Conditional differential measurement error: partial identifiability and estimation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Differential measurement error, which occurs when the level of error in the measured outcome is correlated with the treatment renders the causal effect unidentifiable from observational data. We study conditional differential measurement error, where a subgroup of the population is known to be prone to differential measurement error. Under an assumption about the direction (but not magnitude) of the measurement error, we derive sharp bounds on the conditional average treatment effect, and present an approach to estimate them. We empirically validate our approach on semi-synthetic and real data, showing that it gives more credible and informative bound than other approaches.

1 Introduction

One of the promises of estimating individual level causal effects of an intervention is that it can be used to guide decision making in the real world. However, complexities in real world data pose a challenge to the validity of causal analysis. Much of the recent work in causality has focused on addressing these challenges including hidden confounding, limited overlap, inefficiency due to small samples, among others [4, 15, 13, 18, 26, 12].

Differential measurement error, which occurs when the level of error in the measured outcome is correlated with the treatment assignment, is one challenge that has not received much attention. In the presence of differential measurement error, the causal effect is not identifiable from observational data [5, 24]. In this paper, we study causal estimation methods in the presence of conditional differential measurement error, where one subgroup's outcomes are measured with differential error. One example we use throughout the paper, is estimating the causal effect of lifestyle interventions that aim to change the participants' dietary intake or physical exercise. These studies are prone to differential measurement error due to the nature of the intervention itself: participants assigned to the intervention arm might under/over-report dietary intake in order to appear compliant [16]. In particular, older participants are prone to misreporting their dietary intake post intervention [19]. For this group, and without additional assumptions, it is impossible to disentangle the true causal effect of dietary modification on nutritional intake from the differential error in measuring the outcome.

Instead of attempting to fully identify the causal effect for the mismeasured subgroup, we study a partial identification approach. We show that it is possible to estimate sharp, informative bounds on the causal effect when the direction of the measurement error matches the direction of causal effect modification due to membership in the mismeasured group. Here, our assumption of 'matched directions' represents the case where the dietary intervention is more effective among older participants, and the measurement error leads to an inflated estimate of the causal effect among older participants. We present a novel approach for estimating bounds on the conditional average treatment effects (CATE) and show that our bounds are sharp, meaning they cannot be improved without additional assumptions. Our contributions are summarized as follows: (1) We characterize the identification

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

region of the CATE for the mismeasured subgroup. (2) We develop an efficient approach for estimating bounds on CATE. (3) We empirically test our approach on simulated data and real data from the Women's Health Initiative (WHI) [11], and show that our approach gives more reliable estimates than other approaches.

2 Related work

The majority of work on causality in the machine learning community studies the estimation of causal effects when the common assumptions are satisfied, meaning when the CATE is identifiable [1, 8, 14, 3, 18]. Work studying settings where the CATE is not identifiable focuses on lack of identifiability due to violations of the common assumptions like conditional ignorability and overlap [25, 27, 7, 28].

Relative to work on violation of the common assumptions, work on estimating the CATE in the presence of differential measurement error in the outcome is limited. Existing work commonly makes assumptions about the noise model, which are likely to be violated in practice. For example, Díaz and van der Laan [5] give a sensitivity analysis for differential measurement error of either the exposure or outcome. Their approach requires knowledge of a plausible range of the magnitude of differential measurement error, which might be unrealistic in many cases. By contrast, we assume knowledge about the direction rather than the magnitude of differential measurement error is known.

In different work, VanderWeele and Li [24] outlines a strategy to estimate bounds on the causal effect by relying on the boundedness of the observed outcomes. As our empirical analysis shows, this type of approach can give uninformative bound. [23] studies the impact of measurement error for continuous variable outcome. They assume that a small proportion of the population might have measurement error in the outcome. By contrast, our approach allows for an arbitrarily large population to be measured with error, at the cost of requiring knowledge about which subpopulation is mismeasured. We stress however, that this knowlegde is not required at training time. In addition, in contrast to Shu and Yi [23], our approach can accommodate arbitrary types of outcomes (e.g., binary, categorical, etc.)

3 Preliminaries

We follow the convention of using capital letters to denote variables and small letters to denote their values. We assume access to an observational dataset that consists of tuples of random variables $\mathcal{D}\{(x_i, z_i, t_i, y_i) : i = 1, \ldots, n\}$. $x_i \in \mathbb{R}^d$ is a d-dimensional feature vector, $t_i \in \{0, 1\}$ is the treatment assignment indicator, and z_i reflects membership in the subpopulation that is suspected to have mismeasurement error. For simplicity, we will assume that the mismeasured group is defined by $Z = z'$, whereas the non-mismeasured group is defined by $Z \neq z'$. We use $Y(1), Y(0)$ to denote the potential outcome under treatment $t = 0, 1$, respectively. We define the conditional average treatment effect (CATE) defined as:

$$\tau(X, Z) = \mathbb{E}[Y(1) - Y(0) \mid X, Z], \tag{1}$$

where the expectation is taken with respect to the full unobserved distribution.

In addition, we define $\theta_t(X, Z)$ to be the measurement error of the observed outcome for $T = t$, and $f(X, Z)$ to be the magnitude of the differential measurement error, with $f(X, Z) = \theta_1(X, Z) - \theta_0(X, Z)$. In addition, we define $g(X, Z)$ to be the magnitude of the effect modification due to Z, with $g(X, Z) = \tau(X, Z = z) - \tau(X, Z = z')$. Finally, we use $h(X, Z) = g(X, Z) + f(X, Z)$ to denote the *nominal* causal effect modifiction due to Z. We stress that $h(X, Z) \neq g(X, Z)$ due to the differential measurement error. The observed outcome is defined as:

$$y = t \cdot \big(Y(1) + \theta_1(X, Z)\big) + (1 - t) \cdot \big(Y(0) + \theta_0(X, Z)\big). \tag{2}$$

with $\theta_0(X, Z = z) = \theta_1(X, Z = z) = 0$. In this setting, CATE is identifiable for $Z = z$ but not for $Z = z'$ [5]. We define the nominal CATE for $Z = z'$ as τ'

$$\tau'(X, Z = z') = \mathbb{E}[Y(1) - Y(0) + f(X, Z = z') \mid X, Z = z']. \tag{3}$$

Our goal is to give a point estimate of the CATE for $Z = z$, that is $\hat{\tau}(x, Z = z)$ because the estimand is identifiable among the subpopulation defined by $Z = z$. For the subpopulation defined by $Z = z'$, our goal is to give upper and lower bounds on the CATE, since their CATE is not identifiable due to differential measurement error. Specifically, for the population defined by $Z = z'$,

we aim to estimate a lower and upper bound defined as $\hat{\underline{\tau}}(x, z')$, $\hat{\overline{\tau}}(x, z')$, respectively such that $\tau(x, z) \in [\hat{\underline{\tau}}(x, z'), \hat{\overline{\tau}}(x, z')]$ with high probability.

3.1 Assumptions

Consistent with the majority of causal literature (e.g., Shu and Yi [23], Imai and Yamamoto [12], Shalit et al. [22]), we assume conditional ignorability. In addition, we assume a modified version of the typical overlap assumption, stated below.

Assumption 3.1. *(Z, T)-Overlap. We assume that: $0 < Pr(T|X, Z) < 1$ for $T \in [0, 1]$ and $0 < Pr(Z|X, T) < 1$ for $Z \in [z, z']$*

(Z, T)-Overlap is a modification of the typical overlap assumption in that it requires overlap with respect to the mismeasured group in addition to the treatment assignment. We make the following two assumptions:

Assumption 3.2. *The mismeasured group (i.e., the value of z') is given.*

We note that z' need not be known at estimation or training time, as long as assumption 3.1 is satisfied in the training data. Our proposed approach (which assumes that z' is known at estimation time) is trivially extendable to the setting where z' is given after estimation. We discuss this in section4.1.

We assume that the direction of effect modification by Z matches that of the measurement error:

Assumption 3.3. *We assume that the direction of $f(X, Z)$ and $g(X, Z)$ matches, i.e., that: $g(X, Z) \cdot f(X, Z) \geq 0$*

We stress, however, that assumption 3.3 is an assumption on the direction but not the magnitude of the effect modifiction or the measurement error. This makes it a less stringent assumption compared to other work in differential measurement error (e.g.., in Díaz and van der Laan [5]).

In our motivating example, this assumption would be satisfied if (1) the dietary modification intervention is more effective among older women (the group defined by $Z = z'$) compared to younger women, and (2) older women under-report their dietary intake leading to an inflated estimate of the causal effect. Therefore, both measurement error and effect modification act in the positive direction of the causal effect.

4 Characterizing and estimating bounds on CATE

In this section, we characterize the identification region of CATE for the subpopulation with differential measurement error, and present an approach to estimate the upper and lower bounds that define the identification region.

The identification region represents the range of all possible estimates of CATE that are consistent with the observed data and the assumptions outlined in section 3.1. We characterize that region in the following proposition.

Proposition 4.1. ***(Sharp bounds on CATE)*** *Suppose that assumptions 3.1, 3.2, 3.3 and conditional ignorability hold, further suppose without loss of generality that $f(X, Z) \geq 0, \forall X$, then*

$$\tau(X, Z = z') \in [\tau(X, Z = z), \tau(X, Z = z) + h(X, Z)]$$

and this bound is sharp.

The proof is presented in Appendix. Proposition 4.1 shows that the CATE for the mismeasured group can be bounded above and below by $\tau(X, Z = z)$, which is the CATE of the non-mismeasured group, and $\tau'(X, z') = \tau(X, Z = z) + h(X, Z)$, which is the nominal CATE of the mismeasured group. Such a finding is important since both bounds can be estimated from observed data. In addition, the proposition shows that these bounds are sharp, in that they cannot be improved without further assumptions.

4.1 Implementation

Following the theoretical results, we propose an approach to estimate CATE for $Z = z$ and bounds on CATE for $Z = z'$. Our approach proceeds by estimating $\tau(X, Z = z)$ as is done in typical settings

where there is no mismeasurement error (e.g., using G-estimation or doubly robust approaches). This estimate represents the final estimate for the group defined by $Z = z$ and one of the two bounds for the mismeasured group defined by $Z = z'$. As proposition 4.1 shows, the other bound is the nominal causal effect for the mismeasured group, so it can be estimated from the data using the typical causal estimation methods. Below, we outline two possible approaches to obtain point estimates for the two groups defined by $Z = z$ and $Z = z'$.

G-estimation approach. Using any nonparametric estimator $\mathcal{M}$, we estimate the expected potential, $\mu(x, Z, t) = \mathbb{E}_{\mathcal{D}}[y|X = x, Z, T = t], Z \in \{z, z'\}$, where the expectation is taken with respect to the observed data, $\mathcal{D}$ [9]. For $Z = z$, the estimate of the CATE is computed as $\hat{\tau}(X, Z = z) = \mu(x, Z = z, t = 1) - \mu(x, Z = z, t = 0)$, which is also an estimate for one of the two bounds on $\tau(X, Z = z')$.

For $Z = z'$, as we show in proposition 4.1, $\hat{\tau}(X, Z = z)$ represents one of the two bounds on $\tau(X, Z = z')$, the other bound is the nominal causal effect in the $Z = z'$ subpopulation, which can be similiarly estimated: $\hat{\tau}'(x, z') = \mu(x, Z = z', t = 1) - \mu(x, Z = z', t = 0)$. The final estimated bounds on the unindentifiable $\tau(X, Z = z')$ are then:

$$\underline{\hat{\tau}}(x, z') = \min\{\hat{\tau}(x, z), \hat{\tau}'(x, z')\}, \text{ and } \hat{\overline{\tau}}(x, z') = \max\{\hat{\tau}(x, z), \hat{\tau}'(x, z')\}.$$

For a well specified function class, $\hat{\tau}(x, z), \hat{\tau}'(x, z')$ are guaranteed to asymptotically converge to $\tau(x, z), \tau'(x, z')$ [21].

In our implementation, we estimate $\mu(X, Z, T)$ using 4 different models, one for each combination of T, Z. This is similar in spirit to T-Learners [17]. In settings where the mismeasured group is not known at estimation or training time, training 4 different models is not possible. In that case, it is still possible to estimate bounds on $\tau(X, Z = z')$ by estimating 2 models, one for each treatment group. As long as the function space is well specified and assumption 3.1 is satisfied, this estimation procedure should give asymptotically unbiased estimates.

Doubly-robust approach. We outline how a doubly-robust approach can be used to estimate the bounds. Estimation proceeds similar to G-estimation, but during estimation, we inverse weight each data point by its propensity to recieve the treatment t, and its propensity to belong in the mismeausred group. The weighting scheme reweights the observed data using the Radon-Nikodym derivative of the distribution where treatment assignment and membership in the mismeasured group is uncorrelated with X with respect to the observed distribution, i.e., $w_i = \frac{p(x_i)p(z_i)p(t_i)}{p(x_i, z_i, t_i)}$. We use permutation weighting to compute the weights [2]. Details of the implementation are included in the appendix.

5 Experiment

We evaluate our algorithm on a semi-synthetic and a real dataset. We use ACIC data [6] for the former, and data from the dietary modification arm of the WHI [11] for the latter[1]. Additional details about the implementation are included in the appendix.

Baselines. We compare the two variants of our approach, the G-Estimation approach — **GE (ours)**— and the Doubly Robust approach — **DR (ours)**, to the following baselines: (1) Quantile regression model — **GE (QR)**, which uses quantile estimates to construct the upper and lower bounds of the CATE for the subpopulation defined by $Z = z'$. Specifically, the QR proceeds by estimating the 0.1 and 0.9-quantile of $\mathbb{E}[Y|X = x, T = t]$, for $t \in \{0, 1\}$. The upper (lower) bound of $\tau(x, z')$ is given by the difference between 0.9 (0.1)-quantile of $\mathbb{E}[Y|X = x, T = 1]$ and 0.1 (0.9)-quantile of $\mathbb{E}[Y|X = x, T = 0]$. (2) Quantile regression model with doubly robust approach – **DR (QR)** – is similar to the QR model, but in addition, we utilize inverse propensity score weights to estimate $\mu(x, z, t)$ as described in 4.1. (3) Simple Sensitivity Analysis (**SSA**), which is inspired by previous work on differential measurement error [24]. The SSA model finds the maximum and minimum observed outcome for each treatment T. Then the upper (lower) bound is given by difference between maximum (minimum) value of $y_{t=1}$ and minimum (maximum) value of $y_{t=0}$. (4) **Oracle** model, which is an unattainable version of our GE approach that has oracle estimates of $\mu(x, z, t)$.

[1]WHI data is available for use, upon making a request for data access to the National Heart, Lung and Blood Institute at `biolincc.nhlbi.nih.gov`

Evaluation We evaluate algorithms with three different measures. We evaluate credibility with correct coverage rate (**CCR**) and **deviation**. CCR is defined as the proportion of the test sample that the true causal effect is correctly bound by our method, $CCR = n^{-1} \sum_i \tau \in [\underline{\tau}, \overline{\tau}]$. CCR measures how well algorithms correctly cover true CATE. Deviation is defined as the average distance from false coverage to our predicted bound, deviation$= m^{-1} \sum_{i|\tau \notin [\underline{\tau}, \overline{\tau}]} \min\{|\tau - \overline{\tau}|, \tau - \underline{\tau}\}, m = |\{i|i|\tau \notin [\underline{\tau}, \overline{\tau}]\}|$. Deviation measures how close the false coverage lies to true CATE. We evaluate informativeness by **tightness**, which is defined as the average length of the bound, Tightness $= n^{-1} \sum_i |\underline{\tau} - \overline{\tau}|$.

Semi-synthetic data. The ACIC dataset Dorie et al. [6] consists of 4802 observations and 58 covariates. We randomly select 70% of the data as the training sample and the remaining 30% as the testing sample. The original ACIC data did not include measurement errors, so we created conditional differential errors as follows. We select the mismeasured group based on randomly selected variables that were not effect modifiers in the original dataset. The simulated observed outcome is defined as

$$y = \tilde{y} + \mathbb{1}\{Z = z'\} \cdot (t(\kappa_1(X, Z)) + (1 - t)\kappa_0(X, Z)) \tag{4}$$

where $\tilde{y}$ is the original simulation's outcome, $\kappa_t(X, Z)$ denotes the total nominal effect of Z on the treatment t groups. We use $0 \leq \alpha < 1$ to denote the proportion of the nominal effect due to effect modification versus differential measurement error and update CATE as

$$\tilde{\tau}(x, z') = \tau(x, z') + \alpha(\kappa_1(X, Z) - \kappa_0(X, Z)) \tag{5}$$

We evaluate all models' performance at varying levels of differential measurement error by varying α, going from high measurement error ($\alpha = 0$), to low ($\alpha = 1$).

Figure1 and table 1 show the results of this simulation setting. The $x-$axis of the two plots in figure 1 shows the level of the nominal CATE attributable to measurement error. In figure 1(left), the $y-$axis shows the CCR while in figure 1(right), the $y-$axis shows the deviance. Figure 1(left) demonstrates that our method has a higher coverage rate than QR when the level of differential measurement error is high. At low levels of differential measurement error, QR is able to correctly cover a higher proportion of the population compared to our approach, however, as figure 1(right) reveals, our approach maintains a lower deviance compared to QR. This happens because instances that are incorrectly covered by our approach are "barely missed", which is not true for QR. The oracle model, which achieves perfect CCR, deviation and tightness, demonstrates the unattainable performance of our approach if it did not incur any estimation error. This implies that with a larger dataset, our approach could have achieved near perfect performance. The doubly robust approach has a slightly lower CCR than the G-estimation approach since it provides a bound that is even tighter than the oracle method. Finally, SSA achieves perfect CCR but as table 1 reveals, this comes at the cost of tightness: SSA gives credible yet uninformative bounds.

Algorithm	Tightness (STD)
GE (ours)	9.04 (1.28)
DR (ours)	8.84 (1.79)
GE (QR)	10.82 (1.47)
DR (QR)	7.87 (2.22)
Oracle	9.0 (0)
SSA	65.94 (1.04)

Table 1: Results on the ACIC data, showing the absolute width and standard deviation of the estimated bounds for the mismeasured group. SSA gives uninformative bounds, while our approach gives estimates that are close to the oracle.

Real data. Due to the fundemental problem of causal inference [20, 10], it is impossible to adequately evaluate our approach in a real data setting. Still, we conduct this analysis on the WHI data to demonstrate the utility of our approach.

Specifically, we study the causal effect of the dietary modification intervention in the WHI [11]. In this study, 19,541 women were randomly assigned to the intervention arm, where they received group and individual sessions to promote a decrease in fat intake and increase in vegetable consumption. An additional 29,294 were assigned to the non-intervention arm where they received diet-related education materials. By design, this randomized control trial did not have a confounded treatment assignment. To create a more challenging task, we create confounding by downsampling the women with the lowest quantile BMI. We randomly removed 50% of the women in the lowest percentile of BMI to mimic a setting where women with a lower BMI are less likely to attend dietary modification. In addition, we considered the mismeasured group to be women who are in the highest quantile of age.

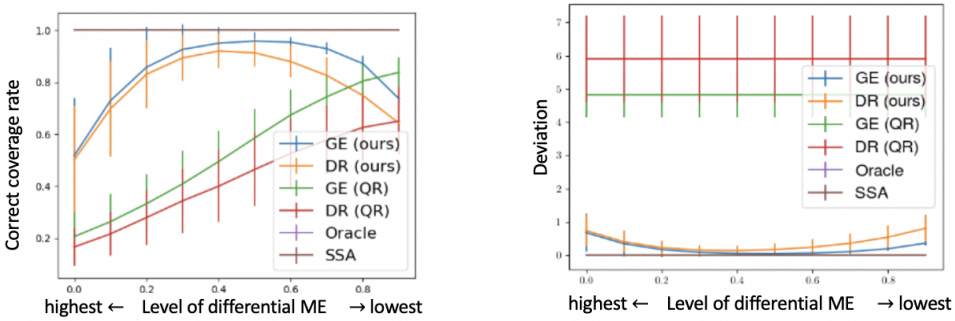

Figure 1: $x-$axis shows the level of the nominal CATE attributable to measurement error. (Left) Plot shows that correct coverage rate ($y-$axis) is higher (i.e., better) for our approach compared to QR algorithm when the level of measurement error is high. The unattainable oracle has a perfect correct coverage rate, revealing that our approach can achieve near perfect coverage if not for estimation error. (Right) Plot shows that deviation for our approach is small compared to others.

This is consistent with existing literature suggesting that older age is correlated with a higher tendency to misreport outcomes [19]. We choose women with the last quantile age as our mismeasured group. We apply our approach and the baselines as outlined in the ACIC experiment.

Because we do not have access to the counterfactual outcomes, we cannot evaluate our approach using the CCR and deviation. Instead, we evaluate how informative the estimated bounds are for guiding intervention decisions among older women. We consider intervention as effective if both predicted upper and lower bounds are greater than 0 ($\overline{\tau} \geq 0$ and $\underline{\tau} \geq 0$), obscure if predicted upper and lower bounds have different signs ($\overline{\tau} \cdot \underline{\tau} < 0$), and ineffective otherwise ($\overline{\tau} < 0, \underline{\tau} < 0$). We show the percentage of each predicted group as follows

Algorithm	Sodium		Fiber		BMI	
	Effective	Uncertain	Effective	Uncertain	Effective	Uncertain
GE (ours)	43.03	49.38	0.01	6.64	1.05	27.53
DR (ours)	39.78	48.50	0.25	6.14	3.24	31.54
GE (QR)	0	100	0.4	99.6	0.16	99.84
DR (QR)	0	99.98	0	100	0	100
SSA	0	100	0	100	0	100

Table 2: Results on the WHI data. SSA and QR give a higher level of uncertainty compared to our approach.

Table 2 shows that the SSA and QR approach predict almost every intervention as uncertain, while our approach predicts the intervention is effective for decreasing the Sodium intake for 40% of the population; ineffective for decreasing BMI for about 70% of the population and ineffective for increasing intake of Fiber.

6 Conclusion

In the paper, we provide a credible and informative bound for CATE by making an assumption on the direction but not the magnitude of the measurement error. Guided by our theoretical analysis, we presented an approach to estimate sharp bounds on the causal effects, and showed settings where it outperforms alternative approaches. On a real dataset, we showed that our approach leads to less uncertainty about intervention decisions when compared to other approaches.

Empirical results show that one limitation of our approach is that estimation error might lead to lower coverage rates when the measurement error is high. Future work can focus on more efficient estimation methods to address this issue. In addition, similar to all other estimation approaches, our approach might give unreliable estimates when the assumptions outlined in section3 are violated. Such unreliable estimates could lead to incorrect treatment decisions with negative societal impact. Future work can focus on developing methods to quantify the sensitivity of our approach to different assumptions made in section3.

References

[1] J. Abrevaya, Y.-C. Hsu, and R. P. Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

[2] D. Arbour, D. Dimmery, and A. Sondhi. Permutation weighting. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 331–341. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/arbour21a.html`.

[3] A. Belloni, V. Chernozhukov, I. Fernández-Val, and C. Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

[4] M. Bonvini and E. H. Kennedy. Sensitivity analysis via the proportion of unmeasured confounding. *Journal of the American Statistical Association*, pages 1–11, 2021.

[5] I. Díaz and M. J. van der Laan. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *The international journal of biostatistics*, 9(2): 149–160, 2013.

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

[7] A. D'Amour, P. Ding, A. Feller, L. Lei, and J. Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

[8] D. P. Green and H. L. Kern. Modeling heterogeneous treatment effects in large-scale experiments using bayesian additive regression trees. In *The annual summer meeting of the society of political methodology*, pages 100–110, 2010.

[9] M. A. Hernán and J. M. Robins. Causal inference, 2010.

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

[11] B. V. Howard, J. E. Manson, M. L. Stefanick, S. A. Beresford, G. Frank, B. Jones, R. J. Rodabough, L. Snetselaar, C. Thomson, L. Tinker, et al. Low-fat dietary pattern and weight change over 7 years: the women's health initiative dietary modification trial. *Jama*, 295(1): 39–49, 2006.

[12] K. Imai and T. Yamamoto. Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis. *American Journal of Political Science*, 54(2):543–560, 2010.

[13] A. Jesson, S. Mindermann, Y. Gal, and U. Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *International Conference on Machine Learning*, pages 4829–4838. PMLR, 2021.

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

[15] N. Kallus, X. Mao, and A. Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2281–2290. PMLR, 16–18 Apr 2019. URL `https://proceedings.mlr.press/v89/kallus19a.html`.

[16] S. I. Kirkpatrick, C. E. Collins, R. H. Keogh, S. M. Krebs-Smith, M. L. Neuhouser, and A. Wallace. Assessing dietary outcomes in intervention studies: pitfalls, strategies, and research needs. *Nutrients*, 10(8):1001, 2018.

[17] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116 (10):4156–4165, 2019.

[18] M. Makar, F. Johansson, J. Guttag, and D. Sontag. Estimation of bounds on potential outcomes for decision making. In *International Conference on Machine Learning*, pages 6661–6671. PMLR, 2020.

[19] K. Pfrimer, M. Vilela, C. M. Resende, F. B. Scagliusi, J. S. Marchini, N. K. Lima, J. C. Moriguti, and E. Ferriolli. Under-reporting of food intake and body fatness in independent older people: a doubly labelled water study. *Age and ageing*, 44(1):103–108, 2014.

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

[21] B. Schölkopf, A. J. Smola, F. Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[22] U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[23] D. Shu and G. Y. Yi. Causal inference with measurement error in outcomes: Bias analysis and estimation methods. *Statistical methods in medical research*, 28(7):2049–2068, 2019.

[24] T. J. VanderWeele and Y. Li. Simple sensitivity analysis for differential measurement error. *American journal of epidemiology*, 188(10):1823–1829, 2019.

[25] J. M. Wooldridge. Violating ignorability of treatment by controlling for too many factors. *Econometric Theory*, 21(5):1026–1028, 2005.

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

[27] Y. Zhu, R. A. Hubbard, J. Chubak, J. Roy, and N. Mitra. Core concepts in pharmacoepidemiology: Violations of the positivity assumption in the causal analysis of observational data: Consequences and statistical approaches. *Pharmacoepidemiology and drug safety*, 30(11): 1471–1485, 2021.

[28] Y. Zhu, N. Mitra, and J. Roy. Addressing positivity violations in causal effect estimation using gaussian process priors. *arXiv preprint arXiv:2110.10266*, 2021.

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]
   (b) Did you describe the limitations of your work? [Yes] See section 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See section 6
   (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 section 3.1
   (b) Did you include complete proofs of all theoretical results? [Yes] See the Appendix, section A.1
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)? [No] Our method is easily reproducible using existing packages. In addition, to facilitate reproducibility, we included psuedocode in the appendix, section A.2
   (b) Did you specify all the training details method.g., data splits, hyperparameters, how they were chosen)? [Yes] See section 5 in the main text and section A.2 in the appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] For the simulated experiments only, see 1 and table 1
   (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)? [Yes] See section A.2 in the appendix.

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]
   (b) Did you mention the license of the assets? [Yes] Both datasets are publicly available. We included information about gaining access to the WHI data in section 5
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Both datasets are completely de-identinfied and available publicly.
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 Appendix

A.1 Proof of proposition 4.1

Proposition A.1. *(**Sharp bounds on CATE**, restated proposition 4.1 from the main text) Suppose that assumptions 3.1, 3.2, 3.3 and conditional ignorability hold, further suppose without loss of generality that $f(X, Z) \geq 0, \forall X$, then*

$$\tau(X, Z = z') \in [\tau(X, Z = z), \tau(X, Z = z) + h(X, Z)]$$

and this bound is sharp.

Proof. Note that by definition:

$$g(X, Z) = h(X, Z) - f(X, Z). \tag{6}$$

By assumption 3.3, and that $f(X, Z) \geq 0$, we have that all three terms, $g(X, Z), h(X, Z), f(X, Z) \geq 0$, which in turn implies that $h(X, Z) \geq f(X, Z)$, with equality only holding in the case where there is no differential measurement error, i.e., $f(X, Z) = 0$. This means that $f(X, Z) \in [0, h(X, Z)]$, and we can rewrite $f(X, Z) := \beta h(X, Z)$, where β is an unidentifiable parameter with taking values between 0, 1. Substituting in equation 6, we get that:

$$g(X, Z) = h(X, Z) - \beta h(X, Z) = (1 - \beta)h(X, Z) := \alpha h(X, Z),$$

where $\alpha \in [0, 1]$ is unidentifiable from observational data.

We can now rewrite the unidentifiable CATE in terms of components that can be estimated from observational data, as well as this unidentifiable but bounded α:

$$\tau(X, Z = z') = \tau(X, Z = z) + g(X, Z) = \tau(X, Z = z) + \alpha h(X, Z).$$

Recall that $\tau(X, Z = z)$ is identifiable since it is simply the CATE for the group that is not mismeasured, $h(X, Z)$ is the nominal CATE which is also identifiable from observational data, and α is bounded between 0, and 1. This allows us to define the identification region as the region defined by extreme values of α, as follows:

$$\min_{\alpha \in [0,1]} \tau(X, Z = z) + \alpha h(X, Z)$$

and

$$\max_{\alpha \in [0,1]} \tau(X, Z = z) + \alpha h(X, Z).$$

By assumption, $h(X, Z) \geq 0$, so it obtains its maximum when $\alpha = 1$ and its minimum when $\alpha = 0$, which means that:

$$\tau(X, Z = z') \in [\tau(X, Z = z), \tau(X, Z = z) + h(X, Z)]$$

Note that these bounds are sharp in that the true $\tau(X, Z = z')$ is equal to the upper bound if there is no differential measurement error (i.e., $f(X, Z) = 0$ making $h(X, Z) = g(X, Z)$), and the true $\tau(X, Z = z')$ is equal to the lower bound if there is no effect modification (i.e., $g(X, Z) = 0$). Hence, the bounds could not be made tighter without additional assumptions. □

A.2 Implementation details

Permutation weighting. Permutation weighting proceeds by making a copy of the original data set, and permuting Z and T randomly in the copied data. Points in the original data are given a label $C = 0$ while the permutated points are given a label $C = 1$. Denote $\mathcal{P}$ as the probability function. We stack the two datasets together and estimate $w_i = \frac{\mathcal{P}(C=1|X=x,Z=z,T=t)}{\mathcal{P}(C=0|X=x,Z=z,T=t)}$.

Algorithm psuedocode The pseudocode for our G-estimation approach is listed as follows:

Algorithm 1

Input: Factual sample $(x_1, t_1, z_1, y_1), \ldots, (x_n, t_n, z_n, y_n)$, non-parametric estimator $\mathcal{M}$

1: $\mu_{t,z} = \mathcal{M}(Y_{t,z} \sim X_{t,z})$
2: $\mu_{t,z'} = \mathcal{M}(Y_{t,z'} \sim X_{t,z'})$
3: $\hat{\tau}_0(x) = \mu_{t=1,z}(x) - \mu_{t=0,z}(x)$
4: $\hat{\tau}_1(x) = \mu_{t=1,z'}(x) - \mu_{t=0,z'}(x)$

Output: $\hat{\underline{\tau}}(x) = \min\{\hat{\tau}_0(x), \hat{\tau}_1(x)\}, \hat{\overline{\tau}}(x) = \max\{\hat{\tau}_0(x), \hat{\tau}_1(x)\}$

Training details Both experiments were implemented on CPUs, with a total compute time of 50 hours. For both experiments, we use logistic regression to compute permutation weighting. We use Random Forest Regressor to train the models and Grid Search Cross Validation to select parameters. We compare parameters $\{1e-3, 1e-2, 1e-1, 1, 1e1, 1e2, 1e3\}$ using R^2 coefficient of determination as regression score function.