Partial Identification of Dose Responses with Hidden Confounders

Abstract

Inferring causal effects of continuous-valued treatments from observational data is a crucial task promising to better inform policy- and decision-makers. A critical assumption needed to identify these effects is that all confounding variables—causal parents of both the treatment and the outcome—are included as covariates. Unfortunately, given observational data alone, we cannot know with certainty that this criterion is satisfied. Sensitivity analyses provide principled ways to give bounds on causal estimates when confounding variables are hidden. While much attention is focused on sensitivity analyses for discrete-valued treatments, much less is paid to continuous-valued treatments. We present novel methodology to bound both average and conditional average continuous-valued treatment-effect estimates when they cannot be point identified due to hidden confounding. A semi-synthetic benchmark on multiple datasets shows our method giving tighter coverage of the true dose-response curve than a recently proposed continuous sensitivity model and baselines. Finally, we apply our method to a real-world observational case study to demonstrate the value of identifying dose-dependent causal effects.

1 INTRODUCTION

Causal inference on observational studies [Hill, 2011, Athey et al., 2019] attempts to predict conclusions of alternate versions of those studies, as if they were actually properly randomized experiments. The causal aspect is unique among inference tasks in that the goal is not prediction per se, as causal inference deals with *counterfactuals*, the problem of predicting unobservables: for example, what would have been a particular patient's health outcome had she taken some medication, versus not, while keeping all else equal (*ceteris paribus*)? There is quite often no way to validate the results without bringing in additional domain knowledge. A set of putative treatments $\mathcal{T}$, often binary with a treated/untreated dichotomy, induces *potential outcomes* $Y_{t\in\mathcal{T}}$. These can depend on covariates X as with heterogeneous treatment effects $\mathbb{E}[Y_1 - Y_0 \mid X]$ in the binary case. Only one outcome is ever observed: that at the assigned treatment T. Potential biases arise from the incomplete observation. This problem is exacerbated with more than two treatment values, especially when there are infinite possibilities, like in a continuum, e.g. $\mathcal{T} = [0, 1]$. Unfortunately, many consequential decisions in life involve this kind of treatment: What dose of drug should I take? How much of ___ should I eat/drink? How much exercise do I really need?

In an observational study, the direct causal link between assigned treatment T and observed outcome Y (also denoted as Y_T) can be influenced by indirect links modulated by *confounding* variables. For instance, wealth is often a confounder in an individual's health outcome from diet, medication, or exercise. Wealth affects access to each of these "treatments," and it also affects health through numerous other paths. Including the confounders as covariates in X allows estimators to condition on them and disentangle the influences [Yao et al., 2021].

It can be challenging to collect sufficient data, in terms of quality and quantity, on confounders in order to adjust a causal estimation to them. Case in point, noisy observations of e.g. lifestyle confounders lead researchers to vacillate on the health implications of coffee [Atroszko, 2019], alcohol [Ystrom et al., 2022], and cheese [Godos et al., 2020].

For consequential real-world causal inference, it is only prudent to allow margins for some amount of hidden confounding. A major impediment to such analysis is that it is impossible to know how a hidden confounder would bias the causal effect. The role of any causal *sensitivity model* [Cornfield et al., 1959, Rosenbaum and Rubin, 1983] is to make reasonable structural assumptions [Manski, 2003]

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

about different levels of hidden confounding. Most sensitivity analyses to hidden confounding require the treatment categories to be binary or at least discrete. This weakens empirical studies that are better specified by dose-response curves [Calabrese and Baldwin, 2001, Bonvini and Kennedy, 2022] from a continuous treatment variable. Estimated dose-response functions are indeed vulnerable in the presence of hidden confounders. Figure 1 highlights the danger of skewed observational studies that lead to biased estimates of personal toxic thresholds of treatment dosages.

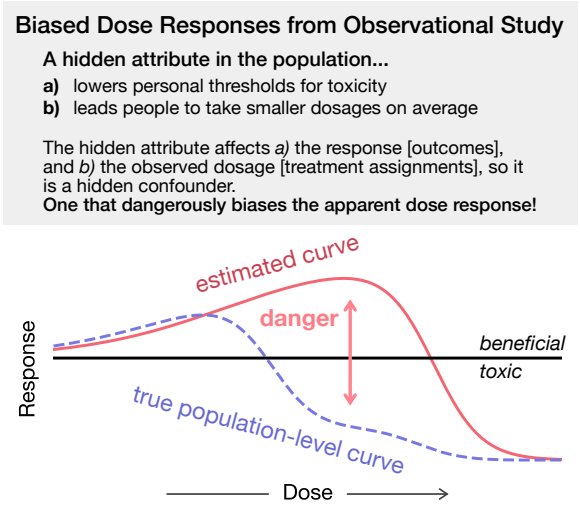

Figure 1. Dose-respone curves in medicine [e.g. Taleb, 2018] can be viewed as expected potential outcomes from continuous treatments. In this simulation (with details in §D,) there is one unobserved confounder. The empirical estimate of the population-level dose responses massively overshoots the maximum effective dosage, and would suggest treatments that were actually toxic to the population. This phenomenon persists even when the vulnerable hidden subgroup occurs more often in the population.

1.1 RELATED WORKS

There is growing interest in causal methodology for continuous treatments (or exposures, interventions), especially in the fields of econometrics [e.g. Huang et al., 2021, Tübbicke, 2022], health sciences [Vegetabile et al., 2021], and machine learning [Chernozhukov et al., 2021, Ghassami et al., 2021, Colangelo and Lee, 2021, Kallus and Santacatterina, 2019]. So far, most scrutiny on partial identification of potential outcomes has focused on the case of discrete treatments [e.g. Rosenbaum and Rubin, 1983, Louizos et al., 2017, Lim et al., 2021]. A number of creative approaches recently made strides in the discrete setting. Most rely on a *sensitivity model* for assessing the susceptibility of causal estimands to hidden-confounding bias. A sensitivity model allows hidden confounders but restricts their possible influence on the data, with an adjustable parameter that controls the overall tightness of that restriction.

The common discrete-treatment sensitivity models are incompatible with continuous treatments, which are needed for estimating dose-response curves. Still, some recent attempts have been made to handle hidden confounding under more general treatment domains [Chernozhukov et al., 2021]. Padh et al. [2022], Hu et al. [2021] optimize generative models to reflect bounds on the treatment effect due to ignorance, inducing an implicit sensitivity model through functional constraints. Instrumental variables are also helpful when they are available [Kilbertus et al., 2020]. The CMSM [Jesson et al., 2022] was developed in parallel to this work, and now serves as a baseline.

Regarding binary treatments, the Marginal Sensitivity Model (MSM) due to Tan [2006] has found widespread usage [Zhao et al., 2019, Veitch and Zaveri, 2020, Yin et al., 2021, Kallus et al., 2019, Jesson et al., 2021]. Variations thereof include Rosenbaum's earlier sensitivity model [2002] that enjoys ties to regression coefficients [Yadlowsky et al., 2020]. Sensitivity models can be avoided through generative [Meresht et al., 2022] and optimization [Guo et al., 2022] perspectives. Other approaches require additional structure to the data-generating (*observed outcome, treatment, covariates*) process. Proximal causal learning [Tchetgen et al., 2020, Mastouri et al., 2021] requires additional proxy structures. Chen et al. [2022] rely on multiple large dataset partitions.

1.2 CONTRIBUTIONS

We propose a novel sensitivity model for continuous treatments in §2. Next, we derive general formulas (§2.1) and solve closed forms for three versions (§2.3) of partially identified dose responses. We devise an efficient sampling algorithm (§3), and validate our results empirically using a semi-synthetic benchmark (§4) and realistic case study (§5).

1.3 PROBLEM STATEMENT

Our goal is the partial identification of causal dose responses under a bounded level of possible hidden confounding. We consider any setup that grants access to two predictors [Chernozhukov et al., 2017] that can be learned empirically and are assumed to output correct conditional distributions. These are (1) a predictor of outcomes conditioned on covariates and the assigned treatment, and (2) a predictor of the propensity of treatment assignments, taking the form of a probability density, conditioned on the covariates. The latter measures (non-)uniformity in treatment assignment for different parts of the population.

Potential outcomes. Causal inference is often cast in the nomenclature of potential outcomes, due to Rubin [1974]. Our first assumption, common to Rubin's framework, is that observation tuples of outcome, assigned treatment, and covariates, $\{(y^{(i)}, t^{(i)}, x^{(i)})\}_{i=1}^{n}$, are *i.i.d* draws from a single

joint distribution. This subsumes the Stable Unit Treatment Value Assumption (SUTVA), where units/individuals cannot depend on one another, since they are *i.i.d.*

The second assumption is overlap/positivity, that all treatments have a chance of assignment for every individual in the data: $p_{T|X}(t \mid x) > 0$ for every $(t, x) \in \mathcal{T} \times \mathcal{X}$.

The third and most challenging fundamental assumption is that of *ignorability*/sufficiency: $\{(Y_t)_{t\in\mathcal{T}} \perp\!\!\!\perp T\} \mid X$. Clearly the outcome should depend on the assigned treatment, but *potential outcomes* ought not to be affected by the assignment, after blocking out paths through covariates.

Our study focuses on dealing with limited violations to ignorability. The situation is expressed formally as $\{(Y_t)_{t\in\mathcal{T}} \not\perp\!\!\!\perp T\} \mid X$, but more specifically, we shall introduce a sensitivity model that governs the shape and extent of that violation.

Let $p(y_t|x)$ denote the probability density function of *potential* outcome $Y_t = y_t$ from a treatment $t \in \mathcal{T}$, given covariates $X = x$. This is what we seek to infer, while observing realized outcomes that allow us to learn the density $p(y_t|\, x,\, T = t)$. If the ignorability condition held, then $p(y_t|\, x,\, T = t) = p(y_t|x)$ due to the conditional independence. However, without ignorability, one has to marginalize over treatment assignment, requiring $p(y_t|\, x,\, T \neq t)$ because

$$p(y_t|x) = \int_{\mathcal{T}} p(y_t|\, \tau, x) p(\tau|x)\, \mathrm{d}\tau, \tag{1}$$

where $p(y_t|\tau, x)$ is the distribution of potential outcomes conditioned on actual treatment $T = \tau \in \mathcal{T}$ that may differ from the potential outcome's index t. The density $p(\tau|x)$ is termed the nominal propensity, defining the distribution of treatment assignments for different covariate values.

On notation. Throughout this study, y_t will indicate the value of the potential outcome at treatment t, and to disambiguate with *assigned* treatment τ will be used for events where $T = \tau$. For instance, we may care about the counterfactual of a smoker's $(\tau = 1)$ health outcome had they not smoked $(y_{t=0})$, where $T = 0$ signifies no smoking and $T = 1$ is "full" smoking. We will use the shorthand $p(\cdots)$ with lowercase variables whenever working with probability densities of the corresponding variables in uppercase:

$$p(y_t|\tau, x) \;\text{ means }\; \frac{\partial}{\partial u}\mathbb{P}[\, Y_t \leq u \mid T = \tau,\; X = x\,]\Big|_{u=y_t}.$$

Quantities of interest. We attempt to impart intuition on the conditional probability densities that may be confusing.

- $p(y_t|\, x)$ [conditional potential outcome]. A person's outcome from a treatment, disentangled from the selection bias of treatment assignment in the population. We seek to characterize this in order to (partially) identify the Conditional Average Potential Outcome (CAPO) and the Average Potential Outcome (APO).
- $p(y_t|\, \tau, x)$ [counterfactual]. What is the potential outcome of a person in the population characterized by x and assigned treatment τ? The answer changes with τ only when x is inadequate to block all backdoor paths through confounders. We can estimate this for $t = \tau$.
- $p(\tau|\, y_t, x)$ [complete propensity] is related to the above by Bayes' rule. We distinguish it from the nominal propensity $p(\tau|x)$ because the unobservable y_t possibly confers more information about the individual, again if x is inadequate. The complete propensity cannot be estimated, *even for* $t = \tau$; hence, this is the target of our sensitivity model.

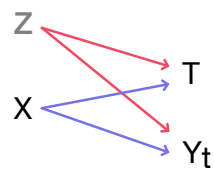

Figure 2. In this example, Z encompasses all hidden confounders. Counterfactual $p(y_t|\, \tau, x)$ diverges from $p(y_t|\, x)$ because of the red path from T to Y_t through Z.

Sensitivity. Explored by Tan [2006] followed by Kallus et al. [2019], Jesson et al. [2021], among many others, the Marginal Sensitivity Model (MSM) serves to bound the extent of (putative) hidden confounding in the regime of binary treatments $T' \in \{0, 1\}$. The MSM limits the discrepancy between the odds of treatment under the nominal propensity and the odds of treatment under the complete propensity.

Definition 1 (The Marginal Sensitivity Model)**.** For binary treatment $t' \in \{0, 1\}$ and violation factor $\Gamma \geq 1$, the following ratio is bounded:

$$\Gamma^{-1} \leq \left[\frac{p(t'|x)}{1 - p(t'|x)}\right]^{-1} \left[\frac{p(t'|\, y_{t'}, x)}{1 - p(t'|\, y_{t'}, x)}\right] \leq \Gamma.$$

The confines of a binary treatment afford a number of conveniences. For instance, one probability value is sufficient to describe the whole propensity landscape on a set of conditions, $p(1 - t'|\cdots) = 1 - p(t'|\cdots)$. As we transfer to the separate context of treatment continua, we must contend with infinite treatments and infinite potential outcomes.

2 CONTINUOUS SENSITIVITY MODEL

The counterfactuals required for Equation 1 are almost entirely unobservable. We look to the Radon-Nikodym derivative ω_δ of a counterfactual with respect to another [Tan, 2006], quantifying their divergence between nearby treatment assignments: (assuming mutual continuity)

$$\omega_\delta(y_t|\, \tau, x) := \frac{p(y_t|\, \tau + \delta, x)}{p(y_t|\, \tau, x)} \overset{\text{(Bayes' rule)}}{=} \frac{p(\tau + \delta|\, y_t, x) p(\tau|x)}{p(\tau|\, y_t, x) p(\tau + \delta|x)}$$
$$= \left[\frac{p(\tau + \delta|x)}{p(\tau|x)}\right]^{-1} \left[\frac{p(\tau + \delta|\, y_t, x)}{p(\tau|\, y_t, x)}\right].$$

As with the MSM, we encounter a ratio of odds, here contrasting τ versus $\tau + \delta$ in the assigned-treatment continuum. Assuming the densities are at least once differentiable,

$$\lim_{\delta \to 0} \delta^{-1} \log \omega_\delta(y_t | \tau, x) = \partial_\tau [\log p(\tau | y_t, x) - \log p(\tau | x)].$$

By bounding ω_δ close to one, we tie the logarithmic derivatives of the nominal and complete propensity densities.

Definition 2 (The Infinitesimal Marginal Sensitivity Model)**.** For treatments $t \in \mathcal{T} \subseteq \mathbb{R}$, where $\mathcal{T}$ is connected, and violation-of-ignorability factor $\Gamma \geq 1$, the δMSM requires

$$\left| \frac{\partial}{\partial \tau} \log \frac{p(\tau | y_t, x)}{p(\tau | x)} \right| \leq \log \Gamma$$

everywhere, for all τ, t, and x combinations. This differs from the CMSM due to Jesson et al. [2022] that considers only $t = \tau$, and which bounds the density ratios directly.

2.1 THE COMPLETE FRAMEWORK

Assumption 1 (Bounded Hidden Confounding)**.** Invoking Definition 2, the violation of ignorability is constrained by a δMSM with some $\Gamma \geq 1$.

Assumption 2 (Anchor Point)**.** A special treatment value designated as zero is not informed by potential outcomes: $p(\tau = 0 \mid y_t, x) = p(\tau = 0 \mid x)$ for all x, t, and y_t.

Assumption 2 states that we look for sensitivity to hidden confounders outside a control group at $T = 0$. The restriction is modeled after situations where we expect a dramatically lessened vulnerability to hidden confounders for a relatively more plentiful control group. It will soon become clear why Assumption 2 is necessary. We discuss ramifications and a workaround in §2.3.

The unobservability of almost all counterfactuals is unique to the case of continuous treatments, since the discrete analogy would be a discrete sum with an observable term. Figure 3 explains our approach to solving Equation 1.

2.2 A PARTIAL APPROXIMATION

We expand $p(y_t|\tau, x)$ around $\tau = t$, where $p(y_t|t, x) = p(y|t, x)$ is learnable from data. Suppose that $p(y_t|\tau, x)$ is twice differentiable in τ. Construct a Taylor expansion

$$\begin{aligned} p(y_t|\tau, x) &= p(y_t|t, x) + (\tau - t)\partial_\tau p(y_t|\tau, x)|_{\tau=t} \\ &\quad + \frac{(\tau - t)^2}{2} \partial_\tau^2 p(y_t|\tau, x)|_{\tau=t} + \mathcal{O}(\tau - t)^3. \end{aligned} \tag{2}$$

Denote with $\tilde{p}(y_t|\tau, x)$ an approximation of second order as laid out above. One could have stopped at lower orders but the difference in complexity is not that large. The intractable derivatives like $\partial_\tau p(y_t|\tau, x)|_{\tau=t}$ will be bounded using the

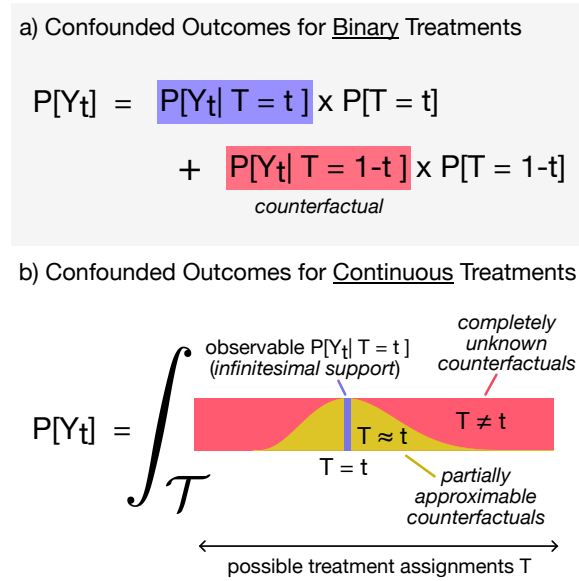

Figure 3. In the binary case, the red part is unobservable, but the MSM condition helps to bound that quantity. In the continuous case the integrand (Equation 1) is unobservable *almost everywhere* in the space of assigned treatments, except for the infinitesimal point $T = t$. In order to divide the integral into two parts (observable and unobservable) like with the binary sum, we must draw an approximation where assigned treatment and potential-outcome index are close enough. We use a soft window (yellow) to mark the validity of the approximation. Our continuous version of the MSM, the δMSM, allows us to bound the red part as well as reason about the yellow part. Covariates X are omitted for brevity.

δMSM machinery. Let us quantify the reliability of this approximation by a trust-weighing scheme $0 \leq w_t(\tau) \leq 1$, where typically $w_t(t) = 1$. This corresponds to the yellow part in Figure 3. We argue that $w_t(\tau)$ should be narrower with lower-entropy (narrower) propensities (§B). The possible forms of $w_t(\tau)$ are elaborated in §2.3.

Splitting Equation 1 along the trusted regime marked by $w_t(\tau)$, and then applying the approximation of Equation 2,

$$\begin{aligned} p(y_t|x) &= \int_{\mathcal{T}} \underbrace{w_t(\tau) p(y_t|\tau, x) p(\tau|x) \, \mathrm{d}\tau}_{\text{“observable” (Fig. 3)}} \\ &\quad + \int_{\mathcal{T}} \underbrace{[1 - w_t(\tau)] p(y_t|\tau, x) p(\tau|x) \, \mathrm{d}\tau}_{\text{“unobservable” (Fig. 3)}} \\ &\approx \int_{\mathcal{T}} \underbrace{w_t(\tau) \tilde{p}(y_t|\tau, x) p(\tau|x) \, \mathrm{d}\tau}_{(A) \text{ the approximated quantity}} \\ &\quad + \int_{\mathcal{T}} \underbrace{[1 - w_t(\tau)] p(\tau|y_t, x) p(y_t|x) \, \mathrm{d}\tau}_{(B) \text{ by Bayes' rule}}. \end{aligned} \tag{3}$$

The intuition behind separating the integral into two parts is

the following. By choosing the weights $w_t(\tau)$ so that they are close to one in the range where approximation Equation 2 is valid (yellow region in Figure 3) and zero outside of this range, we can evaluate the first integral through the approximated counterfactuals. The second integral, which is effectively over the red region in Figure 3 and cannot be evaluated due to unobserved counterfactuals, will be bounded using the δMSM. Simplifying the second integral first,

$$\int_{\mathcal{T}} [1 - w_t(\tau)] p(\tau|\, y_t, x) p(y_t|x)\, \mathrm{d}\tau = p(y_t|x) \left[1 - \int_{\mathcal{T}} w_t(\tau) p(\tau|\, y_t, x)\, \mathrm{d}\tau \right].$$

We witness already that $p(y_t|x)$ shall take the form of

$$p(y_t|x) \approx \frac{\int_{\mathcal{T}} w_t(\tau) \tilde{p}(y_t|\, \tau, x) p(\tau|x)\, \mathrm{d}\tau}{\int_{\mathcal{T}} w_t(\tau) p(\tau|\, y_t, x)\, \mathrm{d}\tau}. \tag{4}$$

Reflecting on Assumptions 1 & 2, the divergence between $p(\tau|\, y_t, x)$ and $p(\tau|x)$ is bounded, allowing characterization of the denominator in terms of the learnable $p(\tau|x)$. Similarly, the derivatives in Equation 2 are bounded for characterizing the numerator. Without loss of generality, consider the unknown quantity γ such that

$$\partial_\tau \log p(\tau|y_t, x) = \partial_\tau \log p(\tau|x) + \gamma(\tau|y_t, x), \quad \text{where } |\gamma(\tau|y_t, x)| \leq \log \Gamma \text{ using the } \delta\text{MSM}. \tag{5}$$

We may attempt to integrate both sides;

$$\int_0^s \partial_\tau \log p(\tau|\, y_t, x)\, \mathrm{d}\tau = \int_0^s \partial_\tau \log p(\tau|x)\, \mathrm{d}\tau + \underbrace{\int_0^s \gamma(\tau|y_t, x)\, \mathrm{d}\tau}_{:= \lambda(s|y_t, x)}.$$

$$\begin{aligned} \therefore \log p(\tau = s|\, y_t, x) - \log p(\tau = 0|\, y_t, x) &= \log p(\tau = s|\, x) - \log p(\tau = 0|\, x) + \lambda(s|y_t, x), \\ \therefore \log p(\tau|\, y_t, x) &= \log p(\tau|x) + \lambda(\tau|y_t, x). \end{aligned}$$

(by Assumption 2)

$$\therefore \quad p(\tau|\, y_t, x) = p(\tau|x)\Lambda(\tau|y_t, x), \quad \Lambda := \exp\{\lambda\}. \tag{6}$$

One finds that $|\lambda(\tau|y_t, x)| \leq |\tau| \log \Gamma$ because λ integrates γ, bounded by $\pm \log \Gamma$, over a support with length τ. Subsequently, Λ is bounded by $\Gamma^{\pm|\tau|}$. These are the requisite tools for bounding $p(y_t|x)$—or an approximation thereof, erring on ignorance via the trusted regime marked by $w_t(\tau)$. The derivation is completed in §A by framing the unknown quantities in terms of γ and Λ, culminating in Equation 7.

Predicting potential outcomes. The recovery of a fully normalized probability density $\tilde{p}(y_t|x)$ via Equation 4 is laid out below. It may be approximated with Monte Carlo or solved in closed form with specific formulations for the weights and propensity. Concretely, it takes on the form $\tilde{p}(y_t|x) = d(t|y_t, x)^{-1} p(y_t|t, x)$, where

$$\begin{aligned} d(t|y_t, x) := \mathbb{E}_\tau[\Lambda(\tau|y_t, x)] - [\gamma\Lambda](t|y_t, x)\ \mathbb{E}_\tau[\tau - t] \\ - \frac{1}{2}[(\dot{\gamma} + \gamma^2)\Lambda](t|y_t, x)\ \mathbb{E}_\tau[(\tau - t)^2], \end{aligned} \tag{7}$$

and said expectations, $\mathbb{E}_\tau[\cdot]$, are with respect to the implicit distribution $q(\tau|t, x) \propto w_t(\tau) p(\tau|x)$. The notation $\dot{\gamma}$ denotes a derivative in the first argument of $\gamma(t|y_t, x)$.

Assumption 3 (Second-order Simplification)**.** The quantity $\dot{\gamma}(\tau|y_t, x)$ cannot be characterized as-is. Granting that γ^2 dominates over the former, and consequently

$$\left|(\dot{\gamma} + \gamma^2)\Lambda\right| \leq \left|\gamma^2 \Lambda\right| + \varepsilon \quad \text{for small } \varepsilon \geq 0.$$

To make use of the formula in Equation 7, one first obtains the set of admissible $d(t|y_t, x) \in \left[\underline{d}(t|y_t, x), \overline{d}(t|y_t, x)\right]$ that violate ignorability up to a factor Γ according to the δMSM. With the negative side of the $\pm$ corresponding to $\underline{d}$ and the positive side to $\overline{d}$, the bounds are expressible as

$$\begin{aligned} \left(\underline{d}, \overline{d}\right) = \int_{\mathcal{T}} \Gamma^{\pm|\tau|} q(\tau|t, x)\, \mathrm{d}\tau \quad \longrightarrow \mathbb{E}_\tau[\Lambda(\tau|y_t, x)] \\ + (\pm \log \Gamma)\Gamma^{|t|} \int_{\mathcal{T}} (\tau - t) q(\tau|t, x)\, \mathrm{d}\tau \\ + \frac{1}{2}\left(0,\ \log^2 \Gamma\right)\Gamma^{|t|} \int_{\mathcal{T}} (\tau - t)^2 q(\tau|t, x)\, \mathrm{d}\tau. \end{aligned} \tag{8}$$

It is evident from the first integral that $\underline{d} \leq 1 \leq \overline{d}$ with equality at $\Gamma = 1$. This is noteworthy because it implies that $p(y|t, x)$ is admissible for the partially identified $\tilde{p}(y_t|x)$. We cannot describe $p(y_t|x)$ in extreme cases where $\underline{d} \leq 0$.

Ensembles. To quantify empirical uncertainties [Jesson et al., 2020] alongside our sensitivity, the predictors could be learned as ensembles, with $\tilde{p}(y_t|x)$ computed as (bootstrap resampled [Lo, 1987]) expectations over them.

2.3 PROPENSITY-TRUST COMBINATIONS

In addition to developing the general framework above, we derive analytical forms for a myriad of paramametrizations that span the relevant supports $\mathcal{T}$ for continuous treatments: the unit interval $[0, 1]$, the nonnegative reals $[0, +\infty)$, and the real number line $(-\infty, +\infty)$. For some nominal propensity distributions $p(\tau|x)$, we propose trust-weighing schemes $w_t(\tau)$ with shared form so that the expectations in Equation 8 are solvable.

For instance, consider the parametrization $(T \mid X = x) \sim \text{Beta}(\alpha(x), \beta(x))$. For the weighing scheme, we select the rescaled and recentered $w_t^{\text{beta}}(\tau) = c_t \tau^{a_t - 1}(1 - \tau)^{b_t - 1}$. Two constraints are imposed on every $w_t(\tau)$ studied herein:

- *(the mode)* that $w_t(\tau)$ peaks at $\tau = t$, and $w_t(t) = 1$.
- *(the precision)* that some $r > 0$ defines a narrowness of the form, and can be set a priori.

Parametrization	Support ($\mathcal{T}$)	Params.	Precision (r)	Bounds for $\mathbb{E}_\tau[\Lambda(\tau\|y_t, x)]$
Beta	$[0,1]$	α, β	$\alpha+\beta-2$	${}_1F_1(\boldsymbol{\alpha}+1;\ \boldsymbol{\alpha}+\boldsymbol{\beta}+2;\ \pm\log\Gamma)$ where $\boldsymbol{\alpha} \coloneqq \bar{\alpha}+\alpha-2,\ \boldsymbol{\beta} \coloneqq \bar{\beta}+\beta-2$
Balanced Beta	$[0,1]$	α, β	$\alpha+\beta-2$	$t \cdot \langle$the Beta above$\rangle + (1-t) \cdot \langle$Beta, mirrored$\rangle$
Gamma	$[0,+\infty)$	α, β	α/β^2	$[1-(\pm\log\Gamma)/\boldsymbol{\beta}]^{-\boldsymbol{\alpha}}$ where $\boldsymbol{\alpha} \coloneqq \bar{\alpha}+\alpha-1,\ \boldsymbol{\beta} \coloneqq \bar{\beta}+\beta$
Gaussian	$(-\infty,+\infty)$	μ, σ	$1/\sigma$	$\exp\{\boldsymbol{\sigma}^2(\log\Gamma)^2/2\}\begin{pmatrix}\Gamma^{\pm\boldsymbol{\mu}}[1+\mathrm{erf}(\frac{\boldsymbol{\mu}\pm\boldsymbol{\sigma}^2\log\Gamma}{\sqrt{2}\boldsymbol{\sigma}})] \\ + \\ \Gamma^{\mp\boldsymbol{\mu}}[1-\mathrm{erf}(\frac{\boldsymbol{\mu}\mp\boldsymbol{\sigma}^2\log\Gamma}{\sqrt{2}\boldsymbol{\sigma}})]\end{pmatrix}$ where $\boldsymbol{\mu} \coloneqq \frac{\mu\bar{\sigma}^2+\bar{\mu}\sigma^2}{\bar{\sigma}^2+\sigma^2},\ \boldsymbol{\sigma}^2 \coloneqq \frac{\bar{\sigma}^2\sigma^2}{\bar{\sigma}^2+\sigma^2}$

Table 1. Candidates for propensity and trust-weighing combinations. Each row specifies the distribution—beta, beta, gamma, and Gaussian respectively—of the propensity model $p(\tau|x)$. The last column lists solutions for the first term of Equation 7 / 8. This is a convolution of the propensity and weighing scheme, which have similar forms (see Bromiley [2003] for the Gaussian case.) We distinguish the replicated parameters between propensity and weight by placing a bar over the propensity parameters. So if the propensity is $x \mapsto (\bar{\alpha}, \bar{\beta})$, then the weighing scheme has $t \mapsto (\alpha, \beta)$. The bold parameters are of the compound density, with respect to which the first and second moments are computed in Equation 7 / 8.

For the beta version we chose $a_t + b_t = r + 2$. These constraints imply that $a_t \coloneqq rt + 1$, $b_t \coloneqq r(1-t) + 1$, and $c_t^{-1} \coloneqq t^{rt}(1-t)^{r(1-t)}$.

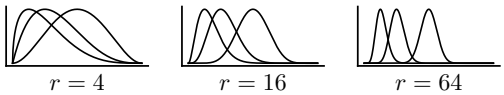

Figure 4. Beta parametrizations for $w_t(\tau)$ in the unit square, plotted for $t = 0.125, 0.25, 0.5$. Trust declines with r.

The choices. We present solutions for propensity-trust combinations in Table 1. Balanced Beta stands out by not strictly obeying Assumption 2. Rather, it adheres to a symmetrified mixture that is more versatile to realistic situations.

Balanced Beta. Formally, for all t, y_t, and x, we balance the Beta parametrization by replacing Assumption 2 with

$$\begin{cases} p(\tau = 0 \mid y_t, x) = p(\tau = 0 \mid x) & \text{w.p.} \quad t, \\ p(\tau = 1 \mid y_t, x) = p(\tau = 1 \mid x) & \text{w.p.} \quad 1-t. \end{cases}$$

This special parametrization deserves further justifying. The premise is that distant treatments are decoupled; treatment assignment τ shares less information with a distal potential outcome y_t than a proximal one. If that were the case, then the above linear interpolation favors the less informative anchor points for a given t. This is helpful because the sensitivity analysis is vulnerable to the anchor points. Stratifying the anchor points eventually leads to an arithmetic mixture of $d(t|y_t, x)$ in Equation 7 with its mirrored version about $t \mapsto 1-t$, and $(\alpha, \beta) \mapsto (\beta, \alpha)$.

Controlling trust levels. The absolute error of the approximation in Equation 3.A is bounded above by a form that could grow with narrower propensities (see §B), in the Beta parametrization. Intuitively, the error also depends on the smoothness of the complete propensity (following the Taylor expansion.) For that reason we decided to set the trust-weighing precision r to that of the nominal propensity.

3 ESTIMATING THE INTERVALS

We seek to bound partially identified expectations with respect to the true potential-outcome densities, which are constrained according to Equation 7 / 8. The quantities of interest are the Average Potential Outcome (APO), $\mathbb{E}[f(Y_t)]$, and Conditional Average Potential Outcome (CAPO), $\mathbb{E}[f(Y_t)|X = x]$, for any task-specific $f(y)$. We use Monte Carlo over n realizations y_i drawn from proposal density $q(y)$, and m covariate instances:

$$\tilde{\mathbb{E}}[f(Y_t) \mid X \in \{x\}_m] = \frac{\sum_{i=1}^n \sum_{j=1}^m f(y_i)\tilde{p}(y_t = y_i|x_j)/q(y_i)}{\sum_{i=1}^n \sum_{j=1}^m \tilde{p}(y_t = y_i|x_j)/q(y_i)}. \quad (9)$$

The vague notation $\{x\}$ conveys an abitrary subset of the covariates with cardinality m. One sets $m = 1$ for the CAPO, and for the APO m is the size of the whole sample. The partially identified $\tilde{p}(y_t|x)$ really encompasses a set of probability densities that includes $p(y|t, x)$ and smooth deviations from it. Our importance sampler ensures normalization [Tokdar and Kass, 2010], but still some amount of sharpness is left on the table [Dorn et al., 2021]. For current

purposes, a greedy algorithm may be deployed to maximize (or minimize) Equation 9 by optimizing the weights w_i attached to each $f(y_i)$, within the range

$$\underline{w}_i := \frac{p(y_i|t,x)}{\overline{d}(t|y_i,x)q(y_i)}, \qquad \overline{w}_i := \frac{p(y_i|t,x)}{\underline{d}(t|y_i,x)q(y_i)}.$$

Our Algorithm 1 adapts the method of Jesson et al. [2021], Kallus et al. [2019] to heterogeneous weight bounds $[\underline{w}_i, \overline{w}_i]$ per draw i. View a proof of correctness in §C.

Others have framed the APO as the averaged CAPOs, and left the min/max optimizations on the CAPO level [Jesson et al., 2022]. We optimize the APO directly, but have not studied the impact of one choice versus the other.

Algorithm 1: The expectation maximizer, with $\mathcal{O}(n)$ runtime if intermediate Δ_j results are memoized.

Input : $\{(\underline{w}_i, \overline{w}_i, f_i)\}_{i=1}^n$ ordered by ascending f_i.
Output : $\max_w \mathbb{E}[f(X)]$ estimated by importance sampling with n draws.

Initialize $w_i \leftarrow \overline{w}_i$ for all $i = 1, 2, \ldots n$;
for $j = 1, 2, \ldots n$ **do**
  Compute $\Delta_j := \sum_{i=1}^n w_i(f_j - f_i)$;
  if $\Delta_j < 0$ **then**
    $w_j \leftarrow \underline{w}_j$;
  else
    break;
  end
end
Return $\sum_i w_i f_i / \sum_i w_i$

4 A SEMI-SYNTHETIC BENCHMARK

It is common practice to test causal methods, especially under novel settings, with real datasets but synthetic outcomes [Curth et al., 2021, Cristali and Veitch, 2022]. We adopted four exceedingly diverse datasets spanning health, bioinformatics, and social-science sources. Our variable-generating process preserved the statistical idiosyncracies of each dataset. Confounders and treatment were random projections of the data, which were quantile-normalized for uniform marginals in the unit interval. Half the confounders were observed as covariates and the other half were hidden. The outcome was Bernoulli with random linear or quadratic forms mixing the variables before passing through a normal CDF activation function. Outcome and propensity models were linear and estimated by maximum likelihood. See §E.

Selecting the baselines. The δMSM with Balanced Beta was benchmarked against three relevant baselines.

- (CMSM) Use the recent model by Jesson et al. [2022], where $\underline{d} := \Gamma^{-1}p(\tau|x)$, $\overline{d} := \Gamma^{+1}p(\tau|x)$.
- (uniform) Suppose $\underline{d} := \Gamma^{-1}$, $\overline{d} := \Gamma^{+1}$, as if the propensity were uniform and constant.
- (binary MSM) Shoehorn the propensity into the classic MSM [Tan, 2006] by considering the treatment as binary with indicator $\mathbb{I}[T > 0.5]$.

Note that the CMSM becomes equivalent to the "uniform" baseline above when CAPOs are concerned (Equation 9 with $m = 1$), which are not studied in this benchmark.

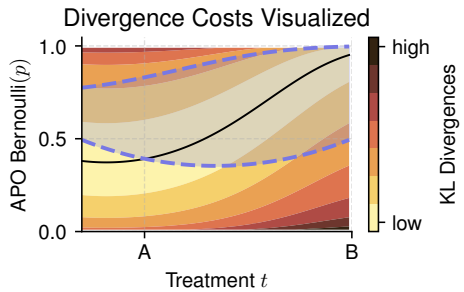

Figure 5. Divergence cost measures the size of the ignorance intervals (blue), weighted by the badness of each estimate (red). The black line is the true APO. Coverage is the portion of treatments contained between A and B in this example.

Scoring the coverages. A reasonable goal would be to achieve a certain amount of coverage [McCandless et al., 2007] of the true APOs, like having 90% of the curve be contained in the ignorance intervals. Since violation factor Γ is not entirely interpretable, nor commensurable across sensitivity models, we measure the size of an ignorance interval via a cost incurred in terms of actionable inference. For each point t of the dose-response curve, we integrated the KL divergence of the actual APO (which defines the Y_t Bernoulli parameter) against the predicted APO uniformly between the bounds. This way, each additional unit of ignorance interval is weighed by its information-theoretic approximation cost. This score is a *divergence cost* of a target coverage.

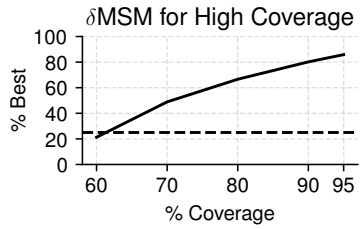

Figure 6. The δMSM outperforms the other three methods at high target coverages. Dashed line is random.

Analysis. The main results are displayed in Table 2. There were ten confounders and the true dose-response curve was a random quadratic form in the treatment and confounders. Other settings are shown in §E, Table 4. Each trial exhibited completely new projections and outcome function. Hence there were different levels and types of confounding each time, as well as varying goodness of fit.

Benchmarks	brain		blood		pbmc		mftc			ratio
	mean	(std.)	mean	(std.)	mean	(std.)	mean	(std.)	% best	to best
δMSM (ours)	**138**	(120)	**141**	(129)	**138**	(121)	**144**	(124)	**78.4**	**1.03** (0.08)
CMSM	186	(153)	188	(156)	205	(169)	182	(145)	7.8	1.81 (2.15)
uniform	158	(137)	162	(146)	157	(136)	167	(141)	4.8	1.20 (0.10)
binary MSM	211	(128)	213	(131)	222	(127)	214	(127)	9.0	2.57 (2.34)

Table 2. Semi-synthetic benchmark: divergence costs of 90% coverage of the Average Potential Outcome (APO), multiplied by 1000. The four `datasets` are listed on top. We report averages over 500 trials per experiment. A paired t-test and sign test, roughly corresponding to the mean and median, showed significant improvement by the δMSM over the others with all $P < 10^{-5}$. "% best" counts the proportion of trials that each method outperformed the rest, and "ratio to best" is the average cost ratio to the best method's in each trial—closer to one is better.

5 A REAL-WORLD EXEMPLAR

The UK Biobank [Bycroft et al., 2018] is a large, densely phenotyped epidemiological study with brain imaging. We preprocessed 40 attributes, eight of which were continuous diet quality scores (DQSs) [Said et al., 2018, Zhuang et al., 2021] valued 0–10 and serving as treatments, on 42,032 people. The outcome was thicknesses of 34 cortical brain regions. A poor DQS could translate to noticeable atrophy in the brain of some older individuals, depending on their attributes [Gu et al., 2015, Melo Van Lent et al., 2022].

Continuous treatments enable the (Conditional) Average Causal Derivative, (C)ACD $:= \partial\,\mathbb{E}[Y_t|X]\,/\,\partial t$. The CACD informs investigators on the incremental change in outcome due to a small change in an individual's given treatment. For instance, it may be useful to identify the individuals who would benefit the most from an incremental improvement in diet. We plotted the age distributions of the top 1% individuals by CACD (diet $\rightarrow$ cortical thickness) in Figure 7.

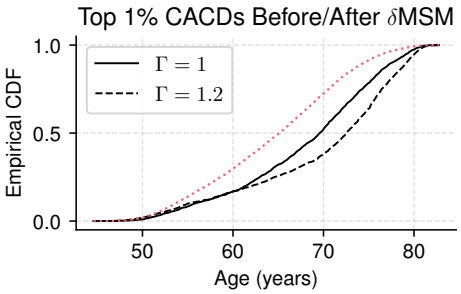

Figure 7. When we apply the δMSM ($\Gamma > 1$) for partial identification, the individuals with the top 1% causal derivatives of cortical thickness with respect to DQSs skew even older. This is expected logically because older people have more years during which they could have revised their diets. Red dotted line corresponds to the entire population.

We also compared the δMSM to an equivalent binary MSM where CACDs are computed in the latter case by thresholding the binary propensity at t. Each model's violation factor Γ was set for an equivalent amount ($\sim$30%) of nonzero CACDs. Under the δMSM, the DQSs with strongest average marginal benefit ranked as vegetables, whole grains, and then meat, for both females and males. They differed under the binary MSM, with meat, then whole grains as the top for females and dairy, then refined grains as the top for males.

6 DISCUSSION

Sensitivity analyses for hidden confounders can help to guard against erroneous conclusions from observational studies. We generalized the practice to causal dose-response curves, thereby increasing its practical applicability. However, there is no replacement for an actual interventional study, and researchers must be careful to maintain a healthy degree of skepticism towards observational results even after properly calibrating the partially identified effects.

Specifically for Average Potential Outcomes (APOs) via the sample-based algorithm, we demonstrated widespread applicability of the δMSM in §4 by showing that it provided tighter ignorance intervals than the recent CMSM and other models for 78% of all trials, notwithstanding the wide variation in scenarios tested. Ablating the approximation in Equation 2 and dropping the quadratic term, that percentage falls slightly to 74%. Even further, keeping just the constant term results in a large drop to 7%. This result suggests that the proposed Taylor expansion (Equation 2) is useful, and that terms of higher order would not give additional value.

We showcased sensical behaviors of the δMSM in a real observational case study (§5), e.g. how older people would be more impacted by (retroactive) changes to their reported diets. Additionally, the top effectual DQSs appeared more consistent with the δMSM rather than the binary MSM.

Looking ahead. Alternatives to sampling-based Algorithm 1 deserve further investigation, for computing ignorance intervals on expectations—but not only. Our analytical solutions bound the density function $p(y_t|x)$ of conditional potential outcomes, which can generate other quantities of interest [Kallus, 2022] or play a role in larger models.

REFERENCES

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

P. A. Atroszko. Is a high workload an unaccounted confounding factor in the relation between heavy coffee consumption and cardiovascular disease risk? *The American Journal of Clinical Nutrition*, 110(5):1257–1258, 2019.

M. Bonvini and E. H. Kennedy. Fast convergence rates for dose-response estimation. *arXiv preprint arXiv:2207.11825*, 2022.

P. Bromiley. Products and convolutions of gaussian probability density functions. *Tina-Vision Memo*, 3(4):1, 2003.

C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

E. J. Calabrese and L. A. Baldwin. U-shaped dose-responses in biology, toxicology, and public health. *Annual Review of Public Health*, 22(1):15–33, 2001. doi: 10.1146/annurev.publhealth.22.1.15. PMID: 11274508.

Y.-L. Chen, L. Minorics, and D. Janzing. Correcting confounding via random selection of background variables. *arXiv preprint arXiv:2202.02150*, 2022.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, J. Robins, et al. Double/debiased machine learning for treatment and causal parameters. Technical report, 2017.

V. Chernozhukov, C. Cinelli, W. Newey, A. Sharma, and V. Syrgkanis. Long story short: Omitted variable bias in causal machine learning. *arXiv preprint arXiv:2112.13398*, 2021.

K. Colangelo and Y.-Y. Lee. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*, 2021.

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

I. Cristali and V. Veitch. Using embeddings for causal estimation of peer influence in social networks. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

A. Curth, D. Svensson, J. Weatherall, and M. van der Schaar. Really doing great at estimating CATE? a critical look at ML benchmarking practices in treatment effect estimation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

J. Dorn, K. Guo, and N. Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *arXiv preprint arXiv:2112.11449*, 2021.

A. Ghassami, N. Sani, Y. Xu, and I. Shpitser. Multiply robust causal mediation analysis with continuous treatments. *arXiv preprint arXiv:2105.09254*, 2021.

J. Godos, M. Tieri, F. Ghelfi, L. Titta, S. Marventano, A. Lafranconi, A. Gambera, E. Alonzo, S. Sciacca, S. Buscemi, et al. Dairy foods and health: an umbrella review of observational studies. *International Journal of Food Sciences and Nutrition*, 71(2):138–151, 2020.

Y. Gu, A. M. Brickman, Y. Stern, C. G. Habeck, Q. R. Razlighi, J. A. Luchsinger, J. J. Manly, N. Schupf, R. Mayeux, and N. Scarmeas. Mediterranean diet and brain structure in a multiethnic elderly cohort. *Neurology*, 85(20):1744–1751, 2015.

W. Guo, M. Yin, Y. Wang, and M. Jordan. Partial identification with noisy covariates: A robust optimization approach. In *Conference on Causal Learning and Reasoning*, pages 318–335. PMLR, 2022.

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

J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A. M. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen, et al. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071, 2020.

Y. Hu, Y. Wu, L. Zhang, and X. Wu. A generative adversarial framework for bounding confounded causal effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12104–12112, 2021.

W. Huang, O. Linton, and Z. Zhang. A unified framework for specification tests of continuous treatment effect models. *Journal of Business & Economic Statistics*, 0(0):1–14, 2021. doi: 10.1080/07350015.2021.1981915.

A. Jesson, S. Mindermann, U. Shalit, and Y. Gal. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33:11637–11649, 2020.

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

A. Jesson, A. R. Douglas, P. Manshausen, M. Solal, N. Meinshausen, P. Stier, Y. Gal, and U. Shalit. Scalable sensitivity and uncertainty analyses for causal-effect estimates of continuous-valued interventions. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=PzI4ow094E.

N. Kallus. Treatment effect risk: Bounds and inference. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 213–213, 2022.

N. Kallus and M. Santacatterina. Kernel optimal orthogonality weighting: A balancing approach to estimating effects of continuous treatments. *arXiv preprint arXiv:1910.11972*, 2019.

N. Kallus, X. Mao, and A. Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd international conference on artificial intelligence and statistics*, pages 2281–2290. PMLR, 2019.

H. M. Kang, M. Subramaniam, S. Targ, M. Nguyen, L. Maliskova, E. McCarthy, E. Wan, S. Wong, L. Byrnes, C. M. Lanata, et al. Multiplexed droplet single-cell rna-sequencing using natural genetic variation. *Nature biotechnology*, 36(1):89–94, 2018.

N. Kilbertus, M. J. Kusner, and R. Silva. A class of algorithms for general instrumental variable models. *Advances in Neural Information Processing Systems*, 33:20108–20119, 2020.

J. Lim, C. X. Ji, M. Oberst, S. Blecker, L. Horwitz, and D. Sontag. Finding regions of heterogeneity in decision-making via expected conditional covariance. *Advances in Neural Information Processing Systems*, 34:15328–15343, 2021.

A. Y. Lo. A large sample study of the bayesian bootstrap. *The Annals of Statistics*, 15(1):360–375, 1987.

C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

C. F. Manski. *Partial identification of probability distributions*, volume 5. Springer, 2003.

A. Mastouri, Y. Zhu, L. Gultchin, A. Korba, R. Silva, M. Kusner, A. Gretton, and K. Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International Conference on Machine Learning*, pages 7512–7523. PMLR, 2021.

L. C. McCandless, P. Gustafson, and A. Levy. Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statist Med*, 26:2331–2347, 2007.

D. Melo Van Lent, H. Gokingco, M. I. Short, C. Yuan, P. F. Jacques, J. R. Romero, C. S. DeCarli, A. S. Beiser, S. Seshadri, J. J. Himali, et al. Higher dietary inflammatory index scores are associated with brain mri markers of brain aging: Results from the framingham heart study offspring cohort. *Alzheimer's & Dementia*, 2022.

V. B. Meresht, V. Syrgkanis, and R. G. Krishnan. Partial identification of treatment effects with implicit generative models. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=8cUGfg-zUnh.

N. Mokhberian, A. Abeliuk, P. Cummings, and K. Lerman. Moral framing and ideological bias of news. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*, pages 206–219. Springer, 2020.

K. Padh, J. Zeitler, D. Watson, M. Kusner, R. Silva, and N. Kilbertus. Stochastic causal programming for bounding treatment effects. *arXiv preprint arXiv:2202.10806*, 2022.

P. R. Rosenbaum. *Observational Studies*. Springer, 2002.

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, 1983.

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

M. A. Said, N. Verweij, and P. van der Harst. Associations of combined genetic and lifestyle risks with incident cardiovascular disease and diabetes in the uk biobank study. *JAMA cardiology*, 3(8):693–702, 2018.

E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

N. N. Taleb. (anti) fragility and convex responses in medicine. In *Unifying Themes in Complex Systems IX: Proceedings of the Ninth International Conference on Complex Systems 9*, pages 299–325. Springer, 2018.

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

E. J. T. Tchetgen, A. Ying, Y. Cui, X. Shi, and W. Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

S. T. Tokdar and R. E. Kass. Importance sampling: A review. *WIREs Computational Statistics*, 2(1):54–60, 2010.

S. Tübbicke. Entropy balancing for continuous treatments. *J Econ Methods*, 11(1):71–89, 2022.

B. G. Vegetabile, B. A. Griffin, D. L. Coffman, M. Cefalu, M. W. Robbins, and D. F. McCaffrey. Nonparametric estimation of population average dose-response curves using entropy balancing weights for continuous exposures. *Health Services and Outcomes Research Methodology*, 21(1):69–110, 2021.

V. Veitch and A. Zaveri. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10999–11009. Curran Associates, Inc., 2020.

S. Yadlowsky, H. Namkoong, S. Basu, J. Duchi, and L. Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*, 2020.

L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.

M. Yin, C. Shi, Y. Wang, and D. M. Blei. Conformal sensitivity analysis for individual treatment effects. *arXiv preprint arXiv:2112.03493v2*, 2021.

E. Ystrom, E. Degerud, M. Tesli, A. Høye, T. Reichborn-Kjennerud, and Ø. Næss. Alcohol consumption and lower risk of cardiovascular and all-cause mortality: the impact of accounting for familial factors in twins. *Psychological Medicine*, pages 1–9, 2022.

G. U. Yule. NOTES ON THE THEORY OF ASSOCIATION OF ATTRIBUTES IN STATISTICS. *Biometrika*, 2(2):121–134, 02 1903. ISSN 0006-3444. doi: 10.1093/biomet/2.2.121. URL https://doi.org/10.1093/biomet/2.2.121.

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

P. Zhuang, X. Liu, Y. Li, X. Wan, Y. Wu, F. Wu, Y. Zhang, and J. Jiao. Effect of diet quality and genetic predisposition on hemoglobin a1c and type 2 diabetes risk: gene-diet interaction analysis of 357,419 individuals. *Diabetes Care*, 44(11):2470–2479, 2021.