0.6714	0.6668	0.7909	0.7195	0.1062
0.90	0.7006	0.7482	0.6726	0.6655	0.7905	0.7192	0.1721
1.00	0.7027	0.7445	0.6735	0.6674	0.7819	0.7168	0.2171

B.2 ICLR Experiments

Table B.2: ICLR: cross-validation along the regularization path

	Out-sample NLL						
λ	CV-0	CV-1	CV-2	CV-3	Avg.	Weighted Avg.	$\sqrt{\epsilon_{PEHE}}$
0.00	1.7480	1.5794	1.1729	1.0259	1.3815	1.3158	0.4043
0.05	1.7786	1.4957	1.1624	1.0343	1.3677	1.3057	0.4420
0.10	1.7018	1.4718	1.1326	1.0434	1.3374	1.2690	0.3783
0.15	1.7231	1.4568	1.1414	1.0122	1.3334	1.2766	0.3382
0.20	1.6967	1.4326	1.1028	1.0057	1.3095	1.2466	0.3966
0.25	1.6656	1.4096	1.1044	1.0127	1.2981	1.2368	0.3598
0.30	1.6948	1.4442	1.0641	1.0178	1.3052	1.2254	0.3778
0.35	1.6158	1.3737	1.0867	1.0052	1.2704	1.2116	0.3541
0.40	1.5755	1.3715	1.0763	1.0448	1.2670	1.2048	0.3770
0.45	1.5169	1.2861	1.0802	1.0483	1.2329	1.1875	0.3748
0.50	1.4733	1.3360	1.0438	1.0530	1.2265	1.1577	0.3053
0.55	1.4538	1.2932	1.0253	1.0446	1.2042	1.1407	0.3222
0.60	1.4818	1.2125	1.0347	1.0265	1.1889	1.1413	0.2961
0.65	1.5182	1.1879	1.0240	1.0097	1.1850	1.1387	0.3798
0.70	1.4609	1.1551	1.0049	1.0397	1.1652	1.1149	0.2843
0.75	1.3715	1.0965	1.0034	1.0415	1.1282	1.0886	0.2975
0.80	1.3494	1.0971	0.9730	1.0348	1.1135	1.0662	0.3252
0.85	1.3638	1.1041	0.9566	1.0529	1.1193	1.0625	0.2709
0.90	1.2824	1.0729	0.9542	1.0685	1.0945	1.0424	0.2302
0.91	1.3243	1.0586	0.9422	1.0394	1.0911	1.0409	0.2901
0.92	1.3246	1.0094	0.9670	1.0277	1.0822	1.0480	0.2659
0.93	1.2944	1.0167	0.9634	1.0243	1.0747	1.0412	0.2409
0.94	1.2392	1.0700	0.9453	1.0511	1.0764	1.0274	0.2755
0.95	1.2431	1.0491	0.9444	1.0641	1.0752	1.0260	0.2300
0.96	1.2491	1.0402	0.9443	1.0418	1.0688	1.0255	0.3011
0.97	1.2529	1.0162	0.9451	1.0635	1.0694	1.0249	0.2394
0.98	1.1947	0.9998	0.9348	1.0498	1.0448	1.0030	0.2524
0.99	1.1668	1.0194	0.9244	1.0633	1.0435	0.9966	0.3505
1.00	1.2999	1.3093	1.3111	1.3017	1.3055	1.3077	1.4646

B.3 Data Heterogeneity in NDS

Due to different sampling designs, there is substantial difference in covariates distribution between the case-cohort sample and the case-crossover sample. The NDS case-crossover sample includes 530 crashes and 1,638 matched baselines. For each crash, up to four baselines were sampled from the same driver under similar driving conditions (`day of week`, `weather`, `traffic density`, `locality`, etc.) within three months prior to the crash. Additional matching criteria includes `time of day` and `speed` (Owens et al., 2018). Each crash was manually validated by trained data reductionist and assigned to a severity level (Hankey et al., 2016). The case-crossover sample includes 78 most severe (*level 1*) crashes , 99 police reportable (*level 2*) crashes, and 353 minor (*level 3*) crashes. For each crash, the number of matched baselines varies from one to four. There are 293 crashes matched with 4 baselines, 80 crashes matched with 3 baselines, 69 crashes matched with 2 baselines, and 88 crash matched with only one baseline. In total, 1,638 baselines were included in the sample.

Table B.3: Number of matched baselines by crash severity

Number of matched baselines	1	2	3	4	Total
I. Most Severe	14	11	10	43	78
II. Police reportable Crash	22	18	11	48	99
III. Minor.Crash	52	40	59	202	353
Total cases	88	69	80	293	530
Total matched baselines	88	138	240	1,172	1,638

The matching mechanism of case-crossover sample inevitably shifts the population towards risk-prone drivers and driving contexts. As is shown in Figure B.1, the case-crossover sample includes higher percentage of young (16-19 years old), low income drivers and drivers

from Florida. For driving contexts, the case-crossover sample contains more driving segments at local roads, those with higher traffic density, and during weekdays (Figure B.2).

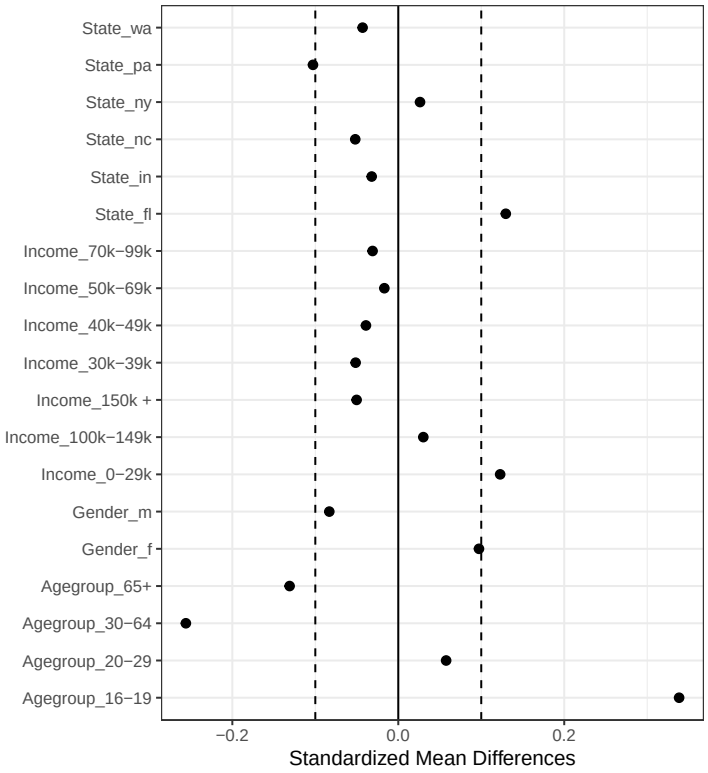

Figure B.1: Covariates shift between population: demographic distributions

The heterogeneity in the population is inevitably reflected in the risk estimation inconsistency (Figure B.3). In general, the crash OR's estimated from case-crossover sample are lower than case-cohort sample. This pattern holds for different models (crude OR's from the two-way table, *mixed-effect logistic regression* (MLR), and *conditional logistic regression* (CLR)) and matching criteria (two different sources of case-crossover samples: CC and CCa).

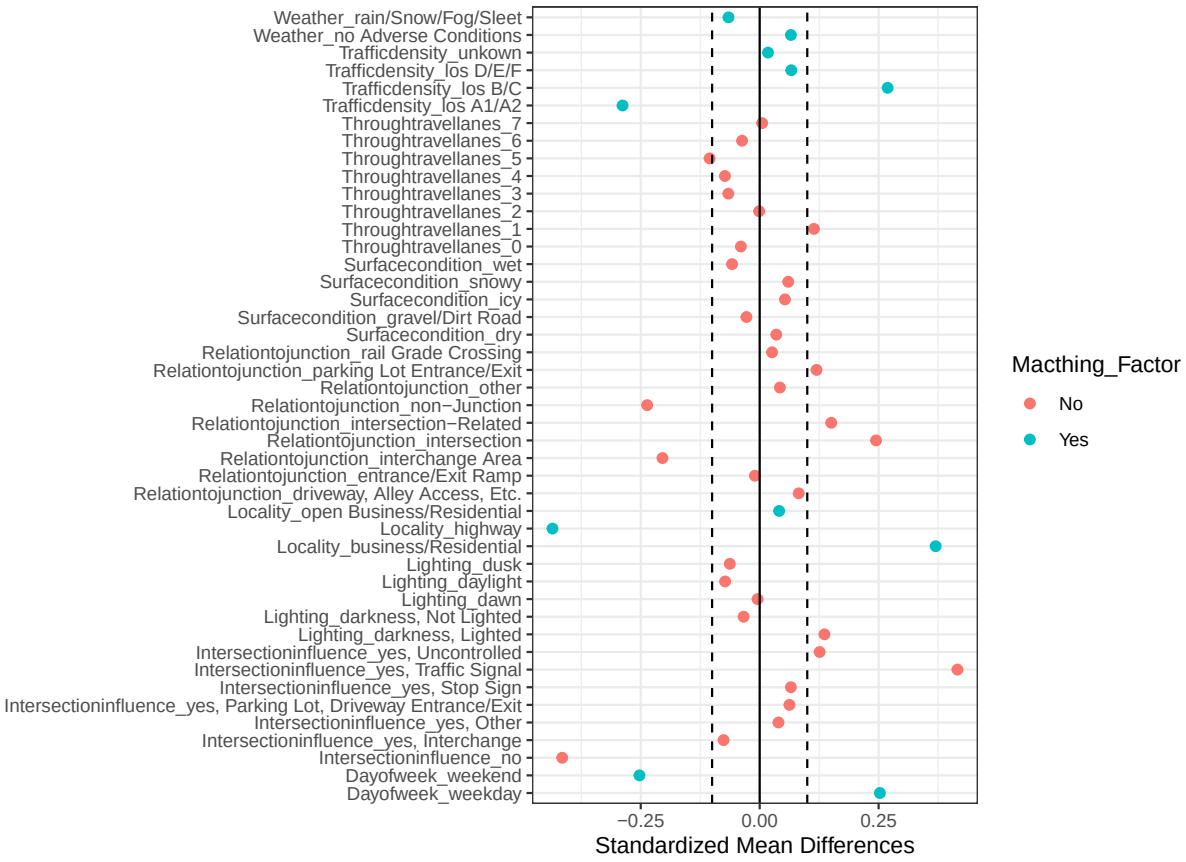

Figure B.2: Covariates shift between population: driving contexts distributions

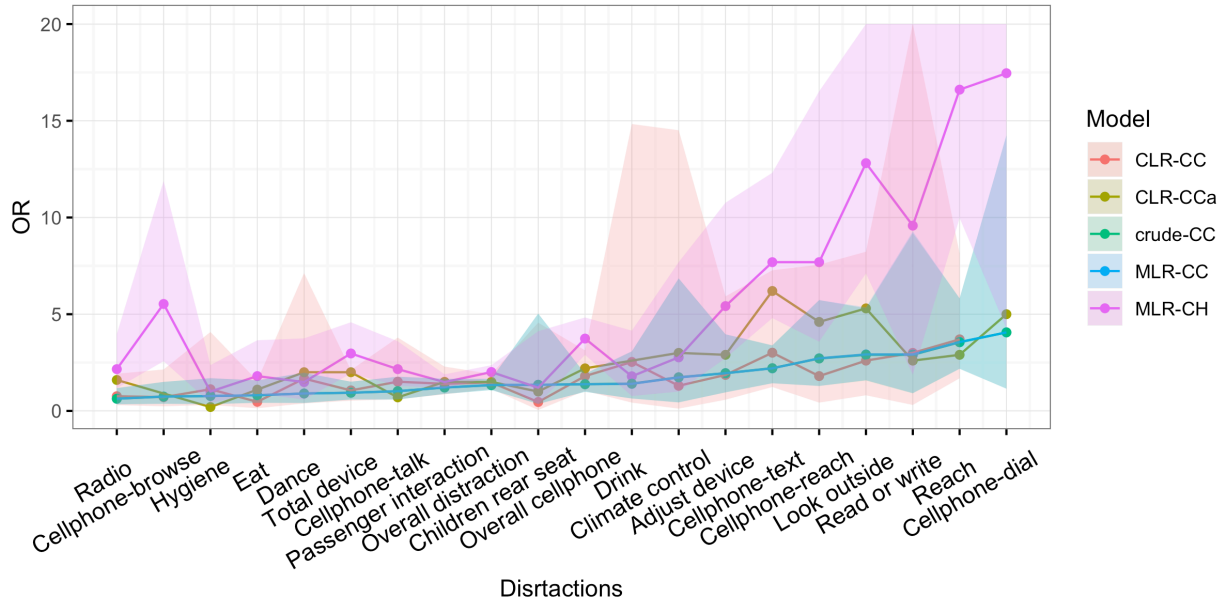

Figure B.3: OR by distractions and models

Table B.4: OR by distraction and models

	Case-crossover (CLR)		Case-crossover (MLR)		Case-crossover (Crude)		Case-cohort (MLR)	
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
Overall distraction	1.47	(1.17 ,1.87)	1.34	(1.08 ,1.66)	1.33	(1.08 ,1.66)	2.01	(1.72 ,2.36)
Look outside	3.73	(1.22 ,11.46)	2.92	(1.59 ,5.36)	2.90	(1.57 ,5.3)	12.81	(7.11 ,23.06)
Dance	1.85	(0.44 ,7.79)	0.90	(0.41 ,1.98)	0.90	(0.41 ,1.96)	1.49	(0.59 ,3.75)
Reach	3.01	(1.51 ,5.97)	3.56	(2.18 ,5.82)	3.55	(2.18 ,5.82)	16.61	(9.98 ,27.65)
Hygiene	1.00	(0.31 ,3.25)	0.77	(0.35 ,1.69)	0.77	(0.35 ,1.69)	1.00	(0.42 ,2.37)
Drink	1.94	(0.52 ,7.21)	1.40	(0.64 ,3.08)	1.41	(0.64 ,3.08)	1.79	(0.77 ,4.15)
Eat	0.44	(0.14 ,1.39)	0.81	(0.42 ,1.58)	0.81	(0.42 ,1.58)	1.80	(0.89 ,3.65)
Read or write	1.76	(0.27 ,11.64)	2.92	(0.92 ,9.28)	2.90	(0.91 ,9.19)	9.58	(1.88 ,48.76)
Passenger interaction	1.29	(0.83 ,2.04)	1.21	(0.88 ,1.66)	1.21	(0.88 ,1.66)	1.49	(1.18 ,1.89)
Children rear seat	0.44	(0.05 ,4.21)	1.35	(0.36 ,5.03)	1.35	(0.36 ,5.03)	1.20	(0.35 ,4.12)
Overall cellphone	2.08	(1.25 ,3.46)	1.38	(0.99 ,1.92)	1.38	(0.99 ,1.92)	3.74	(2.9 ,4.83)
Cellphone-talk	1.64	(0.67 ,4.07)	1.02	(0.59 ,1.77)	1.02	(0.59 ,1.77)	2.16	(1.3 ,3.6)
Cellphone-text	3.84	(1.74 ,8.37)	2.20	(1.43 ,3.39)	2.21	(1.43 ,3.39)	7.69	(4.8 ,12.31)
Cellphone-reach	1.94	(0.6 ,6.28)	2.72	(1.29 ,5.73)	2.71	(1.29 ,5.73)	7.69	(3.58 ,16.52)
Cellphone-dial	-	-	4.06	(1.15 ,14.25)	4.06	(1.15 ,14.25)	17.46	(4.26 ,71.62)
Cellphone-browse	0.75	(0.24 ,2.29)	0.74	(0.37 ,1.5)	0.74	(0.37 ,1.5)	5.53	(2.57 ,11.88)
Total device	1.24	(0.69 ,2.25)	0.94	(0.59 ,1.51)	0.94	(0.59 ,1.51)	2.97	(1.93 ,4.58)
Adjust device	2.36	(0.85 ,6.56)	1.95	(0.96 ,3.96)	1.95	(0.96 ,3.96)	5.42	(2.73 ,10.76)
Radio	0.93	(0.43 ,2.04)	0.63	(0.34 ,1.18)	0.63	(0.34 ,1.18)	2.16	(1.18 ,3.97)
Climate control	1.69	(0.28 ,10.34)	1.73	(0.44 ,6.85)	1.74	(0.44 ,6.85)	2.77	(1 ,7.69)

Appendix C

Glossary

AAE Adversarial auto-encoder. 59

ACIC Atlantic Causal Inference Competition. 67

ATE Average treatment effect. 8

ATO Average treatment effect for the overlap population. 29

ATT Average treatment effect for the treated. 8

BART Bayesian additive regression trees. 63

BNN Balancing neural network. 149

BNP Bayesian nonparametric. 63

BV-NICE Balancing variational neural inference of causal effect. 15, 47

BW Balancing weight. 150

CATE Conditional average treatment effect. 9

CC Case-crossover. 107

CE-VAE Causal estimation variational auto-encoder. 63

CFR Counterfactual regression model. 63

CLR Conditional Logistic regression. 109

CMSE Conditional mean square error. 137

CRM Counterfactual risk minimization. 64

DRO Distribution robust optimization. 78

DSF Double sample forest. 149

EB Entropy-balancing weight. 150

ERM Empirical risk minimization. 78

GAN Generative adversarial network. 46

HTE Heterogenous treatment effect. 9

IB Information bottleneck. 64

ICP Invariant causal prediction. 125

IHDP Infant health and development program. 67

IPW Inverse probability weight. 29

IRM Invariant risk minimization. 81, 83, 86

ITE Individualized treatment effect. 9

IV Instrument variable. 98

JSD Jensen–Shannon divergence. 57

KL Kullback-Leibler divergence. 55

LASSO Least absolute shrinkage and selection operator. 64

LCQP Linear constraint quadratic program. 137

LME Linear mixed effect model. 89

LOO Leave-one-out. 117

LR Logistic regression. 89

MCMC Markov chain Monte Carlo. 145

MLP Multi-layer perceptrons. 67

MLR Mixed-effect Logistic regression. 164

MNAR Missing-not-at-random. 127

MW Matching weight. 150

NDS Naturalistic driving study. 15, 17

NLL Negative log-likelihoood. 90

OOD Out of distribution. 76

OR Odds ratio. 18

OW Overlap weight. 29

PEHE Precision of estimating heterogeneous effects. 62

PF Propensity forest. 149

RCM Rubin causal model. 3

RCT Randomized controlled trial. 7, 76

REx Risk extrapolation. 125

RF Random forest. 68

RICE Risk invariant causal estimation. 15

RL Reinforcement learning. 139

RMSE Residual mean square error. 67

SEM Structural equation model. 3

SHRP2 Second Strategic Highway Research Program. 18

SMD Standardized mean difference. 33

SUTVA The stable unit treatment value assumption . 6

TARNet Treatment agnostic representation net. 149

VAE Variational auto-encoder. 46

VI Variational inference. 49

Bibliography

U. I. Abdullahi, S. Samothrakis, and M. Fasli. Counterfactual domain adversarial training of neural networks. In *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*, pages 151–155. IEEE, 2017.

A. Agresti. *Categorical data analysis*, volume 482. John Wiley & Sons, 2003.

A. Ahmed, A. Husain, T. E. Love, G. Gambassi, L. J. Dell'Italia, G. S. Francis, M. Gheorghiade, R. M. Allman, S. Meleth, and R. C. Bourge. Heart failure, chronic diuretic use, and increase in mortality and hospitalization: An observational study using propensity score methods. *European Heart Journal*, 27(12):1431–1439, 2006. ISSN 0195668X. doi: 10.1093/eurheartj/ehi890.

K. Ahuja, K. Shanmugam, K. Varshney, and A. Dhurandhar. Invariant risk minimization games. *arXiv preprint arXiv:2002.04692*, 2020.

A. Alaa and M. Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pages 129–138, 2018.

A. Alaa and M. Van Der Schaar. Validating causal inference models via influence functions. In *ICML*, pages 191–201, 2019.

A. M. Alaa and M. van der Schaar. Bayesian nonparametric causal inference: Information rates and learning algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 12 (5):1031–1046, 2018.

A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy. Fixing a broken ELBO. In *ICML*, pages 159–168, 2018.

M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. In *NIPS Workshop*. 2017.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

S. Assaad, S. Zeng, C. Tao, S. Datta, N. Mehta, R. Henao, F. Li, and L. Carin. Counterfactual representation learning with balancing weights. *arXiv preprint arXiv:2010.12618*, 2020.

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. W. Imbens. The state of applied econometrics: causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 2017. ISSN 0895-3309. doi: 10.1257/jep.31.2.3. URL http://arxiv.org/abs/1607.00699.

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, J. Tibshirani, S. Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

J. Atwood, F. Guo, G. M. Fitch, and T. A. Dingus. The driver-level crash risk associated with daily cellphone use and cellphone use while driving. *Accident Analysis and Prevention*, 119(June):149–154, 2018. ISSN 00014575. doi: 10.1016/j.aap.2018.07.007. URL https://doi.org/10.1016/j.aap.2018.07.007.

P. C. Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28:3083–3107, 2009. ISSN 02776715. doi: 10.1002/sim.

A. J. Averitt, N. Vanitchanant, R. Ranganath, and A. J. Perotte. The counterfactual chi-gan. *arXiv preprint arXiv:2001.03115*, 2020.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

A. Bellot and M. van der Schaar. Generalization and invariances in the presence of unobserved confounding. *arXiv preprint arXiv:2007.10653*, 2020.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

A. Bennett and N. Kallus. Policy evaluation with latent confounders via optimal balance. In *NeurIPS*, pages 4827–4837, 2019.

D. Bertsimas, M. Johnson, and N. Kallus. The power of optimization over randomization in designing experiments involving small samples. *Operations Research*, 63(4):868–876, 2015.

I. Bica, J. Jordon, and M. van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. In *NeurIPS*, 2020.

M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying mmd gans. In *ICLR*, 2018.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1): 3207–3260, 2013.

N. E. Breslow, N. E. Day, and E. Heseltine. *Statistical methods in cancer research*, volume 1. International Agency for Research on Cancer Lyon, 1980.

N. S. Breslow. The analysis of case-control studies. *Statistical methods in cancer research*, 1, 1980.

M. A. Brookhart, S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Stürmer. Variable selection for propensity score models. *American Journal of Epidemiology*, 163 (12):1149–1156, 2006. ISSN 0002-9262. doi: 10.1093/aje/kwj149.

L. Buesing, T. Weber, Y. Zwols, S. Racaniere, A. Guez, J.-B. Lespiau, and N. Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. *arXiv preprint arXiv:1811.06272*, 2018.

P. Bühlmann. Invariance, causality and robustness. *arXiv preprint arXiv:1812.08233*, 2018.

Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. In *ICLR*, 2016.

Z. Cai and M. Kuroki. On identifying total effects in the presence of latent variables and selection bias. In *UAI*, 2008.

J. Chen, Z. Xiu, B. Goldstein, R. Henao, L. Carin, and C. Tao. Supercharging imbalanced data learning with causal representation transfer. *arXiv preprint arXiv:2011.12454*, 2020.

L. Chen, C. Tao, R. Zhang, R. Henao, and L. C. Duke. Variational inference and model selection with generalized evidence bounds. In *ICML*, pages 893–902, 2018.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

H. A. Chipman, E. I. George, R. E. McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

W. G. Cochran. *Planning and analysis of observational studies*. Wiley, New York, 1983.

Y. Cong, M. Zhao, K. Bai, and L. Carin. Go gradient for expectation-based objectives. In *ICLR*, 2019.

H. Cramér. *Mathematical methods of statistics*, volume 43. Princeton university press, 1999. ISBN 0691005478.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009. ISSN 00063444. doi: 10.1093/biomet/asn055.

I. Csiszár. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten. *Publ. Math. Inst. Hungar. Acad*, 8:85–108, 1963.

E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 113–123, 2019.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

I. J. Dahabreh, S. J.-P. A. Haneuse, J. M. Robins, S. E. Robertson, A. L. Buchanan, E. A. Stuart, and M. A. Hernán. Study designs for extending causal inferences from a randomized trial to a target population. *arXiv:1905.07764*, 2019. URL http://arxiv.org/abs/1905.07764.

B. Dai, H. Dai, N. He, W. Liu, Z. Liu, J. Chen, L. Xiao, and L. Song. Coupled variational bayes via optimization embedding. In *Advances in Neural Information Processing Systems*, pages 9690–9700, 2018.

G. A. Davis. Crash reconstruction and crash modification factors. *Accident Analysis & Prevention*, 62:294–302, 2014.

T. A. Dingus, F. Guo, S. Lee, J. F. Antin, M. Perez, M. Buchanan-King, and J. Hankey. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences*, 113(10):2636–2641, 2016a.

T. A. Dingus, F. Guo, S. E. Lee, J. F. Antin, M. Perez, M. Buchanan-King, and J. M. Hankey. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences*, 113(10):2636–2641, 2016b. ISSN 0027-8424. doi: 10.1073/pnas.1513271113.

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

O. D. Duncan. *Introduction to structural equation models*. Elsevier, 2014. ISBN 148329532X.

G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *UAI*, 2015.

A. D'Amour. On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3478–3486, 2019.

L. Fang, C. Li, J. Gao, W. Dong, and C. Chen. Implicit deep latent variable models for text generation. *arXiv preprint arXiv:1908.11527*, 2019.

C. M. Farmer, S. G. Klauer, J. A. McClafferty, and F. Guo. Secondary Behavior of Drivers on Cell Phones. *Traffic Injury Prevention*, 16(8):801–808, 2015. ISSN 1538957X. doi: 10.1080/15389588.2015.1020422.

R. A. Fisher. The causes of human variability. *The Eugenics Review*, 10(4):213–220, 1919. ISSN 14643685. doi: 10.1093/ije/dyw315.

R. A. Fisher. Statistical methods for research workers. *Edinburgh Oliver Boyd*, 1925.

G. M. Fitch, S. A. Soccolich, F. Guo, J. McClafferty, Y. Fang, R. L. Olson, M. A. Perez, R. J. Hanowski, J. M. Hankey, and T. A. Dingus. The impact of hand-held and hands-free cell phone use on driving performance and safety-critical event risk final report. Technical Report April, 2013.

C. Flannagan, J. Bärgman, and A. Bálint. Replacement of distractions with other distractions: A propensity-based approach to estimating realistic crash odds ratios for driver engagement in secondary tasks. *Transportation Research Part F: Traffic Psychology and Behaviour*, 63:186–192, 2019. ISSN 13698478. doi: 10.1016/j.trf.2019.04.013. URL https://doi.org/10.1016/j.trf.2019.04.013.

A. Franks, A. D'Amour, and A. Feller. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, pages 1–33, 2019.

W. A. Fuller. *Measurement error models*, volume 305. John Wiley & Sons, 2009.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

R. Gao, X. Chen, and A. J. Kleywegt. Distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.

T. A. Glass, S. N. Goodman, M. A. Hernán, and J. M. Samet. Causal inference in public health. *Annual review of public health*, 34:61–75, 2013.

D. Glassbrenner and T. J. Ye. Driver cell phone use in 2006-overall results. *National Highway Traffic Safety Administration*, (July):2005–2006, 2007. URL https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/810790.

J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations research*, 58(4-part-1):902–917, 2010.

A. S. Goldberger. Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, pages 979–1001, 1972. ISSN 0012-9682.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014a.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

J. Greenberg, M. Blommer, K. Kozak, R. Curry, B. Artz, P. Grant, D. Kochhar, L. Tijerina, and L. Cathey. Driver distraction: evaluation with event detection paradigm. *Transportation Research Record: Journal of the Transportation Research Board*, 1843(1):1–9, 2007. ISSN 0361-1981. doi: 10.3141/1843-01.

A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Z. Griliches and J. A. Hausman. Errors in variables in panel data. *Journal of econometrics*, 31(1):93–118, 1986.

R. H. H. Groenwold, E. Hak, and A. W. Hoes. Quantitative assessment of unobserved confounding is mandatory in nonrandomized intervention studies. *Journal of Clinical Epidemiology*, 62(1):22–28, 2009. ISSN 08954356. doi: 10.1016/j.jclinepi.2008.02.011. URL http://dx.doi.org/10.1016/j.jclinepi.2008.02.011.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of Wasserstein GANs. *arXiv preprint arXiv:1704.00028*, 2017.

F. Guo. Statistical Methods for Naturalistic Driving Studies. *Annual Review of Statistics and Its Application*, 6:309–328, 2019a.

F. Guo. Statistical methods for naturalistic driving studies. *Annual review of statistics and its application*, 6:309–328, 2019b.

F. Guo and Y. Fang. Individual driver risk assessment using naturalistic driving data. *Accident Analysis & Prevention*, 61:3–9, 2013.

F. Guo, I. Kim, and S. G. Klauer. Semiparametric Bayesian models for evaluating time-variant driving risk factors using naturalistic driving data and case-crossover approach. *Statistics in Medicine*, (November 2017):1–15, 2017a. ISSN 10970258. doi: 10.1002/sim.7574.

F. Guo, S. G. Klauer, Y. Fang, J. M. Hankey, J. F. Antin, M. A. Perez, S. E. Lee, and T. A. Dingus. The effects of age on crash risk associated with driver distraction. *International Journal of Epidemiology*, 46(1):258–265, 2017b. ISSN 14643685. doi: 10.1093/ije/dyw234.

J. Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

J. M. Hankey, M. A. Perez, and J. A. McClafferty. Description of the SHRP2 naturalistic database and the crash, near-crash, and baseline data sets task report. Technical Report 9, Virginia Tech Transportation Institute, Blacksburg, VA, 2016.

C. Heinze-Deml, J. Peters, and N. Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

M. A. Hernán and J. M. Robins. *Causal Inference : What If*. Boca Raton: Chapman & Hall/CRC, 2020. URL https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2019/10/ci_hernanrobins_23oct19.pdf.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

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

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

G. E. Hinton and R. S. Zemel. Autoencoders, minimum description length and helmholtz free energy. In *NIPS*, pages 3–10, 1994.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

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

A. Hyvarinen, H. Sasaki, and R. E. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. *arXiv preprint arXiv:1805.08651*, 2018.

A. Hyvarinen, H. Sasaki, and R. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *AISTATS*, pages 859–868, 2019.

IIHS. Distracted Driving: Cellphone manipulations up 57 percent over prior survey. *Status Report*, 54(1):1–7, 2019. ISSN 00280496. doi: 10.4324/9781315516530-2.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015a. ISBN 0521885884.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015b.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017.

F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *ICML*, 2016.

F. D. Johansson, N. Kallus, U. Shalit, and D. Sontag. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*, 2018.

F. D. Johansson, D. Sontag, and R. Ranganath. Support and invertibility in domain-invariant representations. *arXiv preprint arXiv:1903.03448*, 2019.

F. D. Johansson, U. Shalit, N. Kallus, and D. Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

N. Kallus. A framework for optimal matching for causal inference. In *AISTATS*, 2017.

N. Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, pages 8895–8906, 2018a.

N. Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. *arXiv preprint arXiv:1802.05664*, 2018b.

N. Kallus. Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):85–112, 2018c.

N. Kallus and A. Zhou. Confounding-robust policy improvement. In *NeurIPS*, pages 9269–9279, 2018.

N. Kallus, X. Mao, and M. Udell. Causal inference with noisy and missing covariates via matrix factorization. In *Advances in neural information processing systems*, pages 6921–6932, 2018a.

N. Kallus, A. M. Puli, and U. Shalit. Removing hidden confounding by experimental grounding. In *NeurIPS*, pages 10888–10897, 2018b.

N. Kallus, X. Mao, and A. Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *AISTATS*, 2019.

V. Karwa, A. B. Slavković, and E. T. Donnell. Causal inference in transportation safety studies: Comparison of potential outcomes and causal diagrams. *Annals of Applied Statistics*, 5(2 B):1428–1455, 2011. ISSN 19326157. doi: 10.1214/10-AOAS440.

I. Khemakhem, D. P. Kingma, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. *arXiv preprint arXiv:1907.04809*, 2019.

I. Khemakhem, D. P. Kingma, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *AISTATS*, 2020.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

D. P. Kingma, T. Salimans, and M. Welling. Improving variational inference with inverse autoregressive flow. In *NIPS*, 2016.

S. G. Klauer, F. Guo, J. Sudweeks, and T. A. Dingus. An Analysis of Driver Inattention Using a Case-Crossover Approach On 100-Car Data: Final Report. Technical Report May, U.S. Department of Transportation, National Highway Traffic Safety Administration, Virginia Tech Transportation Institute, 2010.

S. G. Klauer, F. Guo, B. G. Simons-Morton, M. C. Ouimet, S. E. Lee, and T. A. Dingus. Distracted driving and risk of road crashes among novice and experienced drivers. *The New England Journal of Medicine*, 46(4):600–601, 2014a. ISSN 07364679. doi: 10.1016/j.jemermed.2014.02.017.

S. G. Klauer, F. Guo, B. G. Simons-Morton, M. C. Ouimet, S. E. Lee, and T. A. Dingus. Distracted driving and risk of road crashes among novice and experienced drivers. *New England journal of medicine*, 370(1):54–59, 2014b.

M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*, 2017.

D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, R. L. Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

M. Kuroki and J. Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

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

C. Lee, N. Mastronarde, and M. van der Schaar. Estimation of individual treatment effect in latent confounder models via adversarial learning. *arXiv preprint arXiv:1811.08943*, 2018.

S. Leung, R. J. Croft, M. L. Jackson, M. E. Howard, and R. J. Mckenzie. A comparison of the effect of mobile phone use and alcohol consumption on driving simulation performance. *Traffic Injury Prevention*, 13(6):566–574, 2012. ISSN 15389588. doi: 10.1080/15389588.2012.683118.

C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *NIPS*, 2017a.

F. Li, A. M. Zaslavsky, and M. B. Landrum. Propensity score weighting with multilevel data. *Statistics in Medicine*, 32(19):3373–3387, 2013. ISSN 02776715. doi: 10.1002/sim.5786.

F. Li, K. L. Morgan, and A. M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2017b. ISSN 1537274X. doi: 10.1080/01621459.2016.1260466.

F. Li, K. L. Morgan, and A. M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

F. Li, L. E. Thomas, and F. Li. Addressing extreme propensity scores via the overlap weights. *American journal of epidemiology*, 188(1):250–257, 2019. ISSN 14766256. doi: 10.1093/aje/kwy201.

S. Li and Y. Fu. Matching on balanced nonlinear representations for treatment effects estimation. In *Advances in Neural Information Processing Systems*, pages 929–939, 2017.

X. Li, X. Yan, J. Wu, E. Radwan, and Y. Zhang. A rear-end collision risk assessment model based on drivers' collision avoidance process under influences of cell phone use and gender—A driving simulator based study. *Accident Analysis and Prevention*, 97:1–18, 2016. ISSN 00014575. doi: 10.1016/j.aap.2016.08.021. URL http://dx.doi.org/10.1016/j.aap.2016.08.021.

Y. Li and R. E. Turner. Rényi divergence variational inference. In *NIPS*, 2016.

Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *ICML*, 2015.

R. Limpert. *Motor vehicle accident reconstruction and cause analysis*. LexisNexis, 2020.

B. London and T. Sandler. Bayesian counterfactual risk minimization. In *International Conference on Machine Learning*, pages 4125–4133. PMLR, 2019.

C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. In *NIPS*, pages 6446–6456, 2017.

C. Lu, B. Schölkopf, and J. M. Hernández-Lobato. Deconfounding reinforcement learning in observational settings. *arXiv preprint arXiv:1812.10576*, 2018.

D. Lu, F. Guo, and F. Li. Evaluating the causal effects of cellphone distraction on crash risk using propensity score methods. *Accident Analysis & Prevention*, 143:105579, 2020a.

D. Lu, C. Tao, J. Chen, F. Li, F. Guo, and L. Carin. Reconsidering Generative Objectives For Counterfactual Reasoning. *Nips*, (NeurIPS):1–15, 2020b.

J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment e ects: a comparative study. *Statistics in Medicine*, 2960 (19):2937–2960, 2004. doi: 10.1002/sim.1903.

W. Luo and Y. Zhu. Matching using sufficient dimension reduction for causal inference. *Journal of Business & Economic Statistics*, pages 1–13, 2019.

M. H. Maathuis, M. Kalisch, P. Bühlmann, et al. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

M. Maclure. The case-crossover design : a method for studying transient effects on the risk of acute events. *Amencan Journal of Epidemiolooy*, 133(2):144–153, 1991. doi: 10.1093/aje/kwx105.

M. Maclure, Mittleman, and M. A. Should We Use a Case-Crossover Design? *Annual Review of Public Health*, 21(1):193–221, 2000. ISSN 0163-7525. doi: 10.1146/annurev.publhealth.21.1.193. URL http://www.annualreviews.org/doi/10.1146/annurev.publhealth.21.1.193.

G. S. Maddala and K. Lahiri. *Introduction to econometrics*, volume 2. Macmillan New York, 1992.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

R. Månsson, M. M. Joffe, W. Sun, and S. Hennessy. On the estimation and use of propensity scores in case-control and case-cohort studies. *American Journal of Epidemiology*, 166(3): 332–339, 2007. ISSN 00029262. doi: 10.1093/aje/kwm069.

R. J. Marshall and R. T. Jackson. Analysis of case-crossover design. 12:2333–2341, 1993.

A. T. McCartt and L. L. Geary. Longer term effects of New York State's law on drivers' handheld cell phone use. *Injury Prevention*, 10(1):11–15, 2004. ISSN 13538047. doi: 10.1136/ip.2003.003731.

S. P. McEvoy, M. R. Stevenson, A. T. McCartt, M. Woodward, C. Haworth, P. Palamara, and R. Cercarelli. Role of mobile phones in motor vehicle crashes resulting in hospital attendance: A case-crossover study. *British Medical Journal*, 331(7514):428–430, 2005. ISSN 09598146. doi: 10.1136/bmj.38537.397512.55.

S. P. McEvoy, M. R. Stevenson, and M. Woodward. The contribution of passengers versus mobile phone use to motor vehicle crashes resulting in hospital attendance by the driver. *Accident Analysis and Prevention*, 39(6):1170–1176, 2007. ISSN 00014575. doi: 10.1016/j.aap.2007.03.004.

Q. Meng and J. Weng. Evaluation of rear-end crash risk at work zone using work zone traffic data. *Accident Analysis and Prevention*, 43(4):1291–1300, 2011. ISSN 00014575. doi: 10.1016/j.aap.2011.01.011. URL http://dx.doi.org/10.1016/j.aap.2011.01.011.

W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

O. Miettinen. Estimability and Estimation in Case-Referent Studies. *American Journal of Epidemiology*, 103(2):226–235, 1976.

T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *ICML Implicit Models Workshop*, 2017.

M. R. Montgomery, M. Gragnolati, K. A. Burke, and E. Paredes. Measuring living standards with proxy variables. *Demography*, 37(2):155–174, 2000.

K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

Y. Nagano, S. Yamaguchi, Y. Fujita, and M. Koyama. A wrapped normal distribution on hyperbolic space for gradient-based learning. In *ICML*, 2019.

D. M. Neyens and L. N. Boyle. The effect of distractions on the crash types of teenage drivers. *Accident Analysis and Prevention*, 39(1):206–212, 2007. ISSN 00014575. doi: 10.1016/j.aap.2006.07.004.

J. Neyman. *Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes.* PhD thesis, 1923.

NHTSA. Nationalmotor vehicle crash causation survey: report to Congress Rep. DOT HS 811 059. Technical report, US Dep. Transportation, Washington; DC, 2013.

NHTSA. Distracted driving in fatal crashes, 2017. Technical Report April 2019, 2019a. URL https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812700.

NHTSA. Police-Reported Motor Vehicle Traffic Crashes in 2018. Technical report, U.S. Department of Transportation, Washington, DC, 2019b.

M. Nickel and D. Kiela. Poincaré embeddings for learning hierarchical representations. In *NeurIPS*, pages 6338–6347, 2017.

X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *NIPS*, 2016.

J. M. Ogarrio, P. Spirtes, and J. Ramsey. A hybrid causal search algorithm for latent variable models. In *Conference on Probabilistic Graphical Models*, pages 368–379, 2016.

J. M. Owens, T. A. Dingus, F. Guo, Y. Fang, M. Perez, and J. McClafferty. Crash risk of cell phone use while driving: a case-crossover analysis of naturalistic driving Data. Technical report, AAA Foundation for Traffic Safety, Washington, DC 20005, 2018.

M. Ozery-Flato, P. Thodoroff, M. Ninio, M. Rosen-Zvi, and T. El-Hay. Adversarial balancing for causal inference. *arXiv preprint arXiv:1810.07406*, 2018.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

S. Parbhoo, M. Wieser, and V. Roth. Causal deep information bottleneck. *arXiv preprint arXiv:1807.02326*, 2018.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. ISSN 00063444. doi: 10.1093/biomet/82.4.669.

J. Pearl. *Causality: Models, reasoning and inference*. Cambridge University Press, Cambridge, MA, USA, 2000.

J. Pearl. *Causality*. Cambridge university press, 2009.

J. Pearl. On measurement bias in causal inference. *arXiv preprint arXiv:1203.3504*, 2012.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *JRSS-B*, 2016.

S. Powers, J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie, and R. Tibshirani. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine*, 37(11):1767–1787, 2018.

A. M. Puli and R. Ranganath. Generalized control functions via variational decoupling. *arXiv preprint arXiv:1907.03451*, 2019.

J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

T. Rainforth, T. A. Le, M. I. C. J. Maddison, and Y. W. T. F. Wood. Tighter variational bounds are not necessarily better. In *NIPS workshop*. 2017.

R. Ranganath and A. Perotte. Multiple causal inference with latent confounding. *arXiv preprint arXiv:1805.08273*, 2018.

D. A. Redelmeier and R. J. Tibshirani. Association between Cellular-Telephone Calls and Motor Vehicle Collisions. *New England Journal of Medicine*, 336(7):453–458, 1997a. ISSN 0028-4793. doi: 10.1056/NEJM199702133360701. URL http://www.nejm.org/doi/abs/10.1056/NEJM199702133360701.

D. A. Redelmeier and R. J. Tibshirani. Interpretation and bias in case-crossover studies. *Journal of Clinical Epidemiology*, 50(11):1281–1287, 1997b. ISSN 08954356. doi: 10.1016/S0895-4356(97)00196-0.

D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. In *ICML*, 2015.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89 (427):846–866, 1994.

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

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

P. R. Rosenbaum. From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *J. Am. Stat. Ass.*, 79(385):41–48, 1984.

P. R. Rosenbaum. Overt bias in observational studies. In *Observational studies*, pages 71–104. Springer, 2002.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika Biometrika*, 70(1):41–55, 1983a. ISSN 00063444. doi: 10.2307/2335942. URL http://www.jstor.org/stable/2335942%5Cnhttp://about.jstor.org/terms.

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)*, 45(2):212–218, 1983b.

N. Rosenfeld, Y. Mansour, and E. Yom-Tov. Predicting counterfactuals from large historical data and small randomized trials. In *WWW*, pages 602–609, 2017.

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

K. J. Rothman, S. Greenland, T. L. Lash, et al. *Modern epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008.

D. B. Rubin. Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974a. URL http://www.fsb.muohio.edu/lij14/420_paper_Rubin74.pdf.

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

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. ISSN 00063444. doi: 10.1093/biomet/63.3.581.

D. B. Rubin. Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, 2(1):1–26, 1977. ISSN 1098-6596.

D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 1(1):34–58, 1978.

D. B. Rubin. Randomization analysis of experimental data: The fisher randomization test. *Journal of the American Statistical Association*, 75(371):591–593, 1980. ISSN 1537274X. doi: 10.1080/01621459.1980.10477512.

D. B. Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990. ISSN 0883-4237. URL http://www.hss.caltech.edu/~jiji/Causation-Explanation/Rubin.pdf.

D. B. Rubin. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Matched Sampling for Causal Effects*, 2:169–188, 2001. doi: 10.1017/CBO9780511810725.030.

Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

M. Saito, B. K. Brimley, and G. G. Schultz. Calibration of the Highway Safety Manual and Development of New Safety Performance Functions. *Transportation Safety Data and Analysis*, 2, 2011.

Y. Saito and S. Yasui. Counterfactual cross-validation: Effective causal model selection from observational data. *arXiv preprint arXiv:1909.05299*, 2019.

L. Sasidharan and E. T. Donnell. Application of propensity scores and potential outcomes to estimate effectiveness of traffic safety countermeasures: Exploratory analysis using intersection lighting data. *Accident Analysis and Prevention*, 50:539–553, 2013. ISSN 00014575. doi: 10.1016/j.aap.2012.05.036. URL http://dx.doi.org/10.1016/j.aap.2012.05.036.

D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims. Recommendations as treatments: Debiasing learning and evaluation. *arXiv preprint arXiv:1602.05352*, 2016.

B. Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

A. Schuler, M. Baiocchi, R. Tibshirani, and N. Shah. A comparison of methods for model selection when estimating individual treatment effects. *arXiv preprint arXiv:1804.05146*, 2018.

P. Schwab, L. Linhardt, and W. Karlen. Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks. 2018a. URL http://arxiv.org/abs/1810.00656.

P. Schwab, L. Linhardt, and W. Karlen. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*, 2018b.

U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*, pages 3076–3085. JMLR. org, 2017.

X. Shi, W. Miao, J. C. Nelson, and E. J. T. Tchetgen. Multiply robust causal inference with double negative control adjustment for categorical unmeasured confounding. *arXiv preprint arXiv:1808.04906*, 2018.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

R. Singh, M. Sahani, and A. Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, pages 4595–4607, 2019.

A. Sinha, H. Namkoong, R. Volpi, and J. Duchi. Certifying some distributional robustness with principled adversarial training. In *ICLR*, 2018.

P. L. Spirtes, C. Meek, and T. S. Richardson. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*, 2013.

M. Staib and S. Jegelka. Distributionally robust optimization and generalization in kernel methods. In *Advances in Neural Information Processing Systems*, pages 9134–9144, 2019.

H. Steck. Training and testing of recommender systems on data missing not at random. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 713–722, 2010.

E. a. Stuart. Matching methods for causal inference. *Statistical Science*, 25(June):1–21, 2010. doi: 10.1214/09-STS313T4-AreviewandalookforwardM4-Citavi.

T. Stürmer, K. J. Rothman, J. Avorn, and R. J. Glynn. Treatment effects in the presence of unmeasured confounding: Dealing with observations in the tails of the propensity score distribution-A simulation study. *American Journal of Epidemiology*, 172(7):843–854, 2010. ISSN 14766256. doi: 10.1093/aje/kwq198.

M. Sugiyama, M. Krauledat, and K.-R. MÃžller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440, 2008.

C. Tao, L. Chen, R. Zhang, R. Henao, and L. C. Duke. Variational inference and model selection with generalized evidence bounds. In *ICML*, 2018.

C. Tao, L. Chen, S. Dai, J. Chen, K. Bai, D. Wang, J. Feng, W. Lu, G. Bobashev, and L. Carin. On fenchel mini-max learning. In *NeurIPS*, pages 10427–10439, 2019.

T. Teshima, I. Sato, and M. Sugiyama. Few-shot domain adaptation by causal mechanism transfer. 2020.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

J. M. Tomczak and M. Welling. Vae with a vampprior. In *AISTATS*, 2018.

D. Tran and D. M. Blei. Implicit causal models for genome-wide association studies. *arXiv preprint arXiv:1710.10742*, 2017.

T. J. Vander Weele and I. Shpitser. On the definition of a confounder. *Annals of Statistics*, 41(1):196–220, 2013. ISSN 00905364. doi: 10.1214/12-AOS1058.

T. J. VanderWeele and P. Ding. Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine*, 167(4):268–274, 2017.

V. Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992.

V. Veitch, Y. Wang, and D. M. Blei. Using embeddings to correct for unobserved confounding. In *NeurIPS*, 2019.

R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in neural information processing systems*, pages 5334–5344, 2018.

S. Wager and S. Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018a. ISSN 1537274X. doi: 10.1080/01621459.2017.1319839. URL https://doi.org/10.1080/01621459.2017.1319839.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018b.

D. Wang, Y. Yang, C. Tao, F. Kong, R. Henao, and L. Carin. Proactive pseudo-intervention: Causally informed contrastive learning for interpretable vision models. *arXiv preprint arXiv:2012.03369*, 2020.

H. Wang, Z. He, Z. C. Lipton, and E. P. Xing. Learning robust representations by projecting superficial statistics out. *arXiv preprint arXiv:1903.06256*, 2019.

Y. Wang and D. M. Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, pages 1–71, 2019.

Y. Wang, D. Liang, L. Charlin, and D. M. Blei. The deconfounded recommender: A causal inference approach to recommendation. *arXiv preprint arXiv:1808.06581*, 2018.

D. Westreich, S. R. Cole, M. J. Funk, M. A. Brookhart, and T. Sturmer. The role of the c-statistic in variable selection for propensity score models. *pharmacoepidemiology and drug safety*, (December 2010):317–320, 2011. ISSN 00221902. doi: 10.1002/pds. URL http://eprints.dinus.ac.id/15172/1/jurnal_14778.pdf.

J. Witte, L. Henckel, M. H. Maathuis, and V. Didelez. On efficient adjustment in causal graphs. *arXiv preprint arXiv:2002.06825*, 2020.

J. S. Wood and E. T. Donnell. Causal inference framework for generalizable safety effect estimates. *Accident Analysis and Prevention*, 104(February):74–87, 2017. ISSN 00014575. doi: 10.1016/j.aap.2017.05.001. URL http://dx.doi.org/10.1016/j.aap.2017.05.001.

J. S. Wood, J. P. Gooch, and E. T. Donnell. Estimating the safety effects of lane widths on urban streets in Nebraska using the propensity scores-potential outcomes framework. *Accident Analysis and Prevention*, 82:180–191, 2015. ISSN 00014575. doi: 10.1016/j.aap.2015.06.002. URL http://dx.doi.org/10.1016/j.aap.2015.06.002.

H. Wu and M. Wang. Variance regularized counterfactual risk minimization via variational divergence minimization. In *ICML*, 2018.

Z. Xiu, C. Tao, M. Gao, C. Davis, B. Goldstein, and R. Henao. Variational disentanglement for rare event modeling. In *AAAI*, 2021.

L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643, 2018.

M. Yin and M. Zhou. Arm: Augment-reinforce-merge gradient for discrete latent variable models. 2019.

J. Yoon, J. Jordon, and M. van der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *ICLR*, 2018.

S. Zeng, M. A. Bayir, J. Pfeiffer, D. Charles, and E. Kiciman. Causal transfer random forest: Combining logged data and randomized experiments for robust prediction. *arXiv preprint arXiv:2010.08710*, 2020.

C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

H. Zhang, D. D. Yao, and N. Ramakrishnan. Detection of stealthy malware activities with traffic causality and scalable triggering relation discovery. In *Proceedings of the 9th ACM symposium on Information, computer and communications security*, pages 39–50, 2014.

H. Zhang, D. D. Yao, N. Ramakrishnan, and Z. Zhang. Causality reasoning about network events for detecting stealthy malware activities. *computers & security*, 58:180–198, 2016.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655. AUAI Press, 2009.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827, 2013.

K. Zhang, M. Gong, and B. Schölkopf. Multi-source domain adaptation: A causal view. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

Y. Zhang, Z. Gan, K. Fan, Z. Chen, R. Henao, D. Shen, and L. Carin. Adversarial feature matching for text generation. In *ICML*, 2017.

Y. Zhang, A. Bellot, and M. van der Schaar. Learning overlapping representations for the estimation of individualized treatment effects. In *AISTATS*, 2020.

Q. Zhao, D. S. Small, and A. Ertefaie. Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*, 2017a.

Y. Zhao, X. Fang, and D. Simchi-Levi. A practically competitive and provably consistent algorithm for uplift modeling. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1171–1176. IEEE, 2017b.

J. R. Zubizarreta. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500): 1360–1371, 2012.

J. R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.