Disentangling Misreporting from Genuine Adaptation in Strategic Settings: A Causal Approach

Dylan Zapzalka[1] **Trenton Chang**[1] **Lindsay Warrenburg**[2] **Sae-Hwan Park**[2]
Daniel K. Shenfeld[2] **Ravi B. Parikh**[2,3] **Jenna Wiens**[1] **Maggie Makar**[1]
[1]University of Michigan [2]University of Pennsylvania [3]Emory University
{dylanz, mmakar}@umich.edu

Abstract

In settings where ML models are used to inform the allocation of resources, agents affected by the allocation decisions might have an incentive to strategically change their features to secure better outcomes. While prior work has studied strategic responses broadly, disentangling misreporting from genuine adaptation remains a fundamental challenge. In this paper, we propose a causally-motivated approach to identify and quantify how much an agent misreports on average by distinguishing deceptive changes in their features from genuine adaptation. Our key insight is that, unlike genuine adaptation, misreported features do not causally affect downstream variables (i.e., causal descendants). We exploit this asymmetry by comparing the causal effect of misreported features on their causal descendants as derived from manipulated datasets against those from unmanipulated datasets to identify the misreporting rate. We empirically validate our theoretical results using a semi-synthetic and real Medicare dataset with misreported data, demonstrating that our approach can be employed to identify misreporting in real-world scenarios.

1 Introduction

Machine learning models are increasingly used by decision-makers to guide decisions about the allocation of critical resources, such as in loan applications, or determining government payouts to private insurers [1, 33, 9]. In these contexts, organizations—referred to as agents—may have an incentive to strategically change their features to secure better outcomes [20]. They can do so through genuine adaptation or misreporting. *Genuine adaptation* refers to agents genuinely changing their behavior, causing the actual values of their features to change. This leads to real improvements and authentic changes. Misreporting refers to agents not changing their behavior but instead reporting incorrect feature values to manipulate the allocation process. Genuine adaptation may be desirable as it can lead to improvements in the target outcome [28, 12]. Misreporting, however, is never desirable to the decision-maker as it leads to incorrect predictions and inefficient resource allocation.

In this work, we develop a causal framework for detecting and quantifying misreporting in the presence of genuine adaptation. Our key insight is that misreporting, unlike genuine adaptation, does not causally affect the descendants of a given feature. Consequently, misreporting leads to biased causal effect estimates between the feature and its descendants. We exploit this asymmetry by comparing the estimated causal effect of a feature on its descendants in both manipulated and unmanipulated datasets to infer a misreporting rate. Our contributions are summarized as follows. (1) We recast the problem of quantifying the misreporting rate as a causal problem, showing that causal descendants of features can be used to distinguish changes due to genuine adaptation and misreporting. (2) We propose a novel estimator for the misreporting rate that leverages discrepancies in causal effect estimates from manipulated and unmanipulated data. (3) We empirically validate our estimator, showing that it outperforms baselines on a semi-synthetic and real Medicare dataset.

Accepted to the NeurIPS 2025 Workshop on CauScien: Uncovering Causality in Science.

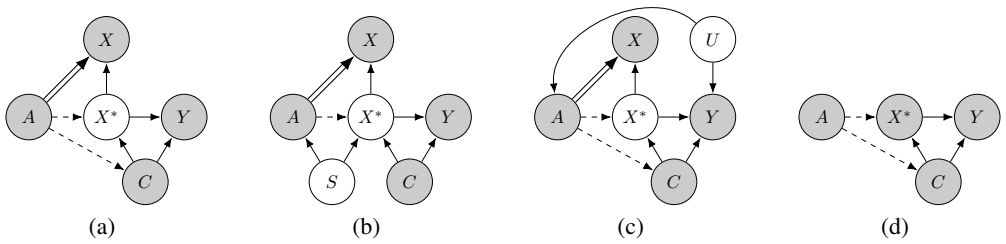

Figure 1: Causal DAGs that describe the setting. White nodes are unobserved, whereas grey nodes are observed. Double-line arrows denote misreporting, while dashed arrows denote genuine adaptation. DAGs (a)-(c) are manipulated data-generating processes; DAG (d) represents unmanipulated data.

2 Preliminaries

Setup. We study a setting where some agents may either genuinely adapt and/or misreport their features. Let A denote the agent identity, X^* the true features, X the (potentially misreported) agent-reported features, Y a downstream variable causally influenced by X^*, and C observed features that may act as confounders or effect modifiers for the relationship between X^* and Y. We use uppercase letters to denote variables and lowercase for their realizations. We assume that we have access to two datasets: (1) A possibly *manipulated dataset* $\mathcal{D} = \{(x_i, y_i, c_i, a_i)\}_{i=1}^N \sim P$, where P follows any of the DAGs in 1(a)-(c). (2) An *unmanipulated dataset* $\mathcal{D}^* = \{(x_i^*, y_i, c_i)\}_{i=1}^M \sim P^*$ generated according to DAG 1(d). $\mathcal{D}^*$ may be pre-deployment data used to train the decision-making model, as agents have no incentive to manipulate their features before the model is deployed.

In Figure 1, dashed arrows indicate genuine adaptation and double-line arrows indicate misreporting. In DAG 1(d), which represents the unmanipulated distribution, only genuine adaptation are allowed. For clarity, we present our main analysis assuming D is sampled according to the DAG in Figure 1(a). However, our results apply without modification to the more complicated DAGs in Figure 1(b)-(c), which include selection bias via an unobserved variable S influencing A and X^*, or an unobserved confounder U between A and Y. Additional allowable DAGs are included in Appendix B.

Assumptions. We assume that X^* and X are binary whereas all other variables may be continuous. Without loss of generality, we assume that $X = 1$ is associated with a higher payout than $X = 0$ which means that agents are not incentivized to misreport features where $X^* = 1$, as formally stated in Assumption 1. We adopt the notation of the Neyman-Rubin potential outcomes framework [25], where $X(X^* = x^*)$ is defined as the counterfactual outcome that we would get if X^* is set to x^*.

Assumption 1 (Optimal Misreporting)**.** $\forall i, \;\; x_i(x_i^* = 1) = 1$

We assume that agents are incentivized to misreport only X^* and genuinely adapt it as follows.

Assumption 2 (Useful Modifications)**.** *Agents may only misreport X^*, or genuinely adapt it by intervening on X^* or its ancestors.*

Our goal is to determine how much an agent misreported their features, without access to X^*. Letting $P_a(V) := P(V|A = a)$ for an arbitrary variable V, we define our estimand of interest as follows:

Definition 1 (Misreporting Rate)**.** *MR* $= P_a(X^* = 0|X = 1)$

The MR quantifies the probability that a reported feature is false. Our approach of estimating the MR relies on estimating the causal effect of X^* on Y, requiring typical causal inference assumptions.

Assumption 3. *The features X^*, C, and the potential outcomes $Y(X^* = 1), Y(X^* = 0)$ satisfy the following properties:*

1. *No unmeasured confounding: $Y(0), Y(1) \perp X^* \mid C$*
2. *Overlap: $P_a(X^* = x^*|C = c), P_a(X = x|C = c), P^*(X^* = x^*|C = c) > 0 \;\; \forall x^*, x, c$*
3. *Consistency: $Y_i(x^*) = y_i$ if $X_i^* = x^*$.*

Finally, to leverage $\mathcal{D}^*$ to recover causal effects needed for estimating misreporting in $\mathcal{D}$, we assume that the conditional average treatment effects are invariant across P_a and P^* for all values of a.

Assumption 4. $\mathbb{E}_{P_a}[Y(1) - Y(0)|C = c] = \mathbb{E}_{P^*}[Y(1) - Y(0)|C = c] \;\; \forall c, a.$

3 Estimating Misreporting Rates

The core challenge of estimating the MR lies in the fact that we only observe the reported features X, but not the true features X^*. This means that the estimand is not identifiable from $\mathcal{D}$ alone. We instead estimate the misreporting rate by leveraging the distinct causal consequences that agent interventions corresponding to genuine adaptation and misreporting have on downstream variables Y.

Our key insight is that genuine adaptation and misreporting affect the causal descendants of X^* differently. When an agent genuinely adapts X^*, this results in a change to its causal descendant Y. In contrast, misreporting only changes the reported feature X, which doesn't causally affect Y. Thus, we can use the causal effect of X on Y and that of X^* on Y as a signature to distinguish misreporting from genuine adaptation. To make progress, we define the "reported" group as those with $X = 1$, and introduce the true and nominal average feature effects on the reported (TAFR and NAFR):

$$\tau_a^* := \int_C (\mathbb{E}_{P_a}[Y(X^* = 1)|C = c] - \mathbb{E}_{P_a}[Y(X^* = 0)|C = c])P_a(C = c|X = 1)dc,$$

$$\tau_a := \int_C (\mathbb{E}_{P_a}[Y|X = 1, C = c] - \mathbb{E}_{P_a}[Y|X = 0, C = c])P_a(C = c|X = 1)dc.$$

These expressions are similar to the commonly studied average treatment effect on the treated. Although observing a difference in τ_a^* and τ_a indicates that an agent misreported, it doesn't give us a rate at which an agent misreports. To obtain the misreporting rate, we must compare the difference in the NAFR and TAFR relative to the baseline causal effect of X^* on Y for the group that is misreported. Since only the variable X^* influences an agent's decision to misreport a datapoint, conditional on the agent, the misreported group will be a random sample of the group where $X^* = 0$. Therefore, the average causal effect of X^* on Y for the misreported will be the average causal effect on the datapoints where $X^* = 0$. We define this as the true average feature effect on the misreported (TAFM):

$$\delta_a^* := \int_C (\mathbb{E}_{P_a}[Y(X^* = 1)|C = c] - \mathbb{E}_{P_a}[Y(X^* = 0)|C = c])P_a(C = c|X^* = 0)dc.$$

Next, we show that the MR can be quantified as the rate in terms of the TAFR, NAFR, and TAFM.

Lemma 1 (Estimator for the misreporting rate)**.** *Let Assumptions 1-3 hold. Then for $\delta_a^* \neq 0$, the MR can be expressed as:*

$$P_a(X^* = 0|X = 1) = \frac{\tau_a^* - \tau_a}{\delta_a^*}.$$

The proof for Lemma 1 and other statements in this section are presented in Appendix C. The lemma states that we can quantify the MR by comparing the true and nominal causal effects of X^* on Y and X on Y. While instructive, Lemma 1 is not very useful as we do not have access to X^* for agent a, and hence we cannot directly estimate τ_a^* or δ_a^* from the manipulated data alone. To resolve this issue, we leverage the unmanipulated dataset $\mathcal{D}^*$ to estimate two other quantities in place of τ_a^* and δ_a^*:

$$\tau'_a := \int_C (\mathbb{E}_{P^*}[Y(X^* = 1)|C = c] - \mathbb{E}_{P^*}[Y(X^* = 0)|C = c])P_a(C = c|X = 1)dc,$$

$$\delta'_a := \int_C (\mathbb{E}_{P^*}[Y(X^* = 1)|C = c] - \mathbb{E}_{P^*}[Y(X^* = 0)|C = c])P_a(C = c|X = 0)dc.$$

Both τ'_a and δ'_a are identifiable because X^* is observed in the unmanipulated dataset and can be used as valid estimators of τ_a^* and δ_a^* to estimate the MR, as we show in Theorem 1.

Theorem 1 (Identifiability)**.** *Let Assumptions 1-4 hold. Then for $\delta'_a \neq 0$, $P_a(X^* = 0|X = 1)$ is identifiable and can be expressed as:*

$$P_a(X^* = 0|X = 1) = \frac{\tau'_a - \tau_a}{\delta'_a}.$$

The proof of Theorem 1 relies on (1) the invariance of conditional causal effects of X^* on Y across P_a and P^* and (2) our assumptions about agent behavior to show that $\tau'_a = \tau_a^*$ and $\delta'_a = \delta_a^*$. We then show that the misreporting rate is identifiable as both δ'_a and τ'_a are identifiable from $\mathcal{D}$, $\mathcal{D}^*$, and standard causal effect assumptions. Guided by Theorem 1, we can now estimate the misreporting rate by comparing causal effect estimates across $\mathcal{D}$ and $\mathcal{D}^*$. We present a formal algorithm that can estimate the misreporting rate for each agent, called the causal misreporting estimator (CMRE), in Appendix F, as well as additional results for the variance of the estimator in Appendix C.

4 Empirical Results

We evaluate the performance of our approach (CMRE) on semi-synthetic and real-world data. We show that CMRE consistently yields reliable estimates of the MR, even when genuine adaptation is present, and outperforms relevant baselines. We compare CMRE against the following baselines: (1) Natural Direct Effect Estimator (**NDEE**): estimates the natural direct effect of the agent A on the feature X. (2) Naive Misreporting Estimator (**NMRE**): is similar to our approach but doesn't control for confounding. Additional simulation details, experiments, and baselines are in the Appendix.

Loan fraud experiments We simulate a scenario where loan applicants may misreport their employment status (X^*) to improve their chances of loan approval. We simulate the true employment status (X^*), reported employment status (X), and if they default (Y). We extract the confounders from a real credit card dataset [31, 32], using age, sex, education, and marital status (C).

We examine how changes in genuine adaptation affect the MR estimates, highlighting the need to account for genuine adaptation when estimating the MR. The results are shown in Figure 2, which shows that our approach (CMRE) gives unbiased, stable estimates of the MR that are unaffected by genuine adaptation. This signals that CMRE can successfully disentangle misreporting from genuine adaptation. By contrast, genuine adaptation affects the estimates of NMRE and NDEE. NMRE gives biased estimates as it does not control for confounding. NDEE also gives biased estimates as it is unable to disentangle the direct causal effect of A on X from the effect mediated through X^*.

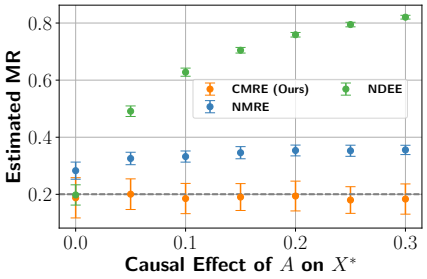

Figure 2: Results from the loan fraud dataset. The x-axis is the causal effect of A on X^*. The y-axis is the estimated misreporting rate. Dashed lines represent the true misreporting rate (MR=0.2).

Misreporting in insurance Next, we highlight the utility of our approach in a real data setting. We aim to identify if private health care insurers misreport enrollees' diagnoses to secure higher payouts. Specifically, the U.S. government calculates how much to pay insurers using a public model based on enrollee diagnoses (X^*), as measured by Hierarchical Condition Categories (HCCs) [24]. This model is trained on an unmanipulated dataset, $\mathcal{D}^*$ where there's no incentive to misreport, whereas data from private insurers, $\mathcal{D}$, may be manipulated. We expect to find evidence of misreporting of HCCs, consistent with existing literature [9, 29]. We gauge the quality of our MR estimates for HCCs by comparing them to estimates of non-payment HCCs: diagnoses where we expect the true MR to be zero, as they are not included in the model. We use mortality as the downstream outcome Y.

Figure 3 shows the MR estimates for two non-payment HCCs and two payment HCCs. Our approach (CMRE) is the only approach that passes the sanity check: it gives MR estimates that are not statistically significantly different from zero for the nonpayment HCCs. This is consistent with our expectation that private insurers have no incentive to misreport nonpayment HCCs. CMRE also estimates significantly high misreporting rates for both of the payment HCCs, which is validated in the health policy literature [2, 17]. In contrast, NMRE estimates a high misreporting rate for all HCCs and NDEE estimates a negative misreporting rate for all HCCs, which does not align with what is expected.

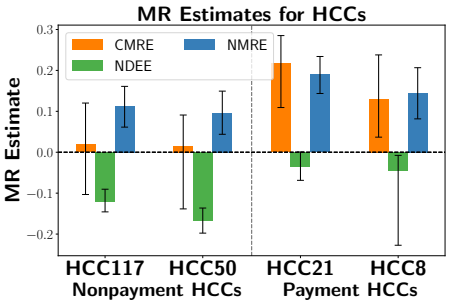

Figure 3: For each plot, the y-axis represents the estimated MR for an HCC code and the error bars represent a 95% confidence interval.

5 Conclusion

In this work, we propose a causal approach to estimating how much strategic agents misreport their features. We show that the misreporting rate is fully identifiable by comparing causal effect estimates between a possibly manipulated and an unmanipulated dataset. We highlight the utility of our approach across empirical experiments over a semi-synthetic and a real Medicare dataset.

Acknowledgments and Disclosure of Funding

We thank the reviewers for their insightful comments. Special thanks to Ezekiel Emanual, Claudia Shi, Cecilia Ehrlichman, and Rohan Singh for their valuable feedback. We also thank Michael Shafir and the Advanced Research Computing team at the University of Michigan for assistance with data access and usage, as well as Will Ferrel for coordinating our meetings. This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor. The authors are supported by a grant from Schmidt Futures (Award No. 70960). The funders had no role in the study design, analysis of results, decision to publish, or preparation of the manuscript. This study was deemed exempt and not regulated by the University of Michigan institutional review board (IRBMED; HUM00230364).

References

[1] Eric P Baumer, JW Andrew Ranson, Ashley N Arnio, Ann Fulmer, and Shane De Zilwa. Illuminating a dark side of the american dream: assessing the prevalence and predictors of mortgage fraud across us counties. *American Journal of Sociology*, 123(2):549–603, 2017.

[2] Caroline S Carlin, Roger Feldman, and Jeah Jung. The mechanics of risk adjustment and incentives for coding intensity in medicare. *Health services research*, 59(3):e14272, 2024.

[3] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[4] Trenton Chang, Lindsay Warrenburg, Sae-Hwan Park, Ravi Parikh, Maggie Makar, and Jenna Wiens. Who's gaming the system? a causally-motivated approach for detecting strategic adaptation. *Advances in Neural Information Processing Systems*, 37:42311–42348, 2024.

[5] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL `http://doi.acm.org/10.1145/2939672.2939785`.

[6] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

[7] Andrew Estornell, Sanmay Das, and Yevgeniy Vorobeychik. Incentivizing truthfulness through audits in strategic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5347–5354, 2021.

[8] Andrew Estornell, Yatong Chen, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. Incentivizing recourse through auditing in strategic classification. In *IJCAI*, 2023.

[9] Michael Geruso and Timothy Layton. Upcoding: evidence from medicare on squishy risk adjustment. *Journal of Political Economy*, 128(3):984–1026, 2020.

[10] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

[11] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2`.

[12] Keegan Harris, Dung Daniel T Ngo, Logan Stapleton, Hoda Heidari, and Steven Wu. Strategic instrumental variable regression: Recovering causal relationships from strategic responses. In *International Conference on Machine Learning*, pages 8502–8522. PMLR, 2022.

[13] Waleed Hilal, S Andrew Gadsden, and John Yawney. Financial fraud: a review of anomaly detection techniques and recent advances. *Expert systems With applications*, 193:116429, 2022.

[14] Guy Horowitz and Nir Rosenfeld. Causal strategic classification: A tale of two shifts. In *International Conference on Machine Learning*, pages 13233–13253. PMLR, 2023.

[15] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9 (3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

[16] Devansh Jalota, Matthew Tsao, and Marco Pavone. Catch me if you can: Combatting fraud in artificial currency based government benefits programs. *arXiv preprint arXiv:2402.16162*, 2024.

[17] Richard Kronick and W Pete Welch. Measuring coding intensity in the medicare advantage program. *Medicare & Medicaid Research Review*, 4(2):mmrr2014–004, 2014.

[18] Sagi Levanon and Nir Rosenfeld. Strategic classification made practical. In *International Conference on Machine Learning*, pages 6243–6253. PMLR, 2021.

[19] Larry M Manevitz and Malik Yousef. One-class svms for document classification. *Journal of machine Learning research*, 2(Dec):139–154, 2001.

[20] John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pages 6917–6926. PMLR, 2020.

[21] The pandas development team. pandas-dev/pandas: Pandas, February 2020. URL `https://doi.org/10.5281/zenodo.3509134`.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[23] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[24] Gregory C Pope, John Kautter, Randall P Ellis, Arlene S Ash, John Z Ayanian, Lisa I Iezzoni, Melvin J Ingber, Jesse M Levy, and John Robst. Risk adjustment of medicare capitation payments using the cms-hcc model. *Health care financing review*, 25(4):119, 2004.

[25] Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[26] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

[27] Naeem Seliya, Azadeh Abdollah Zadeh, and Taghi M Khoshgoftaar. A literature review on one-class classification and its potential applications in big data. *Journal of Big Data*, 8:1–31, 2021.

[28] Yonadav Shavit, Benjamin Edelman, and Brian Axelrod. Causal strategic linear regression. In *International Conference on Machine Learning*, pages 8676–8686. PMLR, 2020.

[29] Elaine Silverman and Jonathan Skinner. Medicare upcoding and hospital ownership. *Journal of health economics*, 23(2):369–389, 2004.

[30] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[31] I-Cheng Yeh. Default of Credit Card Clients. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C55S3H.

[32] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2): 2473–2480, 2009.

[33] Qing Zhan and Hang Yin. A loan application fraud detection method based on knowledge graph and neural network. In *Proceedings of the 2nd international conference on innovation in artificial intelligence*, pages 111–115, 2018.

A Related Work

Strategic Classification and Regression. Our work is related to work on strategic classification and regression, where agents may change their features at some cost [10, 6, 18]. However, it differs in two key aspects. (1) The primary goal is to find a model that is robust to the distribution shifts caused by gaming, often relying on known agents' cost functions and iterative model retraining [23]. In contrast, we seek to estimate how much agents misreport. (2) Our method doesn't require the unrealistic assumptions of known agents' cost function or the iterative model training.

Causal Strategic Classification. Recent work views strategic classification/regression through a causal lens [20, 28, 12, 14] where agents can *only* genuinely adapt their features. They distinguish between two types of genuine adaptation: improvement and gaming, which correspond to adaptations to features that are and are not causally related to the target label, respectively[20]. Unlike us, their focus is on creating models robust to both forms of genuine adaptation [14] and finding models that incentivize improvement over gaming [28, 12]. Closest to our work is Chang et al. [4], who propose an algorithm that can rank agents by their propensity to misreport their features. Unlike our work, their approach can only partially identify how much agents misreport and they do not make a distinction between misreporting and genuine adaptation.

Auditing Policies. Other work seeks to disincentivize agents from misreporting their features through auditing [16, 7, 8]. They define a setting where the decision-maker deploys a transparent auditing policy which allows them to reveal the true features of a limited number of agents selected by the policy. If the agent's true features differ from their reported features, they endure a penalty, which incentivizes them to report their true features. Instead of performing costly audits, our work estimates misreporting by relying on additional unmanipulated data from settings where agents have no incentive to misreport, e.g., data collected before any model was deployed.

Anomaly/Fraud Detection. Our work is closely related to anomaly detection methods aimed at identifying fraudulent instances within a dataset, such as those arising in credit card transactions or insurance claims [13, 3]. Most relevant are one-class classification algorithms [27, 19, 26], which are trained on an unmanipulated dataset to detect anomalies in a manipulated dataset. Unlike our work, these methods focus on identifying specific instances that are anomalous or misreported, whereas our method estimates a rate of misreporting in a dataset. These methods also rely on the assumption that misreported data points differ significantly from normal data points. Our method instead relies on causal assumptions, specifically, that misreporting does not affect the causal descendants of features.

B Additional DAGs

Figures 4(a)-4(g) represent settings in which agents may either genuinely adapt or misreport their features. In contrast, Figure 4(h) represents a scenario involving trustworthy agents that only genuinely adapt their features. In all cases shown in Figures 4(a)-4(g), the decision maker lacks access to X^* but observes X, A, C, and Y. While the main focus of the paper was on the DAG in Figure 4(a), our findings extend naturally to the more complex settings depicted in Figures 4(b)-4(g).

Specifically, the DAGs in Figures 4(b) and 4(f) represent scenarios where some unknown confounding bias may exist between A and X^*, e.g., S. In the context of the Medicare example discussed in the main text, this could arise if enrollees with more chronic conditions (X^*) are more likely to enroll in a private health insurance plan (A). Notably, our approach doesn't require controlling for S, as it's not a confounder between X^* and Y.

The DAGs in Figures 4(c), 4(e), and 4(g) illustrate settings where an unobserved confounder may influence both A and Y. For example, this could occur if enrollees who prefer private insurance

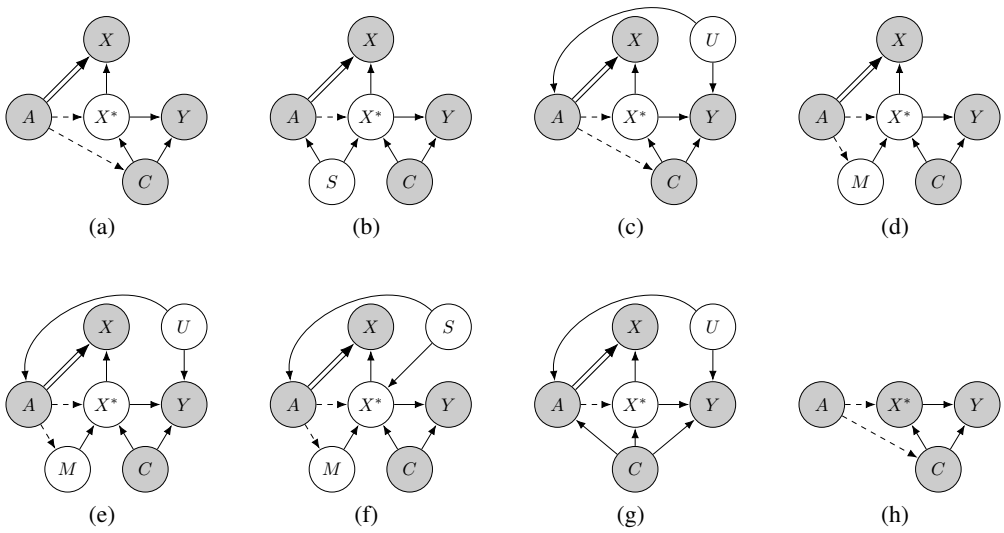

Figure 4: Causal DAGs that describe the setting of this paper. White nodes are unobserved, whereas grey nodes are observed. Double-line arrows represent misreporting, while dashed arrows represent genuine adaptation. DAGs (a)-(g) represent manipulated data-generating processes, while DAG (h) represents unmanipulated data.

plans (A) also happen to have worse health outcomes (Y). Again, our approach does not require controlling for U. Although U is a confounder of X^* and Y, the backdoor path can be blocked by conditioning on A, which means that an adjustment for U is unnecessary.

Finally, the DAGs in Figures 4(d)-4(f) capture settings where an agent may genuinely adapt their features by intervening on a mediator M that lies between A and X^*. For example, this could occur if private health insurers (A) are more likely to offer free gym memberships (M), which influence the true health status of their enrollees (X^*). As before, our approach does not require any knowledge of the mediators an agent intervenes on in order to estimate the misreporting rate, as M is not a confounder of X^* and Y.

C Main Proofs

Each of the proofs within this Section assume that the dataset $\mathcal{D} \sim P_a$ is generated according to any one of the DAGs in Figures 4(a)-4(g).

C.1 Proof for Lemma 1

Lemma 1 is important to build up to Theorem 1. It shows that the MR can be estimated by comparing the true and nominal causal effects of X^* on Y and X on Y.

Lemma A1 (Estimator for the misreporting rate; Lemma 1 in the main text)**.** *Let Assumptions 1-3 hold. Then for $\delta_a^* \neq 0$, the MR can be expressed as:*

$$P_a(X^* = 0|X = 1) = \frac{\tau_a^* - \tau_a}{\delta_a^*}$$

Proof. Our proof proceeds in three main steps. First, we decompose τ_a into two terms: τ_a^* and an additional bias term. Second, we show that this additional term can be written as a function of our target estimand, $P(X^* = 0|X = 1)$. Third and finally, we show that using simple algebra, we can express our target estimand as a function of τ_a^*, τ_a and δ_a^*

Step 1: Decomposing τ_a into τ_a^* and an additional term

$$
\begin{aligned}
\tau_a &= \int_C (\mathbb{E}_{P_a}[Y|X=1,C=c] - \mathbb{E}_{P_a}[Y|X=0,C=c])P_a(C=c|X=1)dc \\
&= \int_C \mathbb{E}_{P_a}[Y|X=1,C=c,X^*=1]P_a(X^*=1|X=1,C=c)P_a(C=c|X=1)dc \\
&+ \int_C \mathbb{E}_{P_a}[Y|X=1,C=c,X^*=0]P_a(X^*=0|X=1,C=c)P_a(C=c|X=1)dc \\
&- \int_C \mathbb{E}_{P_a}[Y|X=0,C=c]P_a(C=c|X=1)dc \\
&= \int_C \mathbb{E}_{P_a}[Y|X^*=1,C=c](1-P_a(X^*=0|X=1,C=c))P_a(C=c|X=1)dc \\
&+ \int_C \mathbb{E}_{P_a}[Y|X^*=0,C=c]P_a(X^*=0|X=1,C=c)P_a(C=c|X=1)dc \\
&- \int_C \mathbb{E}_{P_a}[Y|X=0,C=c]P_a(C=c|X=1)dc \\
&= \int_C \mathbb{E}_{P_a}[Y|X^*=1,C=c]P_a(C=c|X=1)dc \\
&- \int_C \mathbb{E}_{P_a}[Y|X^*=1,C=c]P_a(X^*=0|X=1,C=c)P_a(C=c|X=1)dc \\
&+ \int_C \mathbb{E}_{P_a}[Y|X^*=0,C=c]P_a(X^*=0|X=1,C=c)P_a(C=c|X=1)dc \\
&- \int_C \mathbb{E}_{P_a}[Y|X^*=0,C=c]P_a(C=c|X=1)dc \\
&= \tau_a^* - \int (\mathbb{E}_{P_a}[Y|X^*=1,C] - \mathbb{E}_{P_a}[Y|X^*=0,C])P_a(X^*=0|X=1,C)P_a(C|X=1)dc.
\end{aligned}
$$

The third equality holds as $Y \perp X|X^*, A$ for the DAGs in Figure 4(a)-4(g) and the fourth equality holds due to Assumption 1.

Step 2: Expressing the additional term as a function of $P(X^*=0|X=1)$ Next, to explicitly show that this additional term is a direct consequence of misreporting, we can rewrite it in terms of the misrepoting rate:

$$
\begin{aligned}
&\int_C (\mathbb{E}_{P_a}[Y|X^*=1,C] - \mathbb{E}_{P_a}[Y|X^*=0,C])P_a(X^*=0|X=1,C)P_a(C|X=1)dc \\
&= \int_C (\mathbb{E}_{P_a}[Y|X^*=1,C] - \mathbb{E}_{P_a}[Y|X^*=0,C])\frac{P_a(X^*=0,C|X=1)}{P_a(C|X=1)}P_a(C|X=1)dc \\
&= \int_C (\mathbb{E}_{P_a}[Y|X^*=1,C] - \mathbb{E}_{P_a}[Y|X^*=0,C])P_a(X^*=0|X=1)P_a(C|X^*=0,X=1)dc \\
&= P_a(X^*=0|X=1)\int_C (\mathbb{E}_{P_a}[Y|X^*=1,C] - \mathbb{E}_{P_a}[Y|X^*=0,C])P_a(C|X^*=0,X=1)dc \\
&= P_a(X^*=0|X=1)\int_C \mathbb{E}_{P_a}[Y|X^*=1,C] - \mathbb{E}_{P_a}[Y|X^*=0,C]P_a(C|X^*=0)dc \\
&= P_a(X^*=0|X=1)\int_C \mathbb{E}_{P_a}[Y(X^*=1)|C] - \mathbb{E}_{P_a}[Y(X^*=0)|C=c]P_a(C|X^*=0)dc \\
&= P_a(X^*=0|X=1)\delta_a^*.
\end{aligned}
$$

The fourth equality comes directly from the fact that $C \perp X|X^*, A$ for the DAGs in Figures 4(a)-4(g). The fifth equality comes from Assumption 3, as all confounders are controlled for. Notably, both M and S are not confounders of X^* and Y. The variable U is a confounder of X^* and Y, however, the

backdoor path is blocked by A, so it doesn't need to be directly controlled for. Overall, this shows that any difference between τ_a and τ_a^* is directly related to the misreporting rate.

Step 3: Getting the expression for the final target estimand Finally, we can obtain a way to estimate the misreporting rate by rearanging the terms:

$$\tau_a = \tau_a^* - P_a(X^* = 0|X = 1)\delta_a^* \implies P_a(X^* = 0|X = 1) = \frac{\tau_a^* - \tau_a}{\delta_a^*},$$

for $\delta_a^* \neq 0$. Therefore, by comparing the difference in causal effects, we can identify the misreporting rate. □

C.2 Proof for Theorem 1

We now build upon the result from Lemma 1 as we work toward our main theorem. Before presenting the proof of Theorem 1, we first introduce an additional Lemma which shows that τ_a, τ_a', and δ_a' are identifiable using $\mathcal{D}$ and $\mathcal{D}^*$, along with standard causal estimation assumptions. Then, in Theorem 1, we demonstrate that the misreporting rate is identifiable by showing that $\tau_a' = \tau_a^*$ and $\delta_a' = \delta_a^*$. This proof follows from Assumption 4, which states that the conditional causal effect of X^* on Y will remain invariant across both strategic and non-strategic populations.

Lemma A2 (Identifiability of τ_a, τ_a', and δ_a')**.** *Let Assumption 3 hold. Then τ_a, τ_a', and δ_a' are identifiable using $\mathcal{D}$ and $\mathcal{D}'$.*

Proof. First, recall that

$$\tau_a := \int_C (\mathbb{E}_{P_a}[Y|X = 1, C = c] - \mathbb{E}_{P_a}[Y|X = 0, C = c])P_a(C = c|X = 1)dc.$$

We know that τ_a is identifiable using only $\mathcal{D}$ as Y, X, and C are all known in $\mathcal{D}$.

Next, recall that

$$\tau_a' := \int_C (\mathbb{E}_{P^*}[Y(X^* = 1)|C = c] - \mathbb{E}_{P^*}[Y(X^* = 0)|C = c])P_a(C = c|X = 1)dc$$

and

$$\delta_a' := \int_C (\mathbb{E}_{P^*}[Y(X^* = 1)|C = c] - \mathbb{E}_{P^*}[Y(X^* = 0)|C = c])P_a(C = c|X = 0)dc.$$

Again, we know that $P_a(C = c|X = 1)$ and $P_a(C = c|X = 0)$ are identifiable using only $\mathcal{D}$. Therefore, to show that τ_a' and δ_a' are identifiable, we must show that

$$\mathbb{E}_{P^*}[Y(X^* = 1)|C = c] - \mathbb{E}_{P^*}[Y(X^* = 0)|C = c]$$

is identifiable. This follows immediately from Assumptions 3:

$$\begin{aligned}
&\mathbb{E}_{P^*}[Y(X^* = 1)|C = c] - \mathbb{E}_{P^*}[Y(X^* = 0)|C = c] \\
&= \mathbb{E}_{P^*}[Y(X^* = 1)|X^* = 1, C = c] - \mathbb{E}_{P^*}[Y(X^* = 0)|X^* = 0, C = c] \\
&= \mathbb{E}_{P^*}[Y|X^* = 1, C = c] - \mathbb{E}_{P^*}[Y|X^* = 0, C = c]
\end{aligned}$$

Therefore, τ_a, τ_a', and δ_a' are identifiable using $\mathcal{D}$ and $\mathcal{D}'$. □

Theorem A1 (Identifiability; Theorem 1 in the main text)**.** *Let Assumptions 1-4 hold. Then for $\delta_a' \neq 0$, $P_a(X^* = 0|X = 1)$ is identifiable and can be expressed as:*

$$P_a(X^* = 0|X = 1) = \frac{\tau_a' - \tau_a}{\delta_a'}.$$

Proof. We know that τ_a, τ_a', and δ_a' are identifiable using $\mathcal{D}$ and $\mathcal{D}^*$ by Lemma A2. Therefore, to complete this proof, we only need to show that $\tau_a' = \tau_a^*$ and $\delta_a' = \delta_a^*$, as implied by Lemma 1.

First, we show that $\tau'_a = \tau^*_a$. Recall that

$$\tau'_a := \int_C (\mathbb{E}_{P^*}[Y(X^*=1)|C=c] - \mathbb{E}_{P^*}[Y(X^*=0)|C=c])P_a(C=c|X=1)dc$$

and

$$\tau^*_a := \int_C (\mathbb{E}_{P_a}[Y(X^*=1)|C=c] - \mathbb{E}_{P_a}[Y(X^*=0)|C=c])P_a(C=c|X=1)dc.$$

Since

$$\mathbb{E}_{P_a}[Y(1)-Y(0)|C=c] = \mathbb{E}_{P^*}[Y(1)-Y(0)|C=c]$$

for all c by Assumption 4, it follows immediately that $\tau'_a = \tau^*_a$.

Next, recall that

$$\delta'_a := \int_C (\mathbb{E}_{P^*}[Y(X^*=1)|C=c] - \mathbb{E}_{P^*}[Y(X^*=0)|C=c])P_a(C=c|X=0)dc$$

and

$$\delta^*_a := \int_C (\mathbb{E}_{P_a}[Y(X^*=1)|C=c] - \mathbb{E}_{P_a}[Y(X^*=0)|C=c])P_a(C=c|X^*=0)dc.$$

We already know that the conditional causal effects of X^* on Y are the same across P^* and P_a. It remains to show that $P_a(C=c|X^*=0) = P_a(C=c|X=0)$ for all values of c to show that $\delta'_a = \delta^*_a$. We establish this equality next.

To show this, we simply apply the law of total probability as follows:

$$\begin{aligned} P_a(C=c|X=0) &= P_a(C=c|X=0,X^*=1)P(X^*=1|X=0) \\ &\quad + P_a(C=c|X=0,X^*=0)P(X^*=0|X=0) \\ &= P_a(C=c|X=0,X^*=0) \\ &= P_a(C=c|X^*=0). \end{aligned}$$

The second equality follows because $P_a(X^*=1|X=0)=0$ and $P(X^*=0|X=0)=1$ by Assumption 1. The third equality follows as $C \perp X|A, X^*$ for all DAGs in Figures 4(a)-4(g). Note that this finding is intuitive: it can be traced back to the assumption that the agents pick who to misreport at random, which is implied by the DAGs.

Thus, the MR is identifiable.

□

C.3 Proof for Theorem 2

Theorem A2 (Variance)**.** *Let $\hat{\tau}_a$, $\hat{\tau}'_a$, and $\hat{\delta}'_a$ be asymptotically normal estimators with an asymptotic variance of $\sigma^2_{\tau_a}$, $\sigma^2_{\tau'_a}$, and $\sigma^2_{\delta'_a}$. Also let $\sigma_{\tau_a\tau'_a}$, $\sigma_{\tau_a\delta'_a}$, and $\sigma_{\delta'_a\tau'_a}$ denote the covariance between the estimators and $\xrightarrow{d}$ denote convergence in distribution. Suppose that $N=M=n$, then for $\delta'_a \neq 0$ and $\hat{\delta}'_a \neq 0$,*

$$\sqrt{n}[\frac{\hat{\tau}'_a - \hat{\tau}_a}{\hat{\delta}'_a} - \frac{\tau'_a - \tau_a}{\delta'_a}] \xrightarrow{d} \mathcal{N}(0, \frac{\sigma^2_{\tau'_a} + \sigma^2_{\tau_a} - 2\sigma_{\tau'_a\tau_a}}{{\delta'_a}^2} + 2\frac{\tau_a - \tau'_a}{{\delta'_a}^3}(\sigma_{\tau'_a\delta'_a} - \sigma_{\tau_a\delta'_a}) + \frac{(\tau_a-\tau'_a)^2}{{\delta'_a}^4}\sigma^2_{\delta'_a})$$

Proof. By the definition of asymptotic normality, each estimator has the following asymptotic distributions, where $\sigma^2_{\tau'_a}$ is asymptotic variance of $\hat{\tau}'_a$, $\sigma^2_{\tau_a}$ is asymptotic variance of $\hat{\tau}_a$, and $\sigma^2_{\delta'_a}$ is asymptotic variance of $\hat{\delta}'_a$:

$$\sqrt{n}[\hat{\tau}'_a - \tau'_a] \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\tau'_a}),$$

$$\sqrt{n}[\hat{\tau}_a - \tau_a] \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\tau_a}), \text{ and}$$

$$\sqrt{n}[\hat{\delta}'_a - \delta'_a] \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\delta'_a}).$$

To proceed, we define the function $g(\hat{\tau}'_a, \hat{\tau}_a, \hat{\delta}'_a)$ as an estimator for the misreporting rate:

$$g(\hat{\tau}'_a, \hat{\tau}_a, \hat{\delta}'_a) = \frac{\hat{\tau}'_a - \hat{\tau}_a}{\hat{\delta}'_a}.$$

Since $\hat{\tau}'_a$, $\hat{\tau}_a$, $\hat{\delta}'_a$ are asymptotically normal, we can apply the delta method [30] to find the asymptotic variance of $g(\hat{\tau}'_a, \hat{\tau}_a, \hat{\delta}'_a)$, which states that

$$\sqrt{n}[g(\hat{\tau}'_a, \hat{\tau}_a, \hat{\delta}'_a) - g(\tau'_a, \tau_a, \delta'_a)] \xrightarrow{d} \mathcal{N}(0, \nabla g(\tau'_a, \tau_a, \delta'_a)\Sigma\nabla g(\tau'_a, \tau_a, \delta'_a)^\top),$$

where

$$\nabla g(\tau'_a, \tau_a, \delta'_a) = \begin{pmatrix} \frac{1}{\delta'_a} & \frac{-1}{\delta'_a} & \frac{\tau_a - \tau'_a}{{\delta'_a}^2} \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} \sigma^2_{\tau'_a} & \sigma_{\tau_a\tau'_a} & \sigma_{\delta'_a\tau'_a} \\ \sigma_{\tau'_a\tau_a} & \sigma^2_{\tau_a} & \sigma_{\delta'_a\tau_a} \\ \sigma_{\tau'_a\delta'_a} & \sigma_{\tau_a\delta'_a} & \sigma^2_{\delta'_a} \end{pmatrix}$$

Therefore, we can calculate the asymptotic variance as follows:

$$\begin{aligned}
\nabla g(\tau'_a, \tau_a, \delta'_a)^\top \Sigma \nabla g(\tau'_a, \tau_a, \delta'_a) &= \begin{pmatrix} \frac{1}{\delta'_a} & \frac{-1}{\delta'_a} & \frac{\tau_a - \tau'_a}{{\delta'_a}^2} \end{pmatrix} \begin{pmatrix} \sigma^2_{\tau'_a} & \sigma_{\tau_a\tau'_a} & \sigma_{\delta'_a\tau'_a} \\ \sigma_{\tau'_a\tau_a} & \sigma^2_{\tau_a} & \sigma_{\delta'_a\tau_a} \\ \sigma_{\tau'_a\delta'_a} & \sigma_{\tau_a\delta'_a} & \sigma^2_{\delta'_a} \end{pmatrix} \begin{pmatrix} \frac{1}{\delta'_a} \\ \frac{-1}{\delta'_a} \\ \frac{\tau_a - \tau'_a}{{\delta'_a}^2} \end{pmatrix} \\
&= \begin{pmatrix} \sigma^2_{\tau'_a}\frac{1}{\delta'_a} - \sigma_{\tau'_a\tau_a}\frac{1}{\delta'_a} + \sigma_{\tau'_a\delta'_a}(\frac{\tau_a - \tau'_a}{{\delta'_a}^2}) \\ \sigma_{\tau_a\tau'_a}\frac{1}{\delta'_a} - \sigma^2_{\tau_a}\frac{1}{\delta'_a} + \sigma_{\tau_a\delta'_a}(\frac{\tau_a - \tau'_a}{{\delta'_a}^2}) \\ \sigma_{\delta'_a\tau'_a}\frac{1}{\delta'_a} - \sigma_{\delta'_a\tau_a}\frac{1}{\delta'_a} + \sigma^2_{\delta'_a}(\frac{\tau_a - \tau'_a}{{\delta'_a}^2}) \end{pmatrix} \begin{pmatrix} \frac{1}{\delta'_a} \\ \frac{-1}{\delta'_a} \\ \frac{\tau_a - \tau'_a}{{\delta'_a}^2} \end{pmatrix} \\
&= \sigma^2_{\tau'_a}\frac{1}{{\delta'_a}^2} - \sigma_{\tau'_a\tau_a}\frac{1}{{\delta'}^2_a} + \sigma_{\tau'_a\delta'_a}\frac{\tau_a - \tau'_a}{{\delta'_a}^3} \\
&- \sigma_{\tau_a\tau'_a}\frac{1}{{\delta'}^2_a} + \sigma^2_{\tau_a}\frac{1}{{\delta'}^2_a} - \sigma_{\tau_a\delta'_a}\frac{\tau_a - \tau'_a}{{\delta'_a}^3} \\
&+ \sigma_{\delta'_a,\tau'_a}\frac{\tau_a - \tau'_a}{{\delta'}^3_a} - \sigma_{\delta'_a\tau_a}\frac{\tau_a - \tau'_a}{{\delta'}^3_a} + \sigma^2_{\delta'_a}\frac{(\tau_a - \tau'_a)^2}{{\delta'_a}^4} \\
&= \sigma^2_{\tau'_a}\frac{1}{{\delta'_a}^2} + \sigma^2_{\tau_a}\frac{1}{{\delta'}^2_a} - 2\sigma_{\tau'_a\tau_a}\frac{1}{{\delta'}^2_a} \\
&+ 2\sigma_{\tau'_a\delta'_a}\frac{\tau_a - \tau'_a}{{\delta'_a}^3} - 2\sigma_{\tau_a\delta'_a}\frac{\tau_a - \tau'_a}{{\delta'_a}^3} \\
&+ \sigma^2_{\delta'_a}\frac{(\tau_a - \tau'_a)^2}{{\delta'_a}^4} \\
&= \frac{1}{{\delta'_a}^2}(\sigma^2_{\tau'_a} + \sigma^2_{\tau_a} - 2\sigma_{\tau'_a\tau_a}) \\
&+ 2\frac{\tau_a - \tau'_a}{{\delta'_a}^3}(\sigma_{\tau'_a\delta'_a} - \sigma_{\tau_a\delta'_a}) \\
&+ \frac{(\tau_a - \tau'_a)^2}{{\delta'_a}^4}\sigma^2_{\delta'_a}.
\end{aligned}$$

Therefore, $\sqrt{n}[\frac{\hat{\tau'_a} - \hat{\tau_a}}{\hat{\delta}'_a} - \frac{\tau'_a - \tau_a}{\delta'_a}]$ asymptotically converges to the following normal distribution:

$$\sqrt{n}[\frac{\hat{\tau}'_a - \hat{\tau}_a}{\hat{\delta}'_a} - \frac{\tau'_a - \tau_a}{\delta'_a}] \xrightarrow{d} \mathcal{N}(0, \frac{\sigma^2_{\tau'_a} + \sigma^2_{\tau_a} - 2\sigma_{\tau'_a\tau_a}}{{\delta'_a}^2} + 2\frac{\tau_a - \tau'_a}{{\delta'_a}^3}(\sigma_{\tau'_a\delta'_a} - \sigma_{\tau_a\delta'_a}) + \frac{(\tau_a - \tau'_a)^2}{{\delta'_a}^4}\sigma^2_{\delta'_a})$$

□

D Additional Estimands

In this section, we show that if we can identify the main estimand of interest, $P_a(X = 1|X^* = 0)$, we can also identify other useful estimands, which are defined below.

Definition 2 (Difference in Marginals)**.** *DIM* $= P_a(X = 1) - P_a(X^* = 1)$.

Definition 3 (False Positive Rate)**.** *FPR* $= P_a(X = 1|X^* = 0)$.

The estimand in definition 3 can simply be interepreted as the false positive rate whereas the estimand in definition 2 can be thought of as the probability that a feature was misreported.

To establish that the estimand in definition 2 is identifiable, we first establish that our estimand of interest, $P_a(X = 1) - P_a(X^* = 0)$, can be expressed as the joint distribution $P_a(X = 1, X^* = 0)$ in Lemma A3. Identifiability follows from Theorem 1 and a simple application of Bayes rule as both $P_a(X^* = 0|X = 1)$ and $P_a(X = 1)$ are identifiable.

Additionally, since Lemma A3 implies that both $P_a(X = 1, X^* = 0)$ and $P_a(X^* = 0)$ are identifiable, we can show that the estimand in definition 3 is also identifiable.

Lemma A3. *Let Assumption 1 hold. Then* $P_a(X = 1) - P_a(X^* = 1) = P_a(X = 1, X^* = 0)$

Proof.

$$\begin{aligned} P_a(X = 1, X^* = 0) &= P_a(X = 1, X^* = 0) + P_a(X^* = 1) - P_a(X^* = 1) \\ &= P_a(X = 1, X^* = 0) + P_a(X = 1|X^* = 1)P_a(X^* = 1) - P_a(X^* = 1) \\ &= P_a(X = 1, X^* = 0) + P_a(X = 1, X^* = 1) - P_a(X^* = 1) \\ &= P_a(X = 1) - P_a(X^* = 1), \end{aligned}$$

where the second equality follows because $P_a(X = 1|X^* = 1) = 1$ by Assumption 1. □

Corollary A1 (Identifiability of difference in marginals)**.** *Let Assumptions 1-4 hold. Then for* $\delta'_a \neq 0$*,* $P_a(X = 1) - P_a(X^* = 1)$ *is identifiable and can be expressed as:*

$$P_a(X = 1) - P_a(X^* = 1) = \frac{\tau'_a - \tau_a}{\delta'_a} \times P_a(X = 1).$$

Proof. The proof relys on a simple application of Bayes rule, and results from Lemma A3 and Theorem 1. Specifically, we have that:

$$\begin{aligned} P_a(X = 1) - P_a(X^* = 1) &= P_a(X = 1, X^* = 0) \\ &= P_a(X^* = 0|X = 1)P_a(X = 1), \end{aligned}$$

where the first equality follows by Lemma A3 and the second equality follows by Bayes rule. By theorem 1, the first term ($P_a(X^* = 0|X = 1)$) is identifiable, and $P_a(X = 1)$ is identifiable because all variables required for estimation are observed. □

Corollary A2 (Identifiability of false positive rate)**.** *Let Assumptions 1-4 hold. Then for* $\delta'_a \neq 0$*,* $P_a(X = 1|X^* = 0)$ *is identifiable and can be expressed as:*

$$P_a(X = 1|X^* = 0) = \frac{\tau'_a - \tau_a}{\delta'_a} \times P_a(X = 1).$$

Proof. From Lemma A3, we can derive $P(X^* = 0)$ as follows:

$$\begin{aligned} P_a(X = 1) - P_a(X^* = 1) = P_a(X = 1, X^* = 0) &\implies \\ P_a(X = 1) - P_a(X = 1, X^* = 0) = P_a(X^* = 1) &\implies \\ 1 - \{P_a(X = 1) - P_a(X = 1, X^* = 0)\} = 1 - P_a(X^* = 1) &\implies \\ P_a(X = 0) + P_a(X = 1, X^* = 0) = P_a(X^* = 0) & \end{aligned}$$

By Bayes' theorem, we can write the estimand as

$$P_a(X=1|X^*=0) = \frac{P_a(X^*=0|X=1)P_a(X=1)}{P_a(X^*=0)}$$
$$= \frac{P_a(X^*=0|X=1)P_a(X=1)}{P_a(X=0)+P_a(X=1,X^*=0)}$$

Thus, since $P_a(X^*=0|X=1)$, $P_a(X=1,X^*=0)$, $P_a(X=1)$, and $P_a(X=0)$ are identifiable, $P_a(X=1|X^*=0)$ must be identifiable.

□

E Datasets

E.1 Medicare Dataset

The medicare dataset used in our experiments consists of insurance claims data from real U.S. Medicare enrollees enrolled in either Traditional Medicare or a private medicare insurance plan. The data was provided to the authors under a data usage agreement with the Centers for Medicare and Medicaid Services (CMS). For our experiments, we only use enrollees that had Medicare coverage in both 2019 (t) and 2018 ($t-1$). We exclude enrollees who were dual-eligible (i.e., are eligible for both U.S. Medicaid and Medicare), had end-stage renal disease, or were below the age of 65 for the year $t-1$. In addition, we exclude all enrollees who resided outside of the 50 U.S. states, the District of Columbia, Puerto Rico, or the U.S. Virgin Islands.

Each of the private medicare insurers is treated as a strategic agent that may misreport enrollee features. We used five agents in total for our experiments. Four agents correspond to the largest private insurers based on the total number of enrollees in year t. The fifth agent is created by aggregating the enrollees from all other smaller insurers. In contrast, Traditional Medicare was treated as a trustworthy agent that doesn't manipulate enrollee data, as there is no incentive to misreport.

The goal of our analysis is to assess how much private medicare insurers misreport HCC codes, which are binary variables that indicate if an enrollee has been diagnosed with a specific medical condition. We use V23 HCC codes, as defined by CMS, which are derived by mapping ICD-10 diagnosis codes reported in the claims data. There are two types of HCC codes: payment HCCs, which are used by a risk-adjustment model to predict future healthcare costs, and nonpayment HCCs, which are not used to determine costs. We expect the misreporting rate for each nonpayment HCC to be zero as there is no incentive for private insurers to misreport them.

For our analysis, we partition the enrollees into two different cohorts: stayers and switchers. To derive the stayers cohort, we sampled enrollees enrolled in Traditional Medicare for all 12 months in year $t-1$ and were not enrolled in a private insurance plan in year t. For the switchers cohort, we used enrollees that were enrolled in Traditional Medicare for all 12 months in year $t-1$ and were enrolled in a private insurance plan for at least one month in year t. We only used a 20% random sample of the eligible stayers cohort (868255 samples) and 100% of the eligible switchers cohort (166539 samples). For the outcome (Y), we use the enrollee's death status in year t.

For the features (X), we used both payment and nonpayment HCC codes, consisting of 83 and 99 codes, respectively. As covariates, we used the enrollee's age, race, sex, and the payment HCCs from year $t-1$ to ensure they were not misreported. To obtain low variance estimates, we restricted our analysis to payment and nonpayment HCCs with the largest causal effects on death and where at least 1% of the switchers enrollees in year t had the HCC code.

E.2 Loan Datasets

In our loan dataset simulations, we model a setting where loan applicants may either genuinely adapt or misreport their employment status to improve their chances of getting approved for a loan. For each of our simulations, we simulate a single strategic agent ($A=1$) and a single nonstrategic agent ($A=0$). In addition to the semi-synthetic dataset used for the experiments in Section 5, we generate

additional datasets based on different DAGs in Figure 4. The data generation process for the other datasets is explained in Appendix G.

All of the simulations use the covariates extracted from a real credit card dataset [31]. These include three binary variables: marriage status (C_M), sex (C_S), and education (C_E), as well as another variable representing a person's age (C_A). We use min-max normalization so that C_A is between 0 and 1. The agent variable (A), the variable for employment status (X^*), and the variable indicating if a loan applicant defaulted (Y), are all generated using the covariates. Misreporting is done in accordance with the following equation:

$$X_i \sim X_i^* + A_i(1 - X_i^*)\text{Bernoulli}(\mu)$$

where μ is picked to target a desired MR (default = 0.2). Each experiment is repeated 100 times, with new draws of A, X, X^*, and Y. Across all experiments, we use an 80/20 train/test split of $\mathcal{D}$.

F Estimators

In this section, we present additional details about our primary method (CMRE) as well as the baseline approaches (NMRE, NDEE, and OC-SVM). We also specify the hyperparameters and libraries used to implement each method in our experiments.

F.1 CMRE

Recall that for a suitable function class $\mathcal{F}$, a loss function ℓ, and N_a – the number of data points in $\mathcal{D}$ for which $A = a$ – we define

$$f_a(c, x) = \arg\min_{f \in \mathcal{F}} \frac{1}{N_a} \sum_{i: i \in \mathcal{D}, a_i = a} \ell(f(c_i, x_i), y_i), \quad \text{and} \quad \theta_a(c) := f_a(c, 1) - f_a(c, 0) \tag{1}$$

and

$$f^*(c, x^*) = \arg\min_{f \in \mathcal{F}} \frac{1}{M} \sum_{i: i \in \mathcal{D}^*} \ell(f(c_i, x_i^*), y_i) \quad \text{and} \quad \theta^*(c) := f^*(c, 1) - f^*(c, 0). \tag{2}$$

Recall that N_{ax} denotes the number of data points in $\mathcal{D}$ for which $A = a$ and $X = x$. Using this, we compute the estimates for τ_a', τ_a and δ_a' as follows:

$$\hat{\tau}_a' = \frac{1}{N_{a1}} \sum_{\substack{i: i \in \mathcal{D}, x_i = 1, \\ a_i = a}} \theta^*(c_i), \quad \hat{\tau}_a = \frac{1}{N_{a1}} \sum_{\substack{i: i \in \mathcal{D}, x_i = 1, \\ a_i = a}} \theta(c_i), \quad \hat{\delta}_a' = \frac{1}{N_{a0}} \sum_{\substack{i: i \in \mathcal{D}, x_i = 0, \\ a_i = a}} \theta^*(c_i). \tag{3}$$

To estimate $\theta_a(c)$ and $\theta^*(c)$, we employ an S-learner, where the models f_a and f^* are implemented using XGBoost. We use the default hyperparameters provided by the XGBoost library in Python to train each model [5], including a learning rate of 0.3, a maximume tree depth of 6, and L2 regularization with a coefficient of 1.

The complete algorithm for CMRE is summarized in 1. We note that for our experiments, we split $\mathcal{D}$ such that the data used to train $f_a(c, x)$ in equation 1 is different than the data used to estimate the MR in equation 3. Specifically, 80% of the data in $\mathcal{D}$ is used to train $f_a(c, x)$ in equation 1 and the other 20% is used to estimate $\hat{\tau}_a'$, $\hat{\tau}_a$, and $\hat{\delta}_a'$.

F.2 NMRE

NMRE adopts a similar strategy to CMRE for estimating the misreporting rate. Specifically, it leverages the differences in causal effect estimates. However, the key distinction between NMRE and CMRE is that NMRE doesn't account for potential confounders or treatment effect modifiers between X^* and Y. As a result, NMRE employs a simple difference-in-means estimator to estimate

Algorithm 1 CMRE algorithm

Input: $\mathcal{D} = \{(x_i, y_i, c_i, a_i)\}_i^N$ and $\mathcal{D}^* = \{(x_i^*, y_i, c_i)\}_i^M$
Output: $\widehat{\text{MR}}$, an estimate of the MR for each agent
  for each agent a **do**
    Estimate $\theta_a(c)$ using equation 1
    Estimate $\theta^*(c)$ using equation 2
    Estimate $\hat{\tau}'_a$, $\hat{\tau}_a$, and $\hat{\delta}'_a$ using equation 3
    return $\frac{\hat{\tau}'_a - \hat{\tau}_a}{\hat{\delta}'_a}$
  end for

the average effect of the feature on Y over both $\mathcal{D}^*$ and $\mathcal{D}$. Therefore, to estimate the MR for a given agent a, we define

$$\hat{\tau}' = \frac{1}{M_1} \sum_{i:i\in\mathcal{D}^*, x_i^*=1} y_i - \frac{1}{M_0} \sum_{i:i\in\mathcal{D}^*, x_i^*=0} y_i \tag{4}$$

and

$$\hat{\tau}_a = \frac{1}{N_{a1}} \sum_{\substack{i:i\in\mathcal{D}, x_i=1, \\ a_i=a}} y_i - \frac{1}{N_{a0}} \sum_{\substack{i:i\in\mathcal{D}, x_i=0, \\ a_i=a}} y_i, \tag{5}$$

where M_x is the number of datapoints in $\mathcal{D}^*$ where $X^* = x$.

The misreporting rate is the estimated as:

$$\hat{\text{MR}} = \frac{\hat{\tau}' - \hat{\tau}_a}{\hat{\tau}'}.$$

The complete algorithm for NMRE is summarized in Algorithm 2.

Algorithm 2 NMRE algorithm

Input: $\mathcal{D} = \{(x_i, y_i, c_i, a_i)\}_i^N$ and $\mathcal{D}^* = \{(x_i^*, y_i, c_i)\}_i^M$
Output: $\widehat{\text{MR}}$, an estimate of the MR for each agent
  for each agent a **do**
    Estimate $\hat{\tau}'$ using equation 4
    Estimate $\hat{\tau}_a$ using equation 5
    return $\frac{\hat{\tau}' - \hat{\tau}_a}{\hat{\tau}'}$
  end for

F.3 NDEE

The NDEE baseline estimates the misreporting rate by computing a quanitity similar to the natural direct effect of A on X, divided by the probability $P_a(X = 1)$. Specifically, we estimate

$$\frac{1}{P_a(X=1)} \int_C (\mathbb{E}_{P_a}[X|C=c] - \mathbb{E}_{P^*}[X|C=c]) P_a(C=c) dC. \tag{6}$$

Assuming that all datapoints in P^* are generated a single trustworthy agent a^*, the integral term,

$$\int_C (\mathbb{E}_{P_a}[X|C=c] - \mathbb{E}_{P^*}[X|C=c]) P_a(C=c) dC,$$

can be interpreted as the natural direct effect of A of X within the treated population (i.e., data points where $A = a$ are the treated whereas a^* refers to the untreated), when C controls for all mediators and confounders between A and X. We next show how to estimate the NDEE in practice, which is as follows.

Dataset Preparation. We modify the original dataset $\mathcal{D}^*$ such that $\mathcal{D}^* = \{(x_i^*, y_i, c_i, a_i)\}_{i=1}^M$ where $a_i = a^*$ for all a_i. Next, we combine both $\mathcal{D}$ and $\mathcal{D}^*$ to create a unified dataset:

$$\mathcal{D}' = \mathcal{D} \cup \mathcal{D}^*.$$

Causal Effect Estimation Let $\mathcal{F}$ denote a suitable function class and ℓ a loss function. We then learn a function

$$f(c, a) := \arg\min_{f' \in \mathcal{F}} \frac{1}{N+M} \sum_{i: i \in \mathcal{D}'} \ell(f'(c_i, a_i), x_i). \tag{7}$$

Next, we estimate the natural direct effect of A over the treated population as follows:

$$\hat{\tau}_{\text{NDE}} := \frac{1}{N_a} \sum_{i: i \in \mathcal{D}, a_i = a} f(c_i, a) - f(c_i, a^*). \tag{8}$$

Probability Estimation The probability $P_a(X = 1)$ can be estimated simply as

$$\pi_a := \frac{1}{N_a} \sum_{i: i \in \mathcal{D}, a_i = a} x_i \tag{9}$$

The full algorithm is summarized in 3. The model $f(c, a)$ is implemented using XGBoost, where we use the default hyperparameters provided by the XGBoost library in Python [5] (a learning rate of 0.3, a maximume tree depth of 6, and L2 regularization with a coefficient of 1).

Algorithm 3 NDEE algorithm

Input: $\mathcal{D} = \{(x_i, y_i, c_i, a_i)\}_i^N$, $\mathcal{D}^* = \{(x_i^*, y_i, c_i)\}_i^M$, and $\mathcal{D}' = \{(x_i, y_i, c_i, a_i)\}_i^{N+M}$
Output: $\widehat{\text{MR}}$, an estimate of the MR for each agent
  for each agent a **do**
    Estimate $f(c, a)$ using equation 7
    Estimate $\hat{\tau}_{\text{NDE}}$ using equation 8
    Estimate π_a using equation 9
    return $\frac{1}{\pi_a} \hat{\tau}_{\text{NDE}}$
  end for

When can the NDEE accurately estimate the MR? We show that if A does not directly causally effect X^*, it is possible to obtain an accurate estimate of the misreporting rate using the NDEE. To show this, we can rewrite the misreporting rate as follows:

$$\begin{aligned}
\text{MR} &= P_a(X^* = 0 | X = 1) \\
&= \frac{P_a(X = 1) - P_a(X^* = 1)}{P_a(X = 1)} \\
&= \frac{1}{P_a(X = 1)} (\mathbb{E}_{P_a}[X] - \mathbb{E}_{P_a}[X^*]) \\
&= \frac{1}{P_a(X = 1)} \int_C (\mathbb{E}_{P_a}[X | C = c] - \mathbb{E}_{P_a}[X^* | C = c]) P_a(C = c) dC.
\end{aligned}$$

Thus, unless $\mathbb{E}_{P_a}[X^* | C = c] = \mathbb{E}_{P^*}[X^* | C = c]$, the NDEE will give a biased estimate of the misreporting rate. This equality will hold if $X^* \perp A | C$, which can only be true if A does not directly affect X^*. Therefore, we should expect the NDEE to only work in settings where agents do not directly genuinely adapt their features.

F.4 OC-SVM

Under our assumptions, no data points where $X = 0$ are misreported. Therefore, we restrict the OC-SVM approach to a subset of the data from $\mathcal{D}^*$ and $\mathcal{D}$ where $X = 1$, which we denote as $\mathcal{D}_1^*$ and $\mathcal{D}_1$, respectively. To train a One-Class SVM model, we use data from $\mathcal{D}_1^*$, which is assumed to contain no misreported instances. The One-Class SVM model, denoted as $g(y, c)$, is trained to identify outliers/misreported instances using only the variables Y and C. The model outputs a 1 if a datapoint is classified an outlier, and 0 otherwise.

For a given agent a, we estimate the misreporting rate using the OC-SVM as:

$$\hat{\text{MR}} = \frac{1}{N_{a1}} \sum_{i: i \in \mathcal{D}_1, a_i = a} g(y_i, c_i).$$

We use the One-Class SVM implementation from the scikit-learn library [22]. Given our assumption that all data points in $\mathcal{D}_1^*$ are correctly reported, we used a small ν parameter (0.01). Additionally, we use an RBF kernel with a bandwidth parameter $\gamma = 0.1$.

G Additional Experiments

G.1 Medicare Experiments

Figure 5 presents our Medicare experiments including the results from the OC-SVM estimator. The estimated misreporting rate for the OC-SVM is consistent across all HCC codes and agents, reflecting the results from our semi-synthetic loan dataset experiments. This suggests that the OC-SVM is unable to distinguish misreported data points from normal data points.

Tables 1 and 2 provide information about each of the nonpayment and payment HCCs that had $\hat{\delta}'_a > 0$ and were present in at least 1% of the switcher enrollees. For each HCC code, the tables report the estimated MR using CMRE, the number of stayer and switcher enrollees in year t, $\hat{\delta}'_a$, and the lower and upper bounds for the 95% confidence interval. We exclude HCCs that were nonpayment in year t but were used as payment HCCs for the risk adjustment model in year $t + 1$, due to the potential incentive for agents to misreport them.

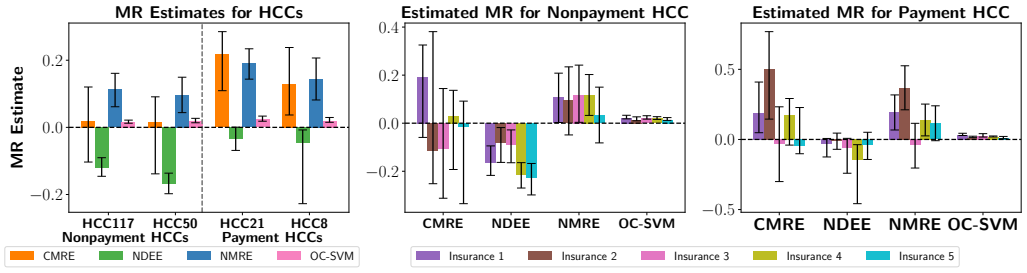

Figure 5: For each plot, the y-axis represents the estimated MR for an HCC code and the error bars represent a 95% confidence interval. **(Left)** The x-axis has two nonpayment HCCs (HCC117 and HCC50) and two payment HCCs (HCC21 and HCC8). Our approach (CMRE) has a MR estimate close to zero for nonpayment HCCs and significantly above zero for the payment HCCs, which aligns with what is expected in current literature. Baselines that fail to distinguish genuine adaptation from strategic adaptation (NDEE) seem to underestimate the MR and baselines that do not control for confounding (NMRE) seem to overestimate the MR. OC-SVM has a similar estimate for each HCC, making it ineffective at identifying misreported data points. **(Middle and Right)** The x-axis represents the baselines. The middle plot represents estimates for HCC50 and the right plot represents MR estimates for HCC8 across different private insurers (agents). Similar to the left plot, NDEE seems to underestimate the MR across most agents, and NMRE overestimates, and the MR estimates for OC-SVM are consistent across all agents and HCC codes.

Table 1: Nonpayment HCCs with $\hat{\delta}'_a > 0.1$ and present in at least 1% of switcher enrollees.

HCC	Full Name	MR	# Stayers	# Switchers	$\hat{\delta}'_a$	LCB	UCB
50	Delirium and Encephalopathy	.015	32294	4535	.175	-.138	.091
117	Pleural Effusion/Pneumothorax	.019	39037	5601	.153	-.103	.120

Table 2: Payment HCCs with $\hat{\delta}'_a > 0.1$ and present in at least 1% of switcher enrollees.

HCC	Full Name	MR	# Stayers	# Switchers	$\hat{\delta}$	LCB	UCB
8	Metastatic Cancer and Acute Leukemia	.130	18762	2646	.276	.037	.238
21	Protein-Calorie Malnutrition	.217	21460	3338	.270	.109	.285
84	Cardio-Respiratory Failure and Shock	.046	40308	6292	.247	-.030	.103
188	Artificial Openings for Feeding or Elimination	.004	10528	1684	.194	-.110	.210
2	Septicemia, Sepsis, SIRS, and Shock	.112	28297	4309	.190	.040	.248
135	Acute Renal Failure	-.004	49021	7868	.130	-.073	.116
103	Hemiplegia/Hemiparesis	.241	12589	2395	.129	.066	.439
86	Acute Myocardial Infarction	.033	20555	3230	.117	-.121	.217

G.2 Loan Dataset Experiments

We conduct additional experiments using alternative versions of the loan fraud dataset to show how well our method and the baselines work under the DAGs defined in Figure 4. We also include two additional baselines that were not in the main paper: NDEE (no C) and NDEE (no S). Unlike the standard NDEE model, which controls for all covariates, NDEE (no C) doesn't control for confounders between X^* and Y, and NDEE (no S) doesn't control for common causes of A and X^*, e.g., S. These variants are used to highlight the importance of controlling for S for NDEE.

G.2.1 Simulation 1

The first simulation replicates the setup used to generate the results in Section 5. It includes four confounders of X^* and Y: education (C_E), sex (C_S), marriage (C_M), and age (C_A). Among these variables, sex and marriage also causally effect A, reflecting a similar scenario represented by the DAG in Figure 4(g). The simulation details are provided below:

$$
\begin{aligned}
A_i &\sim \text{Bernoulli}(0.05 + 0.3(1 - C_{Si}) + 0.3(1 - C_{Mi})),\\
X^*_i &\sim \text{Bernoulli}(0.05 + 0.05C_{Ei} + 0.3C_{Si}C_{Mi} + 0.1{C_A}_i^2 + \beta_A A_i),\\
Y_i &\sim \text{Bernoulli}(0.05 + 0.05C_{Ei} + 0.3C_{Si}C_{Mi} + 0.1{C_A}_i^2 + \beta_{X^*} X^*_i),\\
X_i &\sim X^*_i + A_i(1 - X^*_i)\text{Bernoulli}(\mu),
\end{aligned}
$$

In this simulation, NDEE (no S) doesn't control for either C_S or C_M. Our main method, CMRE, controls for all covariates as they are all confounders between X^* and Y. $\beta_A = 0.3$ and $\beta_{X^*} = 0.4$ unless specified otherwise. The results for this simulation are shown in Figure 6.

G.2.2 Simulation 2

The second simulation includes three confounders of X^* and Y: education (C_E), sex (C_S), and age (C_A). Marriage (C_M) is a common cause of A and X^* and an agent genuinely adapts eduction, reflecting similar scenarios represented by the DAGs in Figure 4(a) and 4(b). In addition, eduction is

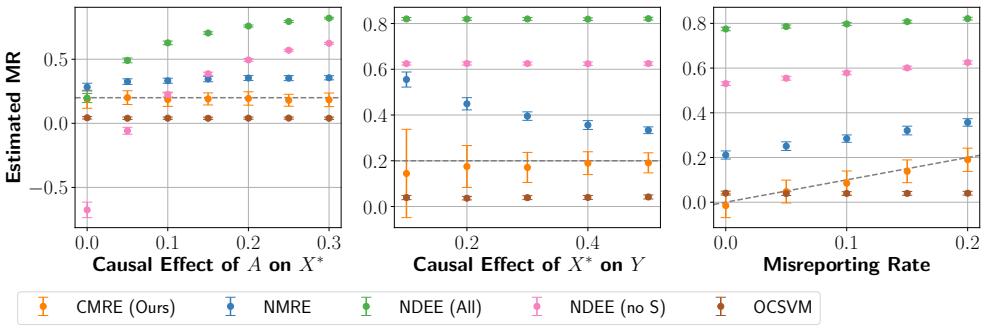

Figure 6: The x-axis is the causal effect of A on X^* **(left)**, causal effect of X^* on Y **(middle)**, and the misreporting rate **(right)**. The y-axis is the estimated misreporting rate. Dashed lines represent the true misreporting rate and the error bars represent the standard deviation. Our approach (CMRE) accurately estimates the MR for all levels of genuine adaptation, the causal effect of X^* on Y, and misreporting rates. The variance for our estimator depends on the magnitude of the causal effect of X^* on Y. Baselines that do not adjust for confounding (NMRE) or do not distinguish between genuine adaptation and misreporting (NDEE) give biased estimates in various cases. NDEE is accurate when there is no genuine adaptation whereas NDEE (no S) is not, highlighting the need for controlling for common causes of A and X^*. Anomaly detection methods (OC-SVM) are unable to distinguish misreported data points from unmanipulated data points.

also a treatment effect modifier. The simulation details are provided below:

$$
\begin{aligned}
A_i &\sim \text{Bernoulli}(0.05 + 0.4(1 - C_{Mi})),\\
C'_{Ei} &\sim C_{Ei} + (1 - C_{Ei})A_i\text{Bernoulli}(\beta_M),\\
X^*_i &\sim \text{Bernoulli}(0.05 + 0.25C_{Mi} + 0.1C'_{Ei}C_{Si} + 0.1{C_A}_i^2 + \beta_A A_i),\\
Y_i &\sim \text{Bernoulli}(0.05 + 0.2C'_{Ei}C_{Si} + 0.1{C_A}_i^2 + (\beta_{X^*} + 0.1C'_{Ei})X^*_i),\\
X_i &\sim X^*_i + A_i(1 - X^*_i)\text{Bernoulli}(\mu),
\end{aligned}
$$

In this simulation, NDEE (no S) doesn't control for C_M and NDEE (no C) only controls for C_M. Our main method, CMRE, only controls for C_E, C_S, and C_A. $\beta_A = 0.1$, $\beta_M = 0.2$, and $\beta_{X^*} = 0.4$ unless specified otherwise. The results for this simulation are shown in Figure 7.

G.2.3 Simulation 3

The third simulation includes three confounders of X^* and Y: education (C_E), sex (C_S), and age (C_A). Marriage (C_M) is a common cause of A and Y and an agent genuinely adapts education, reflecting the scenario represented by the DAG in Figure 4(c). In addition, education is also a treatment effect modifier. $\beta_A = 0.1$, $\beta_M = 0.2$, and $\beta_{X^*} = 0.4$ unless specified otherwise. The simulation details are provided below:

$$
\begin{aligned}
A_i &\sim \text{Bernoulli}(0.05 + 0.4(1 - C_{Mi})),\\
C'_{Ei} &\sim C_{Ei} + (1 - C_{Ei})A_i\text{Bernoulli}(\beta_M),\\
X^*_i &\sim \text{Bernoulli}(0.05 + 0.1C'_{Ei}C_{Si} + 0.1{C_A}_i^2 + \beta_A A_i),\\
Y_i &\sim \text{Bernoulli}(0.05 + 0.2C_{Mi} + 0.1C'_{Ei}C_{Si} + 0.05{C_A}_i^2 + (\beta_{X^*} + 0.1C'_{Ei})X^*_i),\\
X_i &\sim X^*_i + A_i(1 - X^*_i)\text{Bernoulli}(\mu),
\end{aligned}
$$

In this simulation, NDEE (no C) only controls for C_M. Our main method, CMRE, only controls for C_E, C_S, and C_A. The results for this simulation are shown in Figure 8.

G.2.4 Simulation 4

The fourth simulation includes two confounders of X^* and Y: sex (C_S), and age (C_A). Marriage (C_M) is a common cause of A and Y and an agent genuinely adapts education, which is a mediator

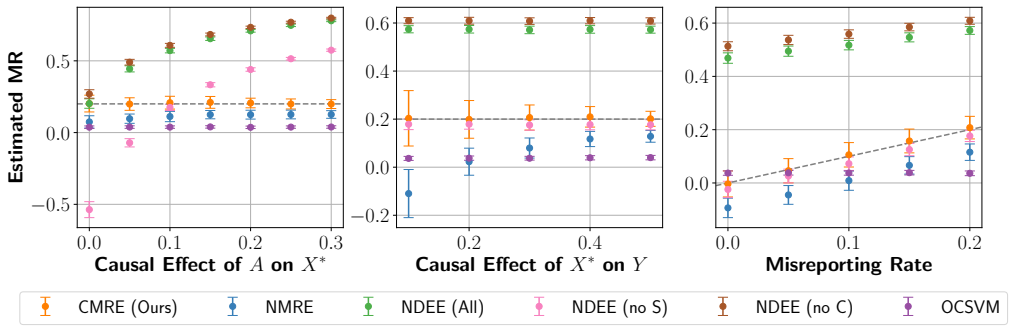

Figure 7: The x-axis is the direct causal effect of A on X^* **(left)**, causal effect of X^* on Y **(middle)**, and the misreporting rate **(right)**. The y-axis is the estimated misreporting rate. Dashed lines represent the true misreporting rate and the error bars represent the standard deviation. Our approach (CMRE) accurately estimates the MR for all levels of genuine adaptation, the causal effect of X^* on Y, and misreporting rates. The variance for our estimator depends on the magnitude of the causal effect of X^* on Y. Baselines that do not adjust for confounding (NMRE) or do not distinguish between genuine adaptation and misreporting (NDEE) give biased estimates in various cases. NDEE is accurate when there is no genuine adaptation whereas NDEE (no S) and NDEE (no C) are not, as they either don't control for common causes of A and X^* or mediators of A and X^*. Anomaly detection methods (OC-SVM) are unable to distinguish misreported data points from unmanipulated data points.

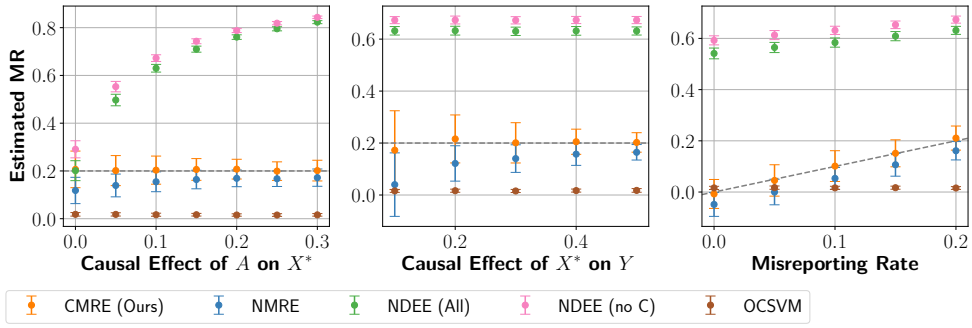

Figure 8: The x-axis is the direct causal effect of A on X^* **(left)**, causal effect of X^* on Y **(middle)**, and the misreporting rate **(right)**. The y-axis is the estimated misreporting rate. Dashed lines represent the true misreporting rate and the error bars represent the standard deviation. Our approach (CMRE) accurately estimates the MR for all levels of genuine adaptation, the causal effect of X^* on Y, and misreporting rates. The variance for our estimator depends on the magnitude of the causal effect of X^* on Y. Baselines that do not adjust for confounding (NMRE) or do not distinguish between genuine adaptation and misreporting (NDEE) give biased estimates in various cases. NDEE is accurate when there is no genuine adaptation whereas NDEE (no C) is not, as it doesn't account for the mediators of A and X^*. Anomaly detection methods (OC-SVM) are unable to distinguish misreported data points from unmanipulated data points.

of A and X^*, reflecting the scenario represented by the DAGs in Figure 4(d) and 4(e). The simulation details are provided below:

$$
\begin{aligned}
A_i &\sim \text{Bernoulli}(0.05 + 0.4(1 - C_{Mi})),\\
C'_{Ei} &\sim C_{Ei} + (1 - C_{Ei})A_i\text{Bernoulli}(\beta_M),\\
X^*_i &\sim \text{Bernoulli}(0.05 + 0.3C'_{Ei}C_{Si} + 0.1C_{A_i}^2 + \beta_A A_i),\\
Y_i &\sim \text{Bernoulli}(0.05 + 0.2C_{Mi} + 0.1C_{Si} + 0.05C_{A_i}^2 + \beta_{X^*}X^*_i),\\
X_i &\sim X^*_i + A_i(1 - X^*_i)\text{Bernoulli}(\mu),
\end{aligned}
$$

In this simulation, NDEE (no C) only controls for C_M and C_E. Our main method, CMRE, only controls for C_S and C_A. $\beta_A = 0.1$, $\beta_M = 0.2$, and $\beta_{X^*} = 0.4$ unless specified otherwise. The results for this simulation are shown in Figure 9.

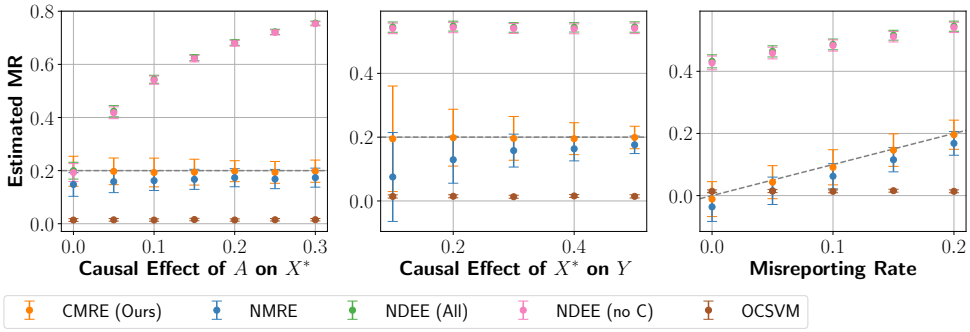

Figure 9: The x-axis is the direct causal effect of A on X^* **(left)**, causal effect of X^* on Y **(middle)**, and the misreporting rate **(right)**. The y-axis is the estimated misreporting rate. Dashed lines represent the true misreporting rate and the error bars represent the standard deviation. Our approach (CMRE) accurately estimates the MR for all levels of genuine adaptation, the causal effect of X^* on Y, and misreporting rates. The variance for our estimator depends on the magnitude of the causal effect of X^* on Y. Baselines that do not adjust for confounding (NMRE) or do not distinguish between genuine adaptation and misreporting (NDEE) give biased estimates in various cases. Both NDEE and NDEE (no C) are accurate when there is no genuine adaptation, as they control for all common causes of A and X^* and mediators of A and X^*. Anomaly detection methods (OC-SVM) are unable to distinguish misreported data points from unmanipulated data points.

G.2.5 Simulation 5

The fifth simulation includes two confounders of X^* and Y: sex (C_S), and age (C_A). Marriage (C_M) is a common cause of A and X^* and an agent genuinely adapts education, which is a mediator of A and X^*, reflecting similar scenarios represented by the DAGs in Figure 4(d) and 4(f). The simulation details are provided below:

$$
\begin{aligned}
A_i &\sim \text{Bernoulli}(0.05 + 0.4(1 - C_{Mi})),\\
C'_{Ei} &\sim C_{Ei} + (1 - C_{Ei})A_i\text{Bernoulli}(\beta_M),\\
X^*_i &\sim \text{Bernoulli}(0.05 + 0.2C_{Mi} + 0.3C'_{Ei}C_{Si} + 0.1{C_A}_i^2 + \beta_A A_i),\\
Y_i &\sim \text{Bernoulli}(0.05 + 0.3C_{Si} + 0.05{C_A}_i^2 + \beta_{X^*}X^*_i),\\
X_i &\sim X^*_i + A_i(1 - X^*_i)\text{Bernoulli}(\mu),
\end{aligned}
$$

In this simulation, NDEE (no C) only controls for C_M and C_E and NDEE (no S) doesn't control for C_M. Our main method, CMRE, only controls for C_S and C_A. $\beta_A = 0.1$, $\beta_M = 0.2$, and $\beta_{X^*} = 0.4$ unless specified otherwise. The results for this simulation are shown in Figure 10.

H Software and Hardware

H.1 Software

All of the code for the experiments was written in Python 3.10.16 (PSF License). The XGBoost models were implemented using the XGBoost 2.1.4 (Apache License 2.0) [5]. The OC-SVM baseline was implemented by using scikit-learn 1.6.1 (BSD License) [22], which used the implementation of the One-Class SVM. To generate the semi-synthetic datasets and for data processing tasks, both numpy 2.0.2 (modified BSD license) [11] and pandas 2.2.3 (BSD license) [21] were employed. For the Medicare dataset, HCCPy 0.1.9 (Apache License 2.0) was employed to extract the HCCs from raw data. All plots were created using matplotlib 3.10.1 (PSF License) [15].

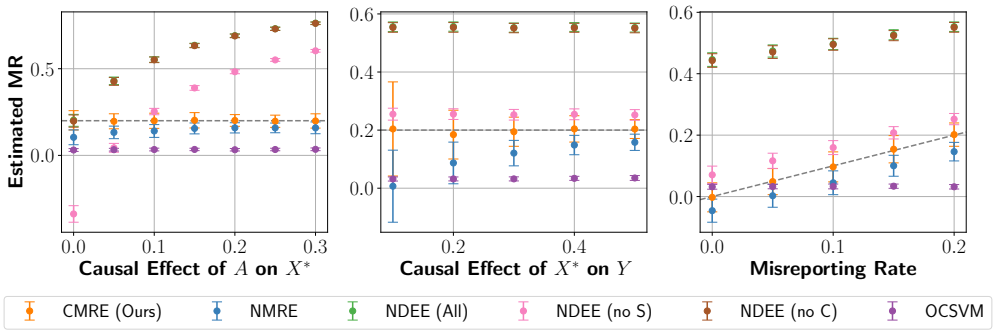

Figure 10: The x-axis is the direct causal effect of A on X^* **(left)**, causal effect of X^* on Y **(middle)**, and the misreporting rate **(right)**. The y-axis is the estimated misreporting rate. Dashed lines represent the true misreporting rate and the error bars represent the standard deviation. Our approach (CMRE) accurately estimates the MR for all levels of genuine adaptation, the causal effect of X^* on Y, and misreporting rates. The variance for our estimator depends on the magnitude of the causal effect of X^* on Y. Baselines that do not adjust for confounding (NMRE) or do not distinguish between genuine adaptation and misreporting (NDEE) give biased estimates in various cases. Both NDEE and NDEE (no C) are accurate when there is no genuine adaptation, as they control for all common causes of A and X^* and mediators of A and X^*. In contrast, NDEE does not control for common causes of A and X^*, which makes it biased. Anomaly detection methods (OC-SVM) are unable to distinguish misreported data points from unmanipulated data points.

H.2 Hardware

All experiments were conducted using 16 CPU cores and 32 GB of memory on a computing cluster with 2 x 2.5 GHz Intel Haswell (Xeon E5-2680v3) processors, which was managed using a Slurm resource manager. The simulations for all of the five semi-synthetic loan experiments took approximately 5 hours to complete, whereas the experiments over the Medicare dataset took approximately 36 hours to complete.