Bounding the Effects of Continuous Treatments for Hidden Confounders

Anonymous Author(s)
Affiliation
Address
email

Abstract

Observational studies often seek to infer the causal effect of a treatment even though both the assigned treatment and the outcome depend on other confounding variables. An effective strategy for dealing with confounders is to estimate a propensity model that corrects for the relationship between covariates and assigned treatment. Unfortunately, the confounding variables themselves are not always observed, in which case we can only bound the propensity, and therefore bound the magnitude of causal effects. In many important cases, like administering a dose of some medicine, the possible treatments belong to a continuum. Sensitivity models, which are required to tie the true propensity to something that can be estimated, have been explored for binary treatments. We propose one for *continuous treatments*. We develop a framework to compute *ignorance intervals* on the partially identified dose-response curves, enabling us to quantify the susceptibility of an inference to hidden confounders. We show with real-world observational studies that our approach can give non-trivial bounds on causal effects from continuous treatments in the presence of hidden confounders.

1 Introduction

The goal of causal inference is to separate the effect of one treatment variable from the influence of many related but irrelevant "confounding" variables. Physical interventions accomplish this most effectively, but practical barriers often force researchers to rely on observational studies and clever statistics. A plethora of methods operate on the basis of a learned propensity model for the assigned treatment conditioned on covariates, for instance to reweigh the sample and remove any visible biases. Usually, the covariates are inadequate to account for all the hidden paths between the treatment and the outcome, and propensity-based approaches may struggle to discern the real effect. The scientific community at large continues to vascillate on the health implications and ideal consumption levels of coffee [Atroszko, 2019], alcohol [Ystrom et al., 2022], and cheese [Godos et al., 2020], to name a few substances.

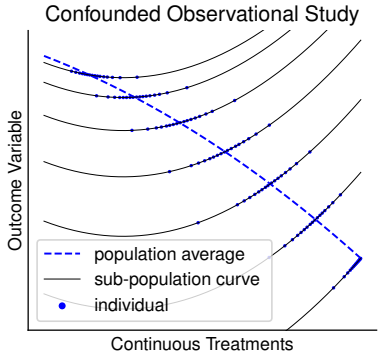

Figure 1. When a confounder is distorting the assigned treatments in sub-populations, the overall population-level trend may appear flipped in comparison to each sub-population's dose response.

Most sensitivity analyses to hidden confounding require the treatment categories to be binary or at least discrete. This weakens the empirical studies that would have been better specified by dose-response curves [Calabrese and Baldwin, 2001] from a continuous treatment variable, for example. Estimated

Submitted to the NeurIPS Causal ML for Real-World Impact Workshop (CML4Impact 2022). Do not distribute.

dose responses are indeed vulnerable in the presence of hidden confounders. A simulated example in Figure 1 demonstrates how a J-shaped treatment effect can appear flipped in observational data due to confounding. The phenomenon is an example of Simpson's paradox [Simpson, 1951, Yule, 1903].

1.1 Related works

There is growing interest in causal methodology for treatments (or exposures, interventions) that take on specific values within a continuum, especially in the fields of econometrics [e.g. Huang et al., 2021, Tübbicke, 2022], health sciences [Vegetabile et al., 2021], and machine learning [Ghassami et al., 2021, Colangelo and Lee, 2021, Kallus and Santacatterina, 2019]. So far, much scrutiny on partially identified potential outcomes has focused on the case of binary treatments, the simplest setting [e.g. Rosenbaum and Rubin, 1983, Louizos et al., 2017, Lim et al., 2021]. A number of creative approaches were exhibited in the past few years to make strides in this binary setting. Most of them relied on a *sensitivity model* for bounding the extent of possible unobserved confounding, to which downstream tasks may be adapted by noting which treatment effects degrade the quickest.

Quite recently, attempts were made to handle unobserved confounding with continuous treatments by optimizing the treatment effect bounds with generative models [Padh et al., 2022, Hu et al., 2021], rather than a sensitivity model. One employs instrumental variables [Kilbertus et al., 2020]. Another, with a sensitivity model, was developed in parallel to the present work [Jesson et al., 2022].

Regarding binary treatments, the so-called Marginal Sensitivity Model (MSM) due to Tan [2006] continues to be studied extensively [Zhao et al., 2019, Veitch and Zaveri, 2020, Yin et al., 2021]. Variations thereof include Rosenbaum's earlier sensitivity model [2002] that enjoys ties to regression coefficients [Yadlowsky et al., 2020]. Other groups have borrowed strategies from deep learning [Wu and Fukumizu, 2022] rather than opting for the MSM. Another active line of work constructs bounds not due to ignorance on confounding but instead viewed from the lens of robustness [Guo et al., 2022, Makar et al., 2020, Johansson et al., 2020]. The MSM is highly interpretable with its single free parameter, and applicable to a wide swath of models.

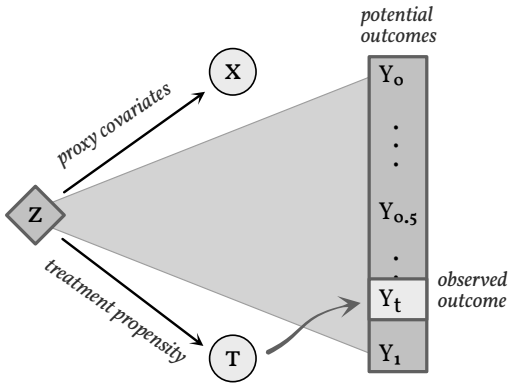

Figure 2. Illustration of true confounder Z determining potential outcomes $Y_{t\in[0,1]}$ with observable covariate X and treatment T. The density $p(y_t|\tau, x)$, found in the integrand of Equation 1, diverges from $p(y_t|x)$ when the covariates are inadequate to block all the links between assigned treatment and potential outcomes.

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

Our first contribution is to propose a unique sensitivity model (§2.1) that extends the MSM to a treatment continuum. Next, we derive general (§3) and specialized (§3.2) formulas. We devise an efficient algorithm (§C) to compute ignorance bounds over dose-response curves, following up with experiments on real (§4) datasets.

2 Potential outcomes

Causal inference is often cast in the nomenclature of potential outcomes, due to Rubin [1974]. The broad goal is to measure a treatment's effect on an individual, marked by a set of covariates, while accounting for all the confounding between the covariates and the treatment variable. Effects could manifest heterogenously across individuals. In non-interventional settings, observed covariates may not entirely overlap across treatment regimens. It is typical to estimate two models, (1) the outcome

predictor and (2) a model for the propensity of treatment conditioned on the covariates. The latter may help account for biases.

The first two assumptions involved in Rubin's framework are that observations of outcome, assigned treatment, and covariates $\{Y^{(i)}, T^{(i)}, X^{(i)}\}$ are i.i.d draws from the same population and that all treatments have a chance to occur for each covariate vector: $p_{T|X}(t \mid x) > 0$ (overlap/positivity) for all $t, x \in [0,1] \times \mathcal{X}$, specifically in our context of continuous treatments. The third and most challenging of these fundamental assumptions is that of *ignorability*, or sufficiency. Our study is concerned with the scenarios where that assumption is violated: when there exists a dependency, not blocked by the covariates, between the assigned treatment and true potential outcomes. Let $p(y_t|x)$ denote the probability density function of *potential* outcome $Y_t = y_t$ from a treatment $t \in [0,1]$, given covariates $X = x$. Formally, we have a violation of ignorability:

$$\{(Y_t)_{t\in\mathcal{T}} \not\perp\!\!\!\perp T\} \mid X.$$

It is only realistic to observe samples of $Y|T = t, X = x$ with density $p(y_t|t,x)$. However, to account for possible hidden confounding, we also require a $p(y_t|\tau \neq t, x)$ for quantifying treatment effects of the general form $\mathbb{E}[f(Y_t)|X]$, involving the density

$$p(y_t|x) = \int_0^1 p(y_t|\tau, x)p(\tau|x)\,\mathrm{d}\tau, \tag{1}$$

where $p(y_t|\tau, x)$ is the distribution of potential outcomes conditioned on actual treatment $T = \tau \in [0,1]$ that may differ from the potential outcome's index t. 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 aim to develop some intuition before introducing the novelties.

On notation. We will use the shorthand $p(\cdots)$ with lowercase variables whenever working with probability densities of the corresponding stochastic variables in uppercase. In other words,

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

Interpretation. How would one interpret $p(y_t|\tau, x)$? The potential-outcomes vector $(Y_t)_{t\in[0,1]}$ of infinite dimensionality is *intrinsic* to each individual with true confounder Z, for which X is a noisy proxy. By "true" confounder we refer to any set of variables that suffice to block all backdoor paths between Y_t and T. The potential-outcomes vector would only change from knowledge of assigned treatment $T = \tau$ if it betrayed additional information about Z, absent in X, that further informed any Y_t. We may express $p(y_t|\tau, x)$ explicitly in terms of hypothetical true confounders as $\int p(y_t|z)p(z|\tau, x)\,\mathrm{d}z$ because z subsumes both x and τ. This way, $p(y_t|z)$ is the true potential outcome and $p(z|\tau, x)$ acts as a filter for how parts of the true confounder mix together into the proxy x and the assigned treatment τ.

Propensities. The probability density $p(\tau|x)$ is termed the *nominal propensity*. A quantity often examined is the *complete propensity*, specifically referring to $p(\tau|\, y_t, x)$ in our realm. The complete propensity can differ from $p(\tau|x)$ because of hidden confounders. In that instance, conditioning on potential outcome y_t modulates the distribution. Similarly, by connection through Bayes' rule, conditioning the potential outcomes $p(y_t|x)$ on assigned treatment τ modulates those distributions. Absent any unobserved confounding, $p(y_t|\tau, x) = p(y_t|x)$ and Equation 1 trivializes. See Figure 2 for a graphical illustration on the runaway influence of τ on the potential outcomes.

Sensitivity. Explored by Kallus et al. [2019] and 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 $T' \in \{0,1\}$. Specifically, it couples the odds of treatment under the nominal propensity to the odds of treatment under complete propensity, limiting the discrepancy:

Definition 1 (The Marginal Sensitivity Model)**.** For binary treatment $t' \in \{0,1\}$ and violation factor $\Gamma \geq 1$, the following ratio is bounded: $\Gamma^{-1} \leq \left[\frac{p(t'|x)}{1-p(t'|x)}\right]^{-1}\left[\frac{p(t'|\, y_{t'}, x)}{1-p(t'|\, y_{t'}, x)}\right] \leq \Gamma$.

Restricting ourselves to binary treatments affords us 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 $t \in [0,1]$, we must contend with infinite treatments and infinite potential outcomes.

2.1 Towards continuous sensitivity

We require a constraint on the quantity $p(\tau|\, y_t, x)$ that is fundamentally unknowable across all the combinations of assigned treatments $T = \tau \in [0,1]$ and potential outcomes $y_{t\in[0,1]}$. As with the MSM, our target is to associate $p(\tau|\, y_t, x)$ to the knowable $p(\tau|x)$. In other words, we seek to constrain the knowledge conferred on propensity by a single potential outcome y_t. It is not necessary for the functions pertaining to $(y_t)_{t\in[0,1]}$ to exhibit any degree of smoothness in t. The potential-outcome variables are treated as entries in an infinitely long vector. However, we do impose that the propensity probability densities $p(\tau|\dots)$ are at least once differentiable in τ. What sort of analogue exists for the notion of "odds" in the MSM?

Contrast treatment τ versus $\tau + \delta$ locally, for some infinitesimal δ, at any part of the curve. A translation of the MSM might appear as $\left[\frac{p(\tau+\delta|x)}{p(\tau|x)}\right]^{-1}\left[\frac{p(\tau+\delta|\, y_t,x)}{p(\tau|\, y_t,x)}\right]$. Let us peer into one of those ratios. In logarithms,

$$\delta^{-1}\log\frac{p(\tau+\delta|x)}{p(\tau|x)} = \frac{\log p(\tau+\delta|x) - \log p(\tau|x)}{\delta} \xrightarrow[\delta\to 0]{} \frac{\partial \log p(\tau|x)}{\partial\tau} \triangleq \partial_\tau \log p(\tau|x).$$

Hence, we introduce the infinitesimal MSM (δMSM), tying $\partial_\tau \log p(\tau|\, y_t, x)$ to $\partial_\tau \log p(\tau|x)$.

Definition 2 (The Infinitesimal Marginal Sensitivity Model)**.** For treatments in the closed unit interval, $t \in [0,1]$, and violation factor $\Gamma \geq 1$, the following inequality holds everywhere:

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

We crafted the δMSM with the intention of functionally mirroring the MSM—locally, on a treatment continuum. Whereas Definition 2 is stated in logarithms, Definition 1 is not; the difference is merely cosmetic and hyperparameter Γ plays an equivalent role in both structures. Nevertheless, the emergent properties are vastly different.

3 The framework

We list the core assumptions surrounding our problem.

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

Assumption 2 (Fully Observed Confounding at No Treatment)**.** The utter lack of treatment is not informed by potential outcomes: $p(\tau = 0|\, y_t, x) = p(\tau = 0|\, x)$ for all t and y_t.

Assumption 2 states that we look for sensitivity to hidden confounders outside the control group at $T = 0$. The restriction is reasonable in situations like the following: we seek to estimate the effect of a prescription drug, and some clinics prescribe different dosages. Our $T = 0$ group would be individuals who have not received any such prescription, and $T > 0$ would place patients on a scale depending on prescription dosage. We expect a dramatically lessened vulnerability to hidden confounders for the well-represented—in observed *and* unobserved attributes—control group. From a technical perspective, Assumption 2 is necessary for our derivations, and should be interpreted as a blind spot in the sensitivity model rather than a requirement for the underlying process. There is no additional constraint, besides the δMSM itself, on how much the complete propensity function may fluctuate around any $T > 0$. We motivate and validate this assumption in the real world with §4.

Next, we proceed with derivations. The key to cracking open Equation 1 is to carve out a region inside the domain of integration where an approximation can be trusted. This will extrapolate from the singular point $\tau = t$ where estimation is feasible.

3.1 Dealing with an unreliable approximation

We approximate $p(y_t|\tau, x)$ 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

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

Denote with $\tilde{p}(y_t|\tau, x)$ an approximation of first or second order as laid out above. We will encounter that even $\partial_\tau p(y_t|\tau, x)|_{\tau=t}$ is intractable. Thankfully, it can be bounded using the δMSM machinery. Let us quantify the reliability of this approximation by a set of weights $0 \leq w_t(\tau) \leq 1$, where typically (but need not necessarily) $w_t(t) = 1$. Decompose the integral of Equation 1—

$$\begin{aligned} p(y_t|x) &= \int_0^1 w_t(\tau)p(y_t|\tau, x)p(\tau|x)\,\mathrm{d}\tau + \int_0^1 [1 - w_t(\tau)]p(y_t|\tau, x)p(\tau|x)\,\mathrm{d}\tau \\ &\approx \int_0^1 \underbrace{w_t(\tau)\tilde{p}(y_t|\tau, x)p(\tau|x)\,\mathrm{d}\tau}_{(A)\text{ the approximated quantity}} + \int_0^1 \underbrace{[1 - w_t(\tau)]p(\tau|y_t, x)p(y_t|x)\,\mathrm{d}\tau}_{(B)\text{ by Bayes' rule}}. \end{aligned} \quad (3)$$

This separation into recoverable (A) and entirely unknown (B), demarcated by the weights, ensures that the inaccurate regimes of the approximation vanish (as $w_t(\tau) \to 0$ away from t) and are replaced with the ignorant quantity. We simplify part B of Equation 3 first, into $p(y_t|x)[1 - \int_0^1 w_t(\tau)p(\tau|y_t, x)\,\mathrm{d}\tau]$. We witness already that $p(y_t|x)$ shall take the form of

$$p(y_t|x) \approx \frac{\int_0^1 w_t(\tau)\tilde{p}(y_t|\tau, x)p(\tau|x)\,\mathrm{d}\tau}{\int_0^1 w_t(\tau)p(\tau|y_t, x)\,\mathrm{d}\tau}. \quad (4)$$

How the approximation error of Equation 2 carries into Equation 4 depends on the peakedness of the weight function. To proceed further demands reflecting on Assumptions 1 & 2, as we do in §A.

A note on ensemble uncertainty. One should quantify empirical uncertainties [Jesson et al., 2020] alongside sensitivity to hidden confounding. In our experiments we learn both the predictor and the propensity model as ensembles from bootstrapped resampled [Lo, 1987] data. Then $\tilde{p}(y_t|x)$ can also be resampled for confidence intervals via its component ensembles.

3.2 Tractable weight combinations

In addition to developing the general framework above, we derive analytical forms for a specific paramametrization to the weighting function and propensity distribution. Here, we look to the Beta function and its associated probability density for a natural solution. Suppose that

$$(T \mid X = x) \sim \text{Beta}(\alpha(x), \beta(x)), \qquad \text{for arbitrary } \alpha(x), \beta(x), \quad (5)$$

$$w_t(\tau) = \frac{\tau^{a_t - 1}(1 - \tau)^{b_t - 1}}{c_t} = \frac{\tau^{rt}(1 - \tau)^{r(1-t)}}{c_t}, \qquad a_t + b_t = r + 2, \quad r > 0. \quad (6)$$

We designed the reliability weights to mirror the propensity's form by rescaling a Beta density. We assert that $w_t(\tau)$ peaks at $\tau = t$, and that $w_t(\tau) = 1$. We find that $c_t \triangleq t^{rt}(1 - t)^{r(1-t)}$, even though the solution is irrelevant for our purposes. The mode is fixed: $(a_t - 1)/(a_t + b_t - 2) = t$. See §B for details on the solution.

4 Results from an observational study

The most pertinent application for the framework laid out above is an observational study with incomplete or noisy covariates and a continuous treatment variable. More concretely, the treatment variable should be transformed and scaled into the unit interval such that $T = 0$ signifies a control with a complete lack of treatment. Every *kind* of individual should be about equally likely to fall in the $(T = 0)$ cohort (Assumption 2.) As for shaping the $(T > 0)$ regime, the domain should inform whether a linear scale is employed, versus an empirical or parametric cumulative density function.

We obtained real observational data from the UK Biobank and performed one semi-synthetic and one fully empirical experiment. In the latter, we relied on discarding covariates to induce greater hidden confounding. See §D for details.

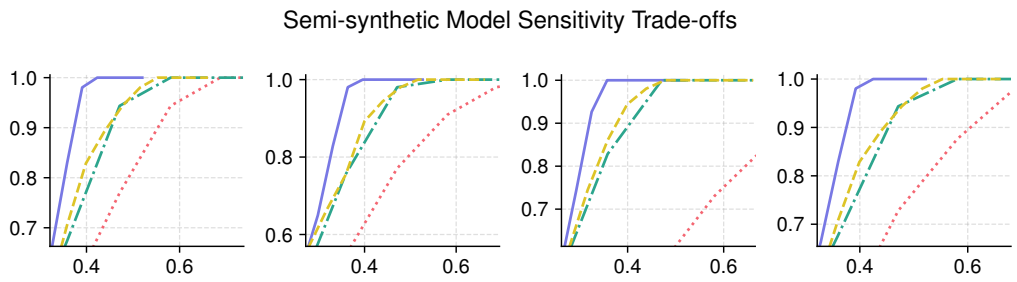

Figure 3. Ignorance (horizontal) versus recall (vertical) in four replications of a counterfactual model estimated on real Biobank covariates with a synthetic binary outcome, and known dose response.

The objective. We chose to investigate the coverage [McCandless et al., 2007] of $\mathbb{E}[Y_t]$ from ignorance intervals on counterfactual models trained with inadequate covariates. In the semi-synthetic case (Figure 3), the covariates were real but the outcome was simulated, and the treatment effect was known to be linear after a logit transformation. Coverage in the empirical case (Figure 4) was more difficult in the absence of a ground truth. There, we trained an *uncensored* model on an expanded set of covariates to act as an approximate target for the smaller model's ignorance intervals. In both cases, the *recall* was expressed as the portion of the dose-response curve that satisfied the relevant objective; *ignorance* was, (*a*) the average width of the intervals in logits for the semi-synthetic, and (*b*) normalized to the 95%-confidence intervals of the uncensored model for the empirical study.

The comparisons. We compared our δMSM with $r = 32$ throughout (**solid** in the figures on this page) to other sensitivity models: namely, (**dotted**) an analogue developed independently and in parallel to the present work, with just one free parameter [Jesson et al., 2022]; (**dashed**) the product of shoehorning a continuous model into the binary MSM by triggering a binary treatment at $T > 0.5$ and discretizing the propensity at the threshold; and (**dot/dash**) a baseline sensitivity model that emerges from Γ-scaling the Algorithm 1 weights without any propensity.

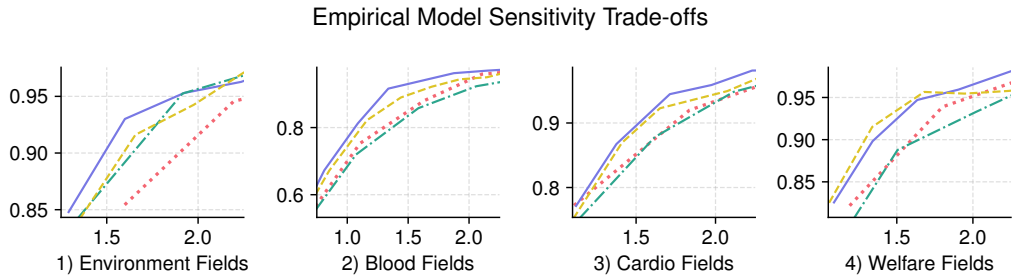

Figure 4. Ignorance (up to 2.25) versus recall for the model estimated on the real censored dataset. Four disjoint sections of the covariates were censored for the different panels. Curves begin at $\Gamma := 1$.

5 Discussion

The utility of our framework is evident in the above showcased results. We demonstrated that, in the presence of a semi-synthetic ground truth, the δMSM covers the full dose-response curve most efficiently. In the empirical study, we showed that discrepancies in the potential-outcomes continuum between a censored model and a full model are most efficiently bridged under the δMSM assumption.

Ethical implications. Sensitivity models for hidden confounders can help to guard against erroneous conclusions from observational studies. We generalized this line of analysis to the regime of continuous treatments, thereby increasing its practical applicability. The method also bears utility in the context of fairness in machine learning. It can help decision makers identify subpopulations for whom the sample is too biased to reliably draw conclusions. Nevertheless, researchers must be careful to maintain a healthy degree of skepticism towards observational results even after properly calibrating the partially identified effects.

References

Paweł 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.

Edward J Calabrese and Linda 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.

You-Lin Chen, Lenon Minorics, and Dominik Janzing. Correcting confounding via random selection of background variables. *arXiv preprint arXiv:2202.02150*, 2022.

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

AmirEmad Ghassami, Numair Sani, Yizhen Xu, and Ilya Shpitser. Multiply robust causal mediation analysis with continuous treatments. *arXiv preprint arXiv:2105.09254*, 2021.

Justyna Godos, Maria Tieri, Francesca Ghelfi, Lucilla Titta, Stefano Marventano, Alessandra Lafranconi, Angelo Gambera, Elena Alonzo, Salvatore Sciacca, Silvio 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.

Wenshuo Guo, Mingzhang Yin, Yixin Wang, and Michael Jordan. Partial identification with noisy covariates: A robust optimization approach. In *First Conference on Causal Learning and Reasoning*, 2022.

Yaowei Hu, Yongkai Wu, Lu Zhang, and Xintao 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.

Wei Huang, Oliver Linton, and Zheng 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.

Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33:11637–11649, 2020.

Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. *ICML*, 2021.

Andrew Jesson, Alyson Douglas, Peter Manshausen, Nicolai Meinshausen, Philip Stier, Yarin Gal, and Uri Shalit. Scalable sensitivity and uncertainty analysis for causal-effect estimates of continuous-valued interventions. *arXiv preprint arXiv:2204.10022*, 2022.

Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

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

Nathan Kallus, Xiaojie Mao, and Angela 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.

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

Justin Lim, Christina X Ji, Michael Oberst, Saul Blecker, Leora Horwitz, and David Sontag. Finding regions of heterogeneity in decision-making via expected conditional covariance. *Advances in Neural Information Processing Systems*, 34:15328–15343, 2021.

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

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Maggie Makar, Fredrik Johansson, John Guttag, and David Sontag. Estimation of bounds on potential outcomes for decision making. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6661–6671. PMLR, 13–18 Jul 2020.

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International Conference on Machine Learning*, pages 7512–7523. PMLR, 2021.

Wesley N. Mathews Jr., Mark A. Esrick, ZuYao Teoh, and James K. Freericks. A physicist's guide to the solution of kummer's equation and confluent hypergeometric functions. *arXiv preprint arXiv:2111.04852*, 2021.

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

Karla L. Miller, Fidel Alfaro-Almagro, Neal K. Bangerter, David L. Thomas, Essa Yacoub, Junqian Xu, Andreas J. Bartsch, Saad Jbabdi, Stamatios N. Sotiropoulos, Jesper L. R. Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Thomas W. Okell, Peter Weale, Iulius Dragonu, Steve Garratt, Sarah Hudson, Rory Collins, Mark Jenkinson, Paul M. Matthews, and Stephen M. Smith. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nat Neurosci*, 19: 1523–1536, 2016.

Kirtan Padh, Jakob Zeitler, David Watson, Matt Kusner, Ricardo Silva, and Niki 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.

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

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

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

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

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

Brian G Vegetabile, Beth Ann Griffin, Donna L Coffman, Matthew Cefalu, Michael W Robbins, and Daniel 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.

Victor Veitch and Anisha 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.

Pengzhou Abel Wu and Kenji Fukumizu. β-intact-VAE: Identifying and estimating causal effects under limited overlap. In *International Conference on Learning Representations*, 2022.

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*, 2020.

Mingzhang Yin, Claudia Shi, Yixin Wang, and David M. Blei. Conformal sensitivity analysis for individual treatment effects. *arXiv preprint arXiv:2112.03493v2*, 2021.

Eivind Ystrom, Eirik Degerud, Martin Tesli, Anne Høye, Ted Reichborn-Kjennerud, and Øyvind 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. Undy 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`.

Qingyuan Zhao, Dylan S. Small, and Bhaswar 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.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We introduce a marginal sensitivity model for continuous treatments.
   (b) Did you describe the limitations of your work? [Yes] Most notably, Assumption 2.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We mentioned in §5 how this line of work can help to guard against erroneous conclusions from observational studies. However, there is no replacement for an actual interventional study, and we must be careful to maintain a healthy degree of skepticism on observational results even after performing the sensitivity analysis.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Assumptions 1–3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See §A–§C.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See §D and the attached source code. A link to the Julia package on Github will be made available upon publication.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See §D.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We studied bootstrapped 95% confidence intervals for all our empirical results, namely in §4.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] These results were not bottlenecked by computational resources. A couple of Nvidia 2080 Ti GPUs were used for training, and a Macbook Pro for the rest of the computations.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] All owners of the datasets used herein were credited in §D. Besides common Julia packages and frameworks, all novel functionality was implemented by the first author.
   (b) Did you mention the license of the assets? [Yes] See §D.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The source code will be included as a Github link in the de-anonymized version.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] Our datasets were acquired from institutions with clear guidelines.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] This study only reported broad aggregations upon the data.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Additional derivations

We will expand the denominator of Equation 4 first, and then repurpose the results for the derivatives of Equation 2 that appear in the numerator. This part shows how the assumed model serves to characterize the unknown quantities, but the impatient reader may skip to §3.2. Without loss of generality, consider

$$\partial_\tau \log p(\tau|y_t, x) = \partial_\tau \log p(\tau|x) + \gamma(\tau|y_t, x), \qquad |\gamma(\tau|y_t, x)| \leq \log \Gamma. \tag{7}$$

We may attempt to integrate both sides;

$$\int_0^{t'} \partial_\tau \log p(\tau|y_t, x)\,\mathrm{d}\tau = \int_0^{t'} \partial_\tau \log p(\tau|x)\,\mathrm{d}\tau + \underbrace{\int_0^{t'} \gamma(\tau|y_t, x)\,\mathrm{d}\tau}_{\triangleq \lambda(t'|y_t, x)}$$

$$\begin{aligned} \iff \log p(\tau = t'|y_t, x) - \log p(\tau = 0\,|y_t, x) &= \log p(\tau = t'|x) - \log p(\tau = 0\,|x) + \lambda(t'|y_t, x), \\ \log p(\tau|y_t, x) &= \log p(\tau|x) + \lambda(\tau|y_t, x) \qquad \text{(by Assumption 2).} \end{aligned}$$

$$\therefore \quad p(\tau|y_t, x) = p(\tau|x)\Lambda(\tau|y_t, x), \qquad \Lambda \triangleq \exp\{\lambda\}. \tag{8}$$

Clearly $|\lambda(\tau|y_t, x)| \leq \tau \log \Gamma$ because it integrates γ, bounded by $\pm \log \Gamma$, over a support with length τ. Subsequently $\Lambda(\tau|y, t)$ is bounded by $\Gamma^{\pm\tau}$. We are now equipped with the requisite tools to properly bound $p(y_t|x)$—or an approximation thereof, erring on ignorance via reliability weights $w_t(\tau)$.

Consider Equation 3.A:

$$\begin{aligned} \int_0^1 w_t(\tau)\tilde{p}(y_t|\tau, x)p(\tau|x)\,\mathrm{d}\tau &= \underbrace{p(y_t|t, x)\int_0^1 w_t(\tau)p(\tau|x)\,\mathrm{d}\tau}_{(A.0)} \\ &+ \underbrace{g_1(y_t|t, x)\int_0^1 w_t(\tau)(\tau - t)p(\tau|x)\,\mathrm{d}\tau}_{(A.1)} + \underbrace{g_2(y_t|t, x)\int_0^1 w_t(\tau)\frac{(\tau - t)^2}{2}p(\tau|x)\,\mathrm{d}\tau}_{(A.2)}, \\ &\text{where} \quad g_k(y_t|t, x) \triangleq \partial_\tau^k p(y_t|\tau, x)|_{\tau=t}. \end{aligned} \tag{9}$$

Lightening the notation with a shorthand for the weighted expectations, $\langle\cdot\rangle_\tau \triangleq \int_0^1 w_t(\tau)(\cdot)p(\tau|x)\,\mathrm{d}\tau$, it becomes apparent that we must grapple with the pseudo-moments $\langle 1\rangle_\tau$, $\langle\tau - t\rangle_\tau$, and $\langle(\tau - t)^2\rangle_\tau$. Note that t should not be mistaken for a "mean" value.

Furthermore, we have yet to fully characterize $g_k(y_t|t, x)$. Observe that

$$p(y_t|\tau, x) = \frac{p(\tau|y_t, x)p(y_t|x)}{p(\tau|x)} \qquad \iff \qquad \partial_\tau p(y_t|\tau, x) = p(y_t|x) \cdot \frac{\partial}{\partial\tau}\frac{p(\tau|y_t, x)}{p(\tau|x)}.$$

The $p(y_t|x)$ will be moved to the other side of the equation as needed; by Equation 8,

$$\frac{\partial}{\partial\tau}\frac{p(\tau|y_t, x)}{p(\tau|x)} = \frac{\partial}{\partial\tau}\Lambda(\tau|y_t, x).$$

Expanding,

$$\begin{aligned} &= \frac{\partial}{\partial\tau}\exp\left\{\int_0^\tau \gamma(\tau|y_t, x)\,\mathrm{d}\tau\right\} \quad = \quad \gamma(\tau|y_t, x)\exp\left\{\int_0^\tau \gamma(\tau|y_t, x)\,\mathrm{d}\tau\right\} \\ &= (\gamma\Lambda)(\tau|y_t, x). \end{aligned}$$

Appropriate bounds will be calculated for $g_2(y_t|t, x)$ next, utilizing the finding above as their main ingredient. Let

$$\tilde{g}_k(y_t|t, x) \triangleq p(y_t|x)^{-1}g_k(y_t|t, x) = \left(\frac{\partial}{\partial\tau}\right)^k \frac{p(\tau|y_t, x)}{p(\tau|x)}\bigg|_{\tau=t.}$$

The second derivative may be calculated in terms of the ignorance quantities γ, Λ:

$$\begin{aligned}\tilde{g}_2(y_t|t,x) =& \partial_\tau \gamma(\tau|y_t,x)\Lambda(\tau|y_t,x) \\ =& \gamma(\tau|y_t,x)^2\Lambda(\tau|y_t,x) + \dot{\gamma}(\tau|y_t,x)\Lambda(\tau|y_t,x) \\ =& (\gamma^2+\dot{\gamma})\Lambda(\tau|y_t,x).\end{aligned}$$

And finally we address $\tilde{p}(y_t|x)$. Carrying over the components of Equation 9 into Equation 3,

$$\begin{aligned}\tilde{p}(y_t|x) &= \frac{p(y_t|t,x)\langle 1\rangle_\tau}{\langle\Lambda(\tau|y_t,x)\rangle_\tau - \tilde{g}_1(y_t|t,x)\langle\tau-t\rangle_\tau - \tilde{g}_2(y_t|t,x)\langle(\tau-t)^2\rangle_\tau} \\ &= \frac{p(y_t|t,x)}{\mathbb{E}_\tau[\Lambda(\tau|y_t,x)] - (\gamma\Lambda)(t|y_t,x)\,\mathbb{E}_\tau[\tau-t] - \frac{1}{2}((\dot{\gamma}+\gamma^2)\Lambda)(t|y_t,x)\,\mathbb{E}_\tau[(\tau-t)^2]},\end{aligned} \tag{10}$$

where these 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 $\gamma(t|y_t,x)$. To make use of this formula, one first procures the set of admissible $d(t|y_t,x) \in [\underline{d}(t|y_t,x), \overline{d}(t|y_t,x)]$ that violate ignorability up to a factor Γ according to the δMSM. Then, considering their reciprocals as importance weights [Tokdar and Kass, 2010], tight bounds on the partially identified expectations over $\tilde{p}(y_t|x)$ may be optimized.

B Analytical solutions for the Beta parametrization

The remaining degree of freedom disappears by a precision constraint $a_t + b_t - 2 = r$ for some $r > 0$. Constraining a more complex dispersion statistic like variance is much more difficult. The expectations found in Equation 10 are now available in closed form, and can be bounded in terms of just two extra free parameters, Γ and r. Guidance on setting the violation factor Γ is discussed elsewhere, e.g. §4; as for the class of weights, high r conveys poor trust in the Equation 2 approximation, as studied in §B.

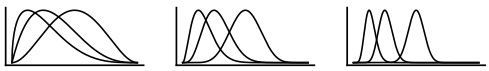

Figure 5. Beta weight schemes $w_t(\tau)$ in the unit square, plotted for centers $t = 0.125, 0.25, 0.5$. Shapes are symmetrical about $t = 0.5$. Trust declines with r.

The findings. We pose a third and final assumption, which enables us to state Proposition 1. The main insight to unlocking those expectations is that each one of them involves an integral with the product $w_t(\tau)p(\tau|x)$ over its normalization constant, yielding the moments of a Beta distribution.

Assumption 3 (Second-order Simplification)**.** The quantity $\dot{\gamma}(\tau|y_t,x)$ cannot be characterized as-is. We grant that γ^2 dominates, and consequently $\left|(\dot{\gamma}+\gamma^2)\Lambda\right| \leq \left|\gamma^2\Lambda\right| + \varepsilon$ for small $\varepsilon \geq 0$.

Proposition 1 (Beta Parametrizations)**.** *The formulations in Equations 5 & 6 admit analytical solutions to the ignorance denominator in Equation 10. With α, β implicitly referring to $\alpha(x), \beta(x)$,*

$$\mathbb{E}_\tau[\Lambda(\tau|y_t,x)] \in {}_1F_1\big(\alpha + a_t - 1;\ \alpha+\beta+r;\ \pm\log\Gamma\big),$$

where again the operator $\mathbb{E}_\tau$ is employed as in Equation 10, and ${}_1F_1$ denotes Kummer's confluent hypergeometric function [Mathews Jr. et al., 2021]. In addition, $\mathbb{E}_\tau[\tau-t]$ and $\mathbb{E}_\tau[(\tau-t)^2]$ can be readily computed by means of the first two moments of the Beta distribution.

C Computing the ignorance intervals

After deriving $\tilde{p}(y_t|x)$ in Equation 10 and a specific solution with Proposition 1, we must find a way to bound the partially identified expectations with respect to this distribution. Concretely, we seek to characterize the Individual Treatment Effect (ITE) $\mathbb{E}[f(Y_t)|X=x]$ or Average Treatment Effect (ATE) $\mathbb{E}[F(Y_t)]$ for any task-specific $f(y)$. This is accomplished with a Monte Carlo importance sampler of n outcome realizations y_i drawn from proposal $q(y)$:

$$\tilde{\mathbb{E}}[f(Y_t)|X=x] = \frac{\sum_{i=1}^n f(y_i)\tilde{p}(y_t=y_i|x)/q(y_i)}{\sum_{i=1}^n \tilde{p}(y_t=y_i|x)/q(y_i)}. \tag{11}$$

For the ATE, one additionally averages over covariates with sample size m:

$$\tilde{\mathbb{E}}[f(Y_t)] = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} f(y_i)\tilde{p}(y_t = y_i|x_j)/q(y_i)}{\sum_{i=1}^{n}\sum_{j=1}^{m} \tilde{p}(y_t = y_i|x_j)/q(y_i)}. \tag{12}$$

Even though $\tilde{p}(y_t|x)$ is a normalized probability density, it contains partially identified quantities. It is untenable to constrain a search along the candidate values for each $d(t|y_t = y_i, x)$ to even approximately ensure $\int_{\mathcal{Y}} \tilde{p}(y_t = y|x)\,\mathrm{d}y = 1$. For this reason the bias of an estimator without the corrective denominator of Equation 11 would be uncontrollable [Tokdar and Kass, 2010]. A greedy algorithm may be deployed to maximize $\tilde{\mathbb{E}}[f(Y_t)|X = x]$ in the form above by optimizing weights w_i attached to each $f(y_i)$, within the range

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

The minimum may be achieved by a trivial adaptation. Maximizing and minimizing $\tilde{\mathbb{E}}[f(Y_t)|X = x]$ with respect to the bounding quantities (γ, Λ) enables the resolution of ignorance bounds on the basis of Γ from Definition 2. Our Algorithm 1 adapts the method of Jesson et al. [2021] to heterogeneous weight bounds $[\underline{w}_i, \overline{w}_i]$ per draw i.

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 $\Delta_j \triangleq \sum_{i=1}^{n} w_i(f_j - f_i)$;
  if $\Delta_j < 0$ **then**
    $w_j \leftarrow \underline{w}_j$;
  else
    break;
  end
end
Return $\sum_i w_i f_i / \sum_i w_i$;

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

D Experimental details

Data from the UK Biobank were accessed under application 11559. From the brain Magnetic Resonance Imaging (MRI) data we extracted the 74 fields corresponding to parcelized cortical volumes on the left and the right hemispheres each [Miller et al., 2016].

Semi-synthetic evaluation. In our first experiment, the synthetic binary outcome was generated by linearly combining the covariates and treatment and then applying a logistic curve for a Bernoulli probability. As the logit-transformed ATE was known to be linear, "recall" was evaluated as the portion of the ignorance intervals that permitted the actual linear effect along each section of the dose response.

Semi-synthetic dataset. The 148 MRI fields were normalized such that values floored at each variable's 25% quantile and ceiled at the 95% quantile were mapped to the range $[0, 1]$. In each of the four replications, a random quarter of the covariates were assigned a nonzero, normally distributed coefficient, and one of them was deemed the treatment variable with unit coefficient. After computing the outcomes, we randomly resampled a third of the feature values, with replacement, in order to introduce noise to the covariates while preserving the marginals.

Semi-synthetic estimators. Both the predictor and the propensity model, censored and uncensored, were trained in ensembles of 32 artificial neural networks with one inner residual layers of 32 activation units each. A dropout of 0.05 was imposed on these layers. Additionally, an L^2-regularization on the inner layers with weight 10^{-3} was applied. All predictors were trained for 10,000 epochs and propensities for 5,000 epochs.

Empirical evaluation. Our test relies on approximating an unconfounded model by collecting a large set of covariates, and then learning another model on a heavily censored version. Our reasoning is that the censored model would suffer from a greater degree of hidden confounding. The censored model could then be assessed along all potential-outcome predictions, by pretending that the full model represented the real dose-response curves. A pertinent metric would be how much a sensitivity model with $\Gamma \geq 1$ swallows the "real" dose responses, as a trade-off against the sheer area of the ignorance bounds. These competing quantities are a form of recall and (the opposite of) precision, respectively.

Denote $(\underline{y_c}, \overline{y_c})$ the partially identified bounds of the censored model and $(\underline{y}, \overline{y})$ the full model's bounds for $\tilde{\mathbb{E}}[Y_t]$, both at 95% confidence from percentile bootstrapping. For some $\Gamma \coloneqq s$ and a $t \in [0, 1]$ grid, of length 17 in our case, **ignorance** is $\sum_t [\overline{y_c}(s,t) - \underline{y_c}(s,t)] / \sum_t [\overline{y}(t) - \underline{y}(t)]$, and **recall** is the normalized intersection between the bounds:

$$\frac{\sum_t \max\{0, \min\{\overline{y_c}(s,t), \overline{y}(t)\} - \max\{\underline{y_c}(s,t), \underline{y}(t)\}\}}{\sum_t [\overline{y}(t) - \underline{y}(t)]}.$$

Empirical dataset. We summed each left/right pair to arrive at 74 positively valued outcomes. Six groups of semantically related fields composed the long covariate vector:

- 6 basic details: age, weight, sex, standing height, seated height, and month of birth.
- 3 reported activity measurements: weekly minutes spent walking, engaged in moderate activity, and vigorous activity.
- 27 *environment*al variables surveying the pollution and greenery surrounding the person's life.
- 42 *blood* measurements from cell counts to calcium concentration.
- 15 *cardiac* measurements including ECG and PWA modalities.
- 8 *welfare* indices for English citizens assessing the following: deprivation, income, employment, health, education, housing, crime, and living environment.

Listwise deletion was employed to handle any missing value. To censor the covariates, the four largest sectors (*italicized*) encompassing various confounding variables were omitted, one at a time. The treatment variable was the walking field taken from the triad of activity measurements, scaled to the unit interval such that any recording of at least two hours per day was set to $T = 0$ and any lesser amount had T increase up to 1 according to an empirical CDF.

Empirical estimators. Both the predictor and the propensity model, censored and uncensored, were trained in ensembles of 32 artificial neural networks with four inner residual layers of 32 activation units each. A dropout of 0.05 was imposed on these layers. Additionally, an L^2-regularization on the inner layers with weight 10^{-3} was applied. All predictors were trained for 10,000 epochs and propensities for 5,000 epochs. The outcome predictor parametrized a Gamma distribution, and the propensity model parametrized a Beta distribution. See Figure 6.

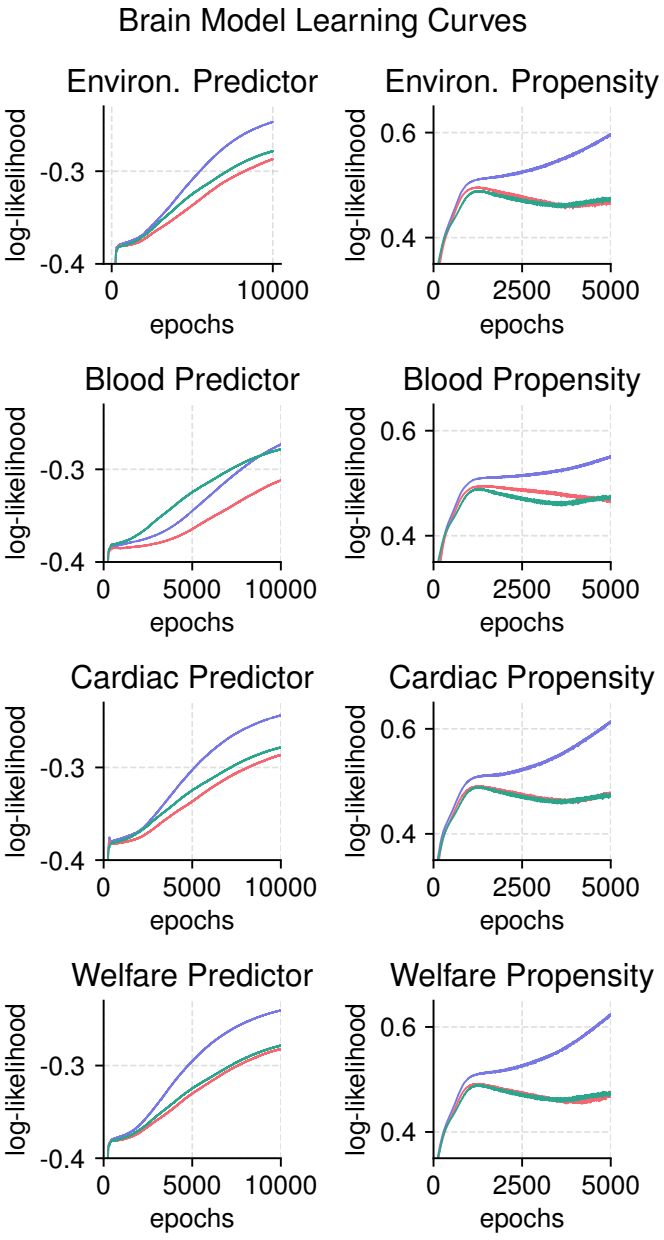

Figure 6. **Blue**: censored-model likelihood in the train set; **red**: censored-model likelihood in the test set; and **green**: full-model likelihood in the test set.