Estimating Upcoding in Medicare Advantage: Identifying Contributors, Costs, and Mechanisms

Dylan Zapzalka DYLANZ@UMICH.EDU
Computer Science and Engineering Division
University of Michigan
Ann Arbor, Michigan, United States

Muskaan Mittal MOKAMITTAL@GMAIL.COM
Computer Science and Engineering
University of Michigan
Ann Arbor, Michigan, United States

Jenna Wiens WIENSJ@UMICH.EDU
Computer Science and Engineering
University of Michigan
Ann Arbor, Michigan, United States

Maggie Makar MMAKAR@UMICH.EDU
Computer Science and Engineering
University of Michigan
Ann Arbor, Michigan, United States

Abstract

Strategic misreporting – the manipulation of features to obtain a better outcome – undermines the integrity of machine learning models used for healthcare resource allocation. In the U.S. Medicare Advantage program, private insurers are reimbursed based on the beneficiary diagnoses they report, creating financial incentives that encourage over-reporting, leading to billions of dollars in annual overpayments. Despite the magnitude of the problem, estimating insurers' misreporting rates is problematic due to the lack of ground-truth diagnoses. Past work has used causality to try to estimate misreporting rates, but assumes access to an unmanipulated dataset and fails to account for unobserved confounders. To mitigate these issues, we introduce the Stitched Causal Misreporting Estimator (SCaMEr), a causally motivated auditing approach to estimate insurer-specific misreporting rates. SCaMEr addresses two key limitations of prior work: (1) it avoids the need for an unmanipulated reference dataset, and (2) it gives reliable upper and lower bounds on misreporting estimates in the presence of hidden confounders. To overcome the lack of an unmanipulated reference, SCaMEr "stitches" multiple datasets together with varying misreporting rates to recover unbiased estimates. To account for unobserved confounding, it incorporates causal sensitivity analysis to produce uncertainty bounds. We validate SCaMEr on both semi-synthetic and real-world Medicare Advantage data, where it achieves lower estimation error than baselines. Our results show that SCaMEr can enable auditing by identifying diagnoses, insurance plans, and reporting mechanisms that are susceptible to misreporting.

1. Introduction

When machine learning (ML) models determine important outcomes for individuals or organizations, which we refer to as agents, those agents are incentivized to adapt the data used to make predictions to secure favorable results (Hardt et al., 2016). This often manifests as strategic misreporting, where agents manipulate reported features to influence the model's output (Barsotti et al., 2022; Chang et al., 2024; Zapzalka et al., 2025). Such behavior is particularly prevalent in healthcare settings, where models are leveraged to determine the allocation of scarce resources, such as financial reimbursements (Cook and Averett, 2020).

A primary example of this phenomenon, and the focus of this work, occurs within the U.S. Medicare Advantage (MA) program, where a risk-adjustment model determines insurer payments based on reported beneficiary diagnoses (Chernew et al., 2026). While the model is used to mitigate adverse selection by disincentivizing the "cherry-picking" of healthy beneficiaries by insurers, the model's reliance on diagnoses inadvertently incentivizes upcoding: the strategic misreporting of diagnoses to inflate reimbursement (Geruso and Layton, 2020). The scale of this issue is immense, with upcoding contributing to an estimated $40 billion in overpayments in 2025 alone (MedPAC, 2025) and a proposed $100 million in auditing efforts in 2026 (CMS, 2025a). Recently, it has been proposed to audit *all* MA contracts for each payment year (CMS, 2025b); however, doing so may be both costly and inefficient.

This setting presents a natural opportunity for ML: by identifying misreporting insurers and estimating insurer-specific misreporting rates at scale, ML-based approaches can enable efficient and reliable targeted audits. However, such a task is challenging due to two main factors. First, observed variations in diagnosis rates may reflect true health disparities rather than manipulation. These differences can arise from *insurance-based confounding*, where sicker beneficiaries systematically enroll in specific plans, or from *genuine adaptation*, where insurers alter a beneficiary's health through interventions like increasing or decreasing preventative care (Agarwal et al., 2021). Second, the absence of ground-truth labels for misreported diagnoses prevents the problem from being framed as a standard supervised classification task, where individual data points are categorized as either truthful or fraudulent.

Prior work has framed the estimation of misreporting rates as a causal inference problem (Chang et al., 2024; Zapzalka et al., 2025). Zapzalka et al. (2025) address both insurance-based confounding and genuine adaptation by exploiting a key asymmetry: unlike genuine adaptation and insurance-based confounding, misreporting does not causally affect downstream variables (anchors). Consequently, misreporting leads to biased causal effect estimates between the potentially misreported feature and its descendants. Their approach leverages this asymmetry by comparing causal effect estimates computed from manipulated versus unmanipulated datasets to infer misreporting rates. For example, falsely reporting an HIV diagnosis will not lead to antiretroviral therapy initiation, whereas a true diagnosis will, creating detectable discrepancies in estimated effects across datasets that can be used to infer misreporting.

However, this approach relies on two restrictive assumptions. First, it requires an unmanipulated, misreporting-free reference population. Zapzalka et al. (2025) use Traditional Medicare (TM) data for this purpose, relying on the notion that its payment structure is not tied to risk adjustment, and hence, there is no incentive to misreport. In practice, TM

data likely suffers from downcoding, e.g., underreporting due to a lack of documentation incentives, which would lead to biased estimates (Frogner et al., 2011). Consequently, several researchers argue that comparing MA to TM can overestimate MA upcoding by failing to account for downcoding (Thorpe et al., 2019; DeBusk et al., 2024). Second, their method assumes no unobserved confounding between diagnoses and anchors, which is not statistically testable within observable data, potentially leading to biased estimates of misreporting.

To address these limitations, we present the Stitched Causal Misreporting Estimator (SCaMEr). Similar to Zapzalka et al. (2025), SCaMEr works by comparing causal effect estimates, but overcomes the limitations of prior work through two key advancements. **(1) Data-Stitching:** Instead of assuming access to a perfectly documented reference population, we introduce data-stitching. Data-stitching combines datasets that may have opposing reporting biases, such as upcoding in one and downcoding in another, to estimate unbiased misreporting rates. **(2) Sensitivity Analysis:** We relax the assumption of "no unobserved confounding" via sensitivity analysis. This allows us to obtain reliable bounds on the misreporting rate, even when some confounding factors remain unobserved.

By identifying the diagnoses and insurers most susceptible to misreporting, our approach provides a precise targeting tool for auditors seeking to recover overpayments. Furthermore, our evaluation of specific reporting modalities, such as chart reviews, uncovers the mechanisms insurers exploit to systematically misreport diagnoses.

Our contributions are summarized as follows: **(1)** we introduce a novel method to estimate misreporting rates by "stitching" together disparate, potentially misreported datasets (e.g., from populations subject to only upcoding or only downcoding); **(2)** provide theoretical guarantees for our estimator, establishing identifiability conditions; **(3)** propose conditional independence tests to empirically determine when multiple features cause the same anchor variables, which can lead to unobserved confounding; **(4)** incorporate causal sensitivity analysis to quantify uncertainty, providing upper and lower bounds on misreporting rates in the presence of unobserved confounding; **(5)** validate SCaMEr through experiments on semi-synthetic and real-world Medicare data, demonstrating that it consistently outperforms existing baselines in accuracy and robustness.

Generalizable Insights about Machine Learning in the Context of Healthcare

Resource allocation models are often deployed in environments where agents are incentivized to manipulate clinical data to secure favorable outcomes. While motivated by and validated on Medicare Advantage, SCaMEr is a general method applicable to other resource-constrained clinical settings shaped by directional misreporting incentives.

Our method provides a way to audit such systems and quantify misreporting rates without requiring a perfectly documented reference population. By leveraging datasets with opposing reporting biases, researchers can identify misreporting even without an unmanipulated dataset. Furthermore, the integration of sensitivity analysis offers a robust way to establish bounds on misreporting in the presence of unobserved confounding factors. These techniques can facilitate the development of more equitable ML models by identifying which features are most susceptible to manipulation. This allows for the selection of more robust clinical features that prioritize beneficiaries' needs over an agent's ability to misreport data.

2. Related Work

Strategic Classification and Regression. Our work aligns with work done in strategic classification, which seeks to develop models that are robust to agents who manipulate features to secure favorable outcomes (Hardt et al., 2016; Levanon and Rosenfeld, 2021). While our method can improve model robustness by identifying and removing susceptible features, our primary objective is to quantify feature misreporting. Recent causal perspectives on strategic classification distinguish improvement (changing causally relevant features) from gaming (changing non-causal features) (Miller et al., 2020; Horowitz and Rosenfeld, 2023). Our work instead focuses on misreporting, which differs from both improvement and gaming, as agents only change their reported features, not their true underlying values.

Sensitivity Analysis Our method can incorporate causal sensitivity analysis methods, such as those introduced by VanderWeele and Arah (2011), Imbens (2003), and Yadlowsky et al. (2022). However, our application is distinct from prior sensitivity analysis work in that, while traditional sensitivity analysis methods aim to bound causal effect estimates (e.g., the average treatment effect), we adapt these tools to bound the misreporting rate.

Misreporting Detection We build upon the method established by Zapzalka et al. (2025) to identify misreporting rates. We relax their core requirements by demonstrating that identification is possible without access to an unmanipulated reference dataset. Furthermore, we introduce sensitivity analysis to provide upper and lower bounds on misreporting rates when unobserved confounding is present. While other causal approaches, such as Chang et al. (2024), attempt to measure misreporting, they only rank agents by their propensity to misreport. Additionally, unlike our method, their framework does not distinguish between genuine modification and misreporting.

3. Preliminaries

3.1. Setup

We consider a setting where a ML model uses N binary features to make a prediction. For each feature i, we denote the observed reported value as X_i and the latent true value as X_i^*. We assume an agent $A = a$ is incentivized toward directional misreporting: either upcoding, where an agent reports $X_i = 1$ when $X_i^* = 0$, or downcoding, where an agent reports $X_i = 0$ when $X_i^* = 1$. We assume that agents may genuinely adapt X_i^* in either direction. We aim to quantify the degree to which an agent misreports a specific feature; while we focus on detecting upcoding, the analysis extends trivially to downcoding. Throughout this work, uppercase letters denote random variables and lowercase letters denote their realizations. Let $P_a(V) := P(V|A = a)$ denote the distribution of agent a. Our primary estimand is:

Definition 1 (Misreporting Rate) $MR_i(a) = P_a(X_i^* = 0|X_i = 1)$

That is, the misreporting rate (MR) is the probability that some reported feature is a false positive. Alternative estimators are detailed in Appendix D. To estimate the misreporting rate, we assume access to some M causal descendants Y_i (referred to as anchors) and covariates C, which represent confounders between X_i^* and Y_i and treatment effect modifiers. We assume access to a dataset $\mathcal{D} = \{(c^{(j)}, a^{(j)}, x_1^{(j)}, ..., x_N^{(j)}, y_1^{(j)}, ..., y_M^{(j)})\}_{j=1}^n$ drawn

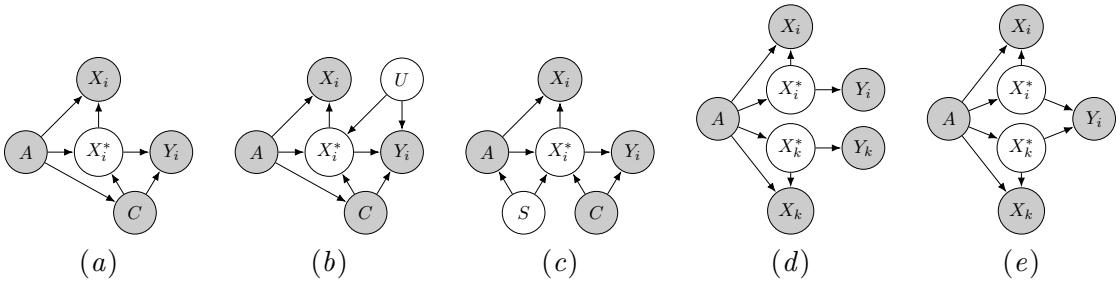

Figure 1: Causal DAGs that describe the assumptions of our settings. Shaded nodes denote observed variables, whereas white nodes denote latent variables. **(a) Observed Confounding:** All confounders C between the true feature X_i^* and anchor Y_i are observed. **(b) Unobserved Confounding:** Presence of an unobserved confounder U between X_i^* and Y_i. **(c) Insurance-based Confounding:** There exists some confounding between the agent and the true features due to sicker patients preferring different plans. **(d) Multiple Misreported Features With Unique Anchors:** Distinct anchors for each feature lead to identifiability of the MR. **(e) Multiple Misreported Features With a Shared Anchor:** Shared anchors necessitate sensitivity analysis.

from a distribution P consistent with the causal DAGs illustrated in Figure 1, where X_i^* is unobserved. We focus the majority of our analysis on the DAGs depicted in Figures 1a, 1c, and 1d, where the confounders between X_i^* and Y_i are fully observed and there are unique anchors. We will also address settings where there is unobserved confounding (Figure 1b) and cases where descendants Y_i may be influenced by multiple features (Figure 1e).

3.2. Assumptions

We assume that while other variables may be continuous, the features X_i and X_i^* are binary and subject to monotonic misreporting, e.g., a given agent can only upcode or downcode:

Assumption 1 (Directional Misreporting) *For all a, either $P_a(X_i = 1|X_i^* = 1) = 1$ or $P_a(X_i = 0|X_i^* = 0) = 1$.*

This aligns with Medicare, where the incentive structure (e.g., payments) may encourage either upcoding or downcoding. This assumption of directional misreporting is grounded in literature detailing the program's systemic financial incentives (DeBusk et al., 2024). We also assume that agents are rational in the sense that they only manipulate features that directly influence the model's output:

Assumption 2 (Useful Modifications) *For $i = 1, ..., N$, agents may only misreport X_i^*, or genuinely adapt it by intervening on X_i^* or its ancestors.*

Our approach for estimating the MR will rely on causality. To that end, we adopt the Neyman-Rubin potential outcomes framework (Rubin, 2005) to define our causal quantities. Let $Y_i(x_i^*)$ denote the counterfactual value of the anchor Y_i had the true feature X_i^* been set to x_i^*. We require the following standard identification assumptions:

Assumption 3 (Overlap) *For $i = 1, ..., N$, $P_a(X_i^* = x_i^* | C = c) > 0$ and $P_a(X_i = x_i | C = c) > 0$ for all x_i^*, x_i, and c.*

Assumption 4 (Consistency) *For all $i = 1, ..., N$, $j = 1, ..., n$, and $k = 1, ..., M$, $y_k^{(j)}(x_i^*) = y_k^{(j)}$ if ${x^*}_i^{(j)} = x_i^*$*

Assumption 5 (No unmeasured confounding) $Y_i(X_i^* = 0), Y_i(X_i^* = 1) \perp X_i^* \mid C$

We later relax Assumption 5 in cases where there are unobserved confounders due to unobserved clinical factors U (Figure 1b) or due to multiple features affecting the same anchor variable (Figure 1e). Finally, we assume that the causal relationship between X_i^* and Y_i is stable across different agents.

Assumption 6 $\mathbb{E}_{P_a}[Y_i(X_i^* = x) \mid C = c] = \mathbb{E}_{P_{a'}}[Y_i(X_i^* = x) \mid C = c] \ \forall c, a, a', x$

In other words, Assumption 6 states that the conditional average treatment effect (CATE) of a true diagnosis on its anchor is invariant across plans. In the Medicare context, this implies the effect of a diagnosis (e.g., HIV) on its anchor (e.g., antiretroviral prescriptions) is the same regardless of the insurance plan.

This assumption could be violated if unobserved effect modifiers (e.g., Body Mass Index) differ between MA and TM populations and systematically influence treatment decisions. Importantly, this does not assume no distribution shift of covariates between MA and TM, only CATE invariance between the two populations.

3.3. Prior work: Zapzalka et al. (2025)

Because X_i^* is latent, the misreporting rate cannot be calculated directly. Instead, we leverage the observed anchors Y_i. Following Zapzalka et al. (2025), we define three key estimands that characterize the relationship between features and anchors:

1. **True Average Feature Effect on the Reported (TAFR):** The causal effect of the true feature X_i^* on Y_i over the population that reported $X_i = 1$:

$$\tau_i^*(a) := \int_C (\mathbb{E}_{P_a}[Y_i(X_i^* = 1) - Y_i(X_i^* = 0) | C = c]) P_a(C = c | X_i = 1) dc.$$

2. **Nominal Average Feature Effect on the Reported (NAFR):** The potentially biased effect of X_i on Y_i over the population that reported $X_i = 1$:

$$\tau_i(a) := \int_C (\mathbb{E}_{P_a}[Y_i | X_i = 1, C = c] - \mathbb{E}_{P_a}[Y_i | X_i = 0, C = c]) P_a(C = c | X_i = 1) dc.$$

3. **True Average Feature Effect on the Misreported (TAFM):** The causal effect of the true feature X_i^* on Y_i over the misreported population:

$$\delta_i^*(a) := \int_C (\mathbb{E}_{P_a}[Y_i(X_i^* = 1) - Y_i(X_i^* = 0) | C = c]) P_a(C = c | X_i^* = 0, X_i = 1) dc.$$

Any discrepancy between the TAFR and NAFR indicates that an agent misreported. Since misreporting renders X_i a noisy proxy for X_i^*, estimating the NAFR in place of the TAFR will result in attenuation bias. The magnitude of this bias is dictated by the TAFM. As established in prior work, these quantities allow us to identify the misreporting rate, with

$$P_a(X_i^* = 0 | X_i = 1) = \frac{\tau_i^*(a) - \tau_i(a)}{\delta_i^*(a)}. \tag{1}$$

However, a fundamental identification challenge remains: $\tau_i^*(a)$ and $\delta_i^*(a)$ depend on the latent X_i^* and are unidentifiable from a single agent's data. Zapzalka et al. (2025) circumvented this by assuming access to data from an agent that doesn't misreport. However, this assumption may not hold in clinical settings such as Medicare, where all available data may be subject to some form of misreporting. They also relied on the unverifiable assumption that there was no hidden confounding between X_i^* and Y_i. In the next section, we propose a method to recover these quantities without requiring an unmanipulated reference population, and we address hidden confounding by applying causal sensitivity analysis.

4. Methods

We now introduce the Stitched Causal Misreporting Estimator (SCaMEr). We first tackle the setting where there is no hidden confounding and then relax that assumption in Section 4.2 and Section 4.3 with sensitivity analysis.

To address the identification challenges posed by unobserved ground-truth diagnoses, we propose *data-stitching*. This approach enables the estimation of $\tau_i^*(a)$ and $\delta_i^*(a)$ without an unmanipulated dataset by combining data from agents with complementary reporting biases. For simplicity, we consider two agents: $\overline{a}$, who does not downcode (e.g., MA), and $\underline{a}$, who does not upcode (e.g., TM), and focus on estimating the misreporting rate for $\overline{a}$. Our approach trivially extends to multiple agents and can estimate misreporting rates for any agent. We define the "stitched" estimands, which are identifiable from observable data (see Appendix A for a derivation):

$$\tau_i'(\overline{a}, \underline{a}) := \int_C (\mathbb{E}_{P_{\underline{a}}}[Y(X_i^* = 1) | C = c] - \mathbb{E}_{P_{\overline{a}}}[Y(X_i^* = 0) | C = c]) P_{\overline{a}}(C = c | X_i = 1) dc, \text{ and}$$

$$\delta_i'(\overline{a}, \underline{a}) := \int_C (\mathbb{E}_{P_{\underline{a}}}[Y(X_i^* = 1) | C = c] - \mathbb{E}_{P_{\overline{a}}}[Y(X_i^* = 0) | C = c]) P_{\overline{a}}(C = c | X_i = 0) dc$$

As we show formally in Theorem 1, τ_i' and δ_i' are unbiased estimators for τ_i^* and δ_i^*, which allows for estimating the misreporting rate using equation 1. The validity of these estimands relies primarily on the directional misreporting assumption (Assumption 1). Because $\underline{a}$ never upcodes, the observed group $X_i = 1$ under $P_{\underline{a}}$ is a trustworthy sample where $X_i^* = 1$. Conversely, because $\overline{a}$ never downcodes, the observed group $X_i = 0$ under $P_{\overline{a}}$ is a trustworthy sample where $X_i^* = 0$. By "stitching" these reliable subsets together, we effectively construct a new dataset where the latent feature X_i^* is known. This "stitched" dataset will allow us to identify the misreporting rate as follows.

Theorem 1 (Identifiability via data-stitching) *Let Assumptions 1-6 hold. Suppose $\overline{a}$ does not downcode and $\underline{a}$ does not upcode. Also, let P follow any of the DAGs in Figures 1a,*

1c, or 1d. Then for $|\delta_i'(\overline{a}, \underline{a})| > 0$, the MR is identifiable and can be expressed as:

$$P_{\overline{a}}(X_i^* = 0 | X_i = 1) = \frac{\tau_i'(\overline{a}, \underline{a}) - \tau_i(\overline{a})}{\delta_i'(\overline{a}, \underline{a})}$$

The proof for Theorem 1 (Appendix A) utilizes the invariance of causal effects (Assumption 6) to establish that $\tau_i^*(\overline{a}) = \tau_i'(\overline{a}, \underline{a})$ and $\delta_i^*(\overline{a}) = \delta_i'(\overline{a}, \underline{a})$: if the causal effect of X_i^* on Y_i is stable across agents, we can estimate it using a "stitched" dataset from multiple agents.

4.1. Identification Limits and Confounding

While Theorem 1 yields a point estimate in many settings, it has two main limitations. **(1) Latent Confounding:** With unobserved confounders U (Figure 1b), we will be unable to measure the causal effect of X_i^* on Y_i. **(2) Cross-Feature Interference:** In Figure 1e, when multiple features influence a single anchor, e.g., X_i^*, X_k^* affect Y_i, we need to control for X_k^* when estimating the causal effect of X_i^* on Y_i. Since X_k^* itself is unobserved, we cannot control for it. This creates unobserved confounding between X_i^* and Y_i identical to latent confounding through U. [1] While we can in principle "stitch" the data with respect to both X_i^* and X_k^*, this is an impractical solution since it would severely reduce the sample size when X_k^* is high dimensional.

To address this, we propose applying sensitivity analysis to provide upper and lower bounds on the misreporting rate due to unobserved confounding, whether through an unobserved U or through cross-feature interference. We also develop conditional independence tests to detect cross-feature interference, indicating that sensitivity analysis should be used.

Beyond the risk of confounding bias, the choice of anchor also impacts the efficiency of our estimator. Specifically, anchors that do not have a strong causal relationship with the true diagnosis will increase the variance of the estimated misreporting rate, as noted in prior work Zapzalka et al. (2025). Therefore, selecting a good anchor requires balancing a strong causal signal (to minimize variance) against the risk of cross-feature interference (to minimize bias).

4.2. Detecting Cross-Feature Interference

We first introduce a conditional independence test to detect cross-feature interference:

Proposition 1 (Conditional Independence Test) *Assume X_i^* causes Y_i, assume that A causes X_j and X_j^*, and assume that X_j^* causes X_j for all $j \in [M]$. Finally, let $X' = \{X_j : j \in [M] \setminus \{i\}\}$, assume P is faithful, and assume that there is no unmeasured confounding between all X_j^* and Y_i, i.e., $Y_i(X_j^* = 0), Y_i(X_j^* = 1) \perp X_j^* | C$ for all $j \in [M]$. Then*

$$X' \perp Y_i | X_i^*, C \implies X_i^* \text{ is the unique causal parent of } Y_i \text{ among all features } X_j^*.$$

This follows from the d-separation properties in Figure 1e. If the conditional independence test fails, it indicates Y_i is influenced by multiple features. In such cases, the best we can do

1. Note that in principle we could control for A to block the open pathway through X_k^*. However, because our data-stitching relies on subsets where $X_i = X_i^*$, we lack the "overlap" required to control for A, i.e., we cannot observe both states of X_i for the same agent without introducing some bias.

is partial identification: sensitivity analysis is necessary to bound the misreporting rate. For our implementation, we use Kernel-based Conditional Independence Test (KCIT) (Zhang et al., 2012) to test if $X' \perp Y_i | X_i^*, C$. We note that KCIT scales poorly to large datasets; extending our approach to more scalable tests is left to future work.

4.3. Sensitivity Analysis

While our method point-identifies the misreporting rate when all assumptions hold, the no unobserved confounding assumption is unverifiable in clinical data. When unobserved confounders are suspected, or when our test indicates cross-feature interference, point identification of the exact misreporting rate is no longer feasible. Instead, we must reframe the problem: given a hypothesized strength for this unobserved confounding, what are the plausible upper and lower bounds on the misreporting rate?

To answer this, recall that our estimator relies on three causal effect estimates: $\tau'(\overline{a}, \underline{a})$, $\tau(\overline{a})$, and $\delta'(\overline{a}, \underline{a})$, where we omit the subscript i for notational simplicity. By applying standard causal sensitivity analysis to these individual components under a specified confounding strength, we can derive valid upper and lower bounds for each. We can then plug in these component-wise upper and lower bounds into our estimator to yield an interval for the misreporting rate, as formalized in Theorem 2:

Theorem 2 (Bounds via sensitivity intervals) *Suppose a sensitivity analysis method provides valid bounds for $\tau'(\overline{a}, \underline{a})$, $\tau(\overline{a})$, and $\delta'(\overline{a}, \underline{a})$, i.e.,*

$$\tau'_L \leq \tau'(\overline{a}, \underline{a}) \leq \tau'_U, \quad \tau_L \leq \tau(\overline{a}) \leq \tau_U, \text{ and } \quad \delta'_L \leq \delta'(\overline{a}, \underline{a}) \leq \delta'_U.$$

For $|\delta'(\overline{a}, \underline{a})| \geq \epsilon$ with $\epsilon > 0$, let $\Delta = \{\delta'_L, \delta'_U, -\epsilon, \epsilon\} \cap [\delta'_L, \delta'_U] \cap \{\tilde{\delta}' : |\tilde{\delta}'| \geq \epsilon\}$. Then we have that the misreporting rate, $\frac{\tau'-\tau}{\delta'}$ is bounded as:

$$\min_{\substack{\tilde{\tau}' \in \{\tau'_L, \tau'_U\} \\ \tilde{\tau} \in \{\tau_L, \tau_U\} \\ \tilde{\delta}' \in \Delta}} \frac{\tilde{\tau}' - \tilde{\tau}}{\tilde{\delta}'} \leq \frac{\tau' - \tau}{\delta'} \leq \max_{\substack{\tilde{\tau}' \in \{\tau'_L, \tau'_U\} \\ \tilde{\tau} \in \{\tau_L, \tau_U\} \\ \tilde{\delta}' \in \Delta}} \frac{\tilde{\tau}' - \tilde{\tau}}{\tilde{\delta}'}. \tag{2}$$

Theorem 2 highlights two important properties: (1) our bounding approach is agnostic to the underlying causal sensitivity analysis method making it fully compatible with established sensitivity analysis methods (e.g., VanderWeele and Arah (2011), Yadlowsky et al. (2022)). And (2) the misreporting rate bounds are attained *at* the boundaries of the valid values for $\tau'(\overline{a}, \underline{a})$, $\tau(\overline{a})$, and $\delta'(\overline{a}, \underline{a})$, making equation 2 a very simple optimization step.

5. Empirical Analysis

In this section, we first validate SCaMEr using a semi-synthetic Medicare dataset with known ground-truth diagnoses. We then apply SCaMEr to real-world Medicare data to identify which diagnoses are most susceptible to misreporting and which diagnosis collection methods are associated with the greatest levels of misreporting.

5.1. Medicare Dataset Background

CMS determines payments for MA insurers using a publicly available risk-adjustment model (Chernew et al., 2026). Originally trained on TM data, this model predicts expenditures based on a beneficiary's diagnoses known as Hierarchical Condition Categories (HCCs) (McWilliams et al., 2023). Because insurer payout is a non-decreasing function of diagnoses, MA insurers might have a financial incentive to upcode. In contrast, providers under TM are not reimbursed based on reported diagnoses and lack incentives to document every diagnosis (Ghoshal-Datta et al., 2024). This may result in downcoding diagnoses due to administrative burden. The incentives (or lack thereof) provide the basis for our data-stitching approach: we assume that MA data is never downcoded, whereas TM data is never upcoded. Importantly, we assume both TM and MA face no incentive to upcode or downcode beneficiary prescriptions, which we use as anchor variables in our analysis.

5.2. Cohort Selection and Features

Our study utilizes a 20% sample of claims data of TM and MA beneficiaries from 2018 and 2019. We restricted the cohort to beneficiaries aged 65 or older at the beginning of 2018. We excluded beneficiaries with end-stage renal disease (ESRD) as CMS uses a separate risk-adjustment model for them (Yeatts and Sangvai, 2016). We also exclude those residing outside the 50 U.S. states or territories, and beneficiaries with an unknown sex. We excluded beneficiaries enrolled in integrated health systems, which integrate a health insurance plan with hospitals and medical groups to deliver care, as their care delivery models may alter treatment patterns (Anderson et al., 2022).

We mapped each beneficiary's ICD-10 diagnosis codes reported in the claims data to V23-HCCs (X_i) using the official CMS crosswalk. Each HCC is a binary indicator of a condition's presence. Similarly, our anchors (Y_i) are binary variables denoting whether a beneficiary received at least one relevant prescription in 2019. Since we rely on prescription data to construct our anchor variables, we restrict our cohort to beneficiaries enrolled in Medicare Part D, which covers prescription drugs. To ensure clinical relevance, we used an FDA database of drug labels, designating a prescription as a valid anchor only if the HCC condition is explicitly mentioned in the label's "Indications and Usage" section. We include a full list of the prescriptions used for each HCC in Appendix B. Because prescription reporting is not tied to reimbursement in either TM or MA, yet remains clinically important, there is no incentive for misreporting prescriptions. This makes prescription data well-suited as anchor variables.

To address potential confounding between diagnoses and prescriptions, we adjusted for covariates (C), including demographics (age, sex, race, and location) and baseline health. We use one-hot encoded Hospital Referral Regions (HRRs) (Wennberg and Cooper, 2022) to account for regional variations in healthcare utilization. For stability, we grouped HRRs with less than 1% of the population into an "Other" group. To control for baseline health, we used 2018 level-one Anatomical Therapeutic Chemical (ATC) classifications (Organization et al., 2018). These indicators capture medication use by anatomical/pharmacological group at any time in 2018. We assume these covariates are free from misreporting, as demographics are inherently difficult to manipulate and ATC codes are derived from prescription data. Additional details about our dataset and anchor variables are provided in Appendix B.

5.3. Baselines

We evaluate the performance of SCaMEr against four baselines. Additional details about each baseline are included in Appendix D. **(1) Natural Direct Effect Estimator (NDEE):** measures the direct effect of the agent identity A on X. NDEE is limited as it conflates misreporting with genuine adaptation. **(2) One-Class SVM (OCSVM):** an outlier detection approach trained on only covariates C from TM data, where misreporting is assumed not to occur ($X^* = 1$), to identify anomalous reporting patterns within the MA population that may signify misreporting. **(3) CMRE (Zapzalka et al., 2025):** this method can be viewed as an ablation of our main approach that does not include data-stitching but instead assumes that TM data is unmanipulated. **(4) SCaMEr-NoC:** to evaluate the necessity of controlling for confounding, this baseline implements the SCaMEr estimator without adjusting for any covariates.

SCaMEr, CMRE, and NDEE require estimation of causal effects as a subroutine. While any valid causal estimation approach could be used, we use S-learners (Künzel et al., 2019) with an XGBoost estimator (Chen and Guestrin, 2016) for simplicity. Since SCaMEr-NoC does not adjust for confounding, we employ a standard difference-in-means estimator. Further baseline details are in Appendix D.

5.4. Semi-synthetic Experiments

To validate SCaMEr against ground-truth misreporting rates, we generated a semi-synthetic dataset using real features extracted from the Medicare dataset. We simulate a simple setting where there is only one MA insurer and one TM insurer and study whether we're able to correctly recover the simulated MA insurer's misreporting rate. We extract age, sex, and race covariates $C = (C_A, C_S, C_R)$ from the Medicare dataset. We then simulate true HCCs (X^*), reported HCCs (X), and relevant prescriptions (Y). To reflect directional misreporting, we generate two datasets ($n = 50,000$ each): one that only upcodes (MA) and another that only downcodes (TM). Because real-world data may contain mild violations of the directional misreporting assumption, we also evaluated SCaMEr under varying degrees of such violations; these experiments are detailed in Appendix C. The data-generating process is defined as follows:

$$
\begin{aligned}
A^{(j)} &\sim \text{Bernoulli}(0.4) + 0.2(1 - C_S^{(j)}), \\
X^{*(j)} &\sim \text{Bernoulli}(0.15 + 0.2 C_S^{(j)} C_R^{(j)} + 0.1 {C_A^{(j)}}^2 + 0.2 A^{(j)}), \\
Y^{(j)} &\sim \text{Bernoulli}(0.05 + 0.25 {C_S}^{(j)} + 0.2 {C_R}^{(j)} + 0.1 {C_A^{(j)}}^2 + 0.4 {X^*}^{(j)}), \\
X^{(j)} &\sim X^{*(j)} + (1 - A^{(j)}) X^{*(j)} \text{Bernoulli}(\mu_0) + A^{(j)} (1 - X^{*(j)}) \text{Bernoulli}(\mu_1).
\end{aligned}
$$

In our simulations, μ_1 is set to achieve an upcoding rate of 0.2 and μ_0 is set to achieve a downcoding rate of 0.1 for agents $A = 1$ (MA) and $A = 0$ (TM), respectively. We report the average estimate across 100 runs, with error bars representing the standard deviation. We draw new values for A, X, X^*, and Y in each iteration, but the coefficients for each variable remain the same. Additional details for our experiments can be found in the Appendix B.

Varying the downcoding rate We first evaluate how the downcoding rate of TM impacts MR estimates, highlighting the limitations of methods that assume access to a clean

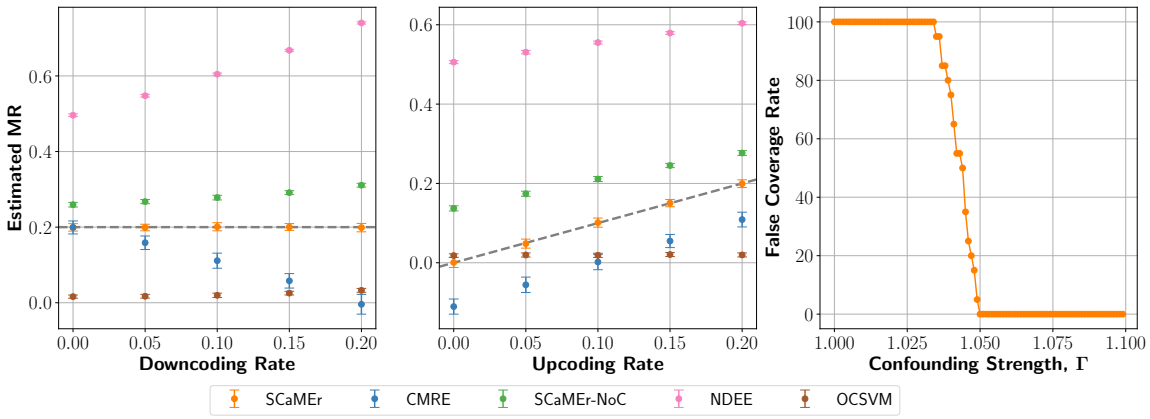

Figure 2: Semi-synthetic results. **Left:** x-axis shows TM's downcoding rate, y-axis shows estimated MR, dashed horizontal line shows true MR value. **Middle:** x-axis shows MA's upcoding rate. In both plots, SCaMEr is the only approach that gives accurate MR estimates for all downcoding/upcoding rates, whereas all baselines give biased estimates. **Right:** Confounding strength on the x-axis and false coverage rate (FCR) on the y-axis. Results show that our approach achieves low FCR for reasonable values of the confounding strength Γ.

reference dataset. To vary the amount of downcoding, we examine simulations where the downcoding rate is between 0.0 and 0.2, reflecting plausible values observed in practice. Our results are shown in Figure 2 (left), where the x-axis represents different downcoding rates and the y-axis represents the estimated MR, with the dashed line showing the true MR. We observe that SCaMEr accurately recovers the true MR across all downcoding rates, while baseline methods are not robust. Specifically, CMRE is biased when the downcoding rate is not 0, as it erroneously assumes that the TM data is reliable. This bias increases as the downcoding rate increases, such that it estimates a misreporting rate of 0 when the downcoding rate is equal to the misreporting rate of MA. Other baselines exhibit bias as well: SCaMEr-NoC is biased due to unaddressed confounding, NDEE is biased as it does not account for genuine adaptation, and OCSVM gives biased estimates as the misreported and truthful data points appear similar. These results demonstrate that failing to account for TM downcoding introduces substantial bias in baseline estimators, whereas SCaMEr recovers the true misreporting rate across all evaluated downcoding levels.

Varying the upcoding rate We further assess the performance across five simulated upcoding rates ranging from 0.0 to 0.2, reflecting reasonable values that might exist in practice. Our results are shown in Figure 2 (middle), where the x-axis represents different upcoding rates, and the y-axis represents the estimated MR. SCaMEr is able to recover the misreporting rate regardless of the upcoding rate. In contrast, CMRE exhibits bias across *all* MR values as it fails to account for downcoding. That is, as long as downcoding occurs, CMRE will always produce biased estimates. Other baselines also continue to exhibit bias: NDEE is biased due to the presence of genuine adaptation and SCaMEr-NoC continues to suffer from confounding bias. These results again highlight how SCaMEr is the only method to accurately estimate MRs regardless of the magnitude of upcoding.

5.5. Synthetic Sensitivity Analysis Experiments

To evaluate the bounds from Theorem 2, we adopt the semi-synthetic setup from Section 5.4, treating sex as a hidden confounder. We bound τ', τ, and δ' following Yadlowsky et al. (2022), who constrain unobserved confounding via its influence on the treatment assignment odds. Specifically, we assume that for all $u, \tilde{u} \in \mathcal{U}$ and $c \in \mathcal{C}$, there exists $1 \leq \Gamma \leq \infty$ such that for any probability distribution $P \in \{P_{\overline{a}}, P_{\underline{a}}\}$ and the treatment variable $V \in \{X^*, X\}$:

$$\frac{1}{\Gamma} \leq \frac{P(V=1|C=c, U=u)P(V=0|C=c, U=\tilde{u})}{P(V=0|C=c, U=u)P(V=1|C=c, U=\tilde{u})} \leq \Gamma.$$

Figure 2 (right) illustrates the false coverage rate (FCR), defined as the frequency with which the true misreporting rate falls outside our estimated uncertainty bounds, on the y-axis and the user-specified confounding strength, Γ, on the x-axis. The leftmost point in the plot where $\Gamma = \frac{1}{\Gamma} = 1$ reflects no confounding. The plot shows that our approach is able to achieve a 0 FCR at very low levels of assumed confounding (≈ 1.05). The results give an empirical validation for Theorem 2: provided valid sensitivity bounds, our approach gives valid uncertainty estimates for the misreporting rate.

5.6. Synthetic Anchor Selection Experiments

We next validate our conditional independence test for anchor uniqueness. Using a similar semi-synthetic Medicare dataset, we simulate ten features $(X_1^*, ..., X_{10}^*)$ and one anchor (Y) but randomly choose which features have a causal effect on Y such that each feature has a 50% chance of affecting Y. Each feature is assigned ground-truth misreporting and adaptation rates. We apply the conditional independence test (Proposition 1) using the kernel conditional independence test Zhang et al. (2012) for all features that causally affect Y . At significance level $\alpha = 0.05$, we achieved 100% accuracy in identifying unique anchors. Additional details about the simulation setup and a table of p-values are in Appendix C.

5.7. Real Medicare Data Experiments

In the subsequent sections, we apply SCaMEr to real Medicare data to audit misreporting within the MA program. For our analysis, we seek to (1) quantify misreporting by diagnosis to identify conditions prone to systematic misreporting and (2) compute subgroup-level rates to uncover the specific mechanisms driving misreporting.

Evaluation. Unlike the semi-synthetic experiments, we do not have ground truth MR for the real data. Instead, we conduct a sanity check: we apply our approach to the three most prevalent non-payment HCCs (those not affecting reimbursement) as negative controls. Because these offer no financial incentive for misreporting, their estimated rate should be zero. We contrast these with five payment HCCs (Specified Heart Arrhythmias, Schizophrenia, Parkinson's and Huntington's Diseases, HIV, and Diabetes with Chronic Complications), which have been identified to be frequent targets of upcoding (Weaver et al., 2024). To account for possible unobserved confounding, we apply sensitivity analysis from Theorem 2 using causal sensitivity analysis from Yadlowsky et al. (2022) setting $\Gamma = 1.05$. We obtain 95% confidence intervals for our estimator using bootstrapping with 200 samples.

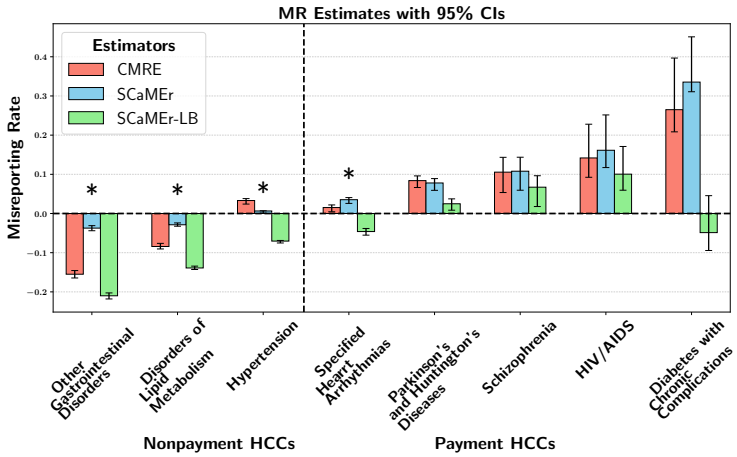

Figure 3: Presents national MR estimates by HCC. The x-axis represents both non-payment (left) and payment (right) HCCs. The y-axis shows the estimated MR, with the stars indicating an HCC had a p-value < 0.05 using our conditional independence test to test for cross-feature interference. SCaMEr produces near-zero estimates for non-payment HCCs, validating our estimator. Conversely, all five of the payment HCCs exhibit significant upcoding (MR > 0), with three remaining significant even under conservative SCaMEr-LB estimates.

5.8. Misreporting by Diagnosis

We estimate MRs across HCCs to identify diagnoses susceptible to misreporting. We report our main estimate (SCaMEr) and a conservative lower bound from our sensitivity analysis (SCaMEr-LB) to account for hidden confounding/cross-feature interference. We compare our two estimates to CMRE. Finally, we verify anchor uniqueness for each of the HCCs using the test in Proposition 1. We report the corresponding p-values in Appendix C.

Figure 3 presents the MR estimates for each HCC. Stars indicate that the p-values from our conditional independence tests were less than 0.05: if the p-value was less than 0.05, we suspect cross-feature interference. In this case, the conservative SCaMEr-LB provides a more reliable lower bound for guiding auditing decisions.

These results highlight the value of our data-stitching approach. If TM suffers from downcoding, existing methods like CMRE give unreliable estimates that do not pass our sanity checks: e.g., for one of our negative controls, "Other Gastrointestinal Disorders," where we expect a 0 MR, CMRE gives a statistically significant negative MR ≈ -0.16.

SCaMEr effectively corrects this bias, yielding an estimate significantly closer to zero. Although SCaMEr corrects for the majority of this bias, the resulting MR estimate remains slightly below zero. This may stem from violated assumptions, such as the presence of unobserved confounding or different treatment patterns between TM and MA. We discuss these limitations further in the limitations section.

While SCaMEr estimates for non-payment HCCs remain near zero, all payment HCCs show significantly positive rates. This disparity suggests that MA insurers strategically upcode these conditions to inflate reimbursements, with "Diabetes with Chronic Complications" exhibiting the most pronounced manipulation. Notably, three of the five payment

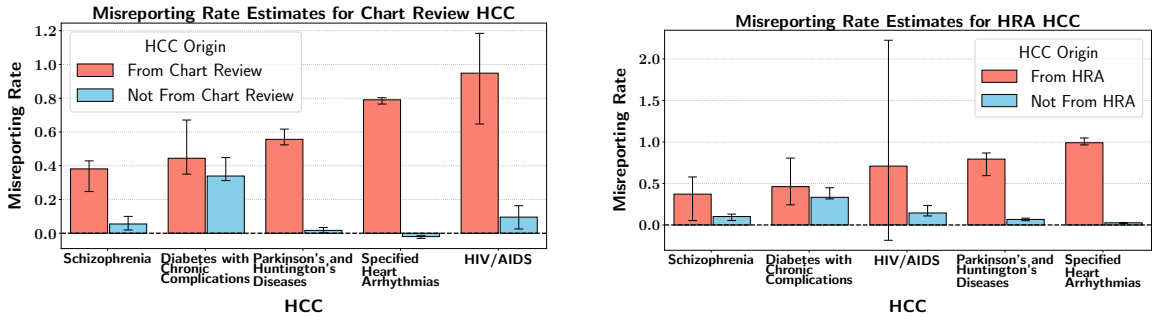

Figure 4: Figure shows estimated MRs by diagnosis documentation method. Specifically, they show the MR conditioned on whether the diagnosis was identified exclusively via chart reviews (left) or Health Risk Assessments (HRAs) (right). The x-axis represents the HCC, and the y-axis represents the estimated misreporting rate. Results demonstrate that diagnoses sourced from chart reviews and HRAs are significantly more likely to be upcoded than those documented during standard clinical encounters. The MR is significantly higher for four HCCs when submitted via chart review, and two HCCs when submitted via HRA.

HCCs remain significantly positive, even under the conservative SCaMEr-LB. This indicates that substantial unobserved confounding is insufficient to explain away the estimated MRs, reinforcing the conclusion that they are driven by strategic misreporting.

Finally, we note that SCaMEr can be applied to estimate misreporting rates across different Medicare Advantage Organizations (MAOs) or specific contracts. Using this approach, we found significant heterogeneity in MR estimates across insurance plans. For example, among the four largest MAOs by beneficiary count, the difference between the highest and lowest estimated MR was approximately 0.6. By identifying these organizational disparities, SCaMEr provides a valuable tool for regulators, enabling targeted audits of the most egregious offenders.

5.9. Variation in Misreporting Rates by Documentation Method

Finally, we investigate the drivers of misreporting by providing MR estimates across different documentation methods. Specifically, we examine chart reviews, where insurers identify additional diagnoses by reviewing a beneficiary's records, and health risk assessments (HRAs), which are used to collect information about a beneficiary's health status. Both chart reviews and HRAs have been identified as potential drivers of upcoding (Meyers and Trivedi, 2021; Jacobs, 2024).

As illustrated in Figure 4 (left), chart reviews are associated with much higher misreporting rates. For four of the five analyzed payment HCCs, MRs for diagnoses documented during direct clinical encounters remain near zero. In contrast, all diagnoses reported via chart reviews exhibit significant misreporting, with four HCCs exceeding a MR estimate of 0.4. This disparity corroborates recent literature suggesting that HCCs identified through chart reviews may be more susceptible to upcoding than those documented during face-to-face clinical encounters (Meyers and Trivedi, 2021). Furthermore, we also see evidence of upcoding in HRAs, as shown in Figure 4 (right). Notably, Parkinson's and Huntington's

Diseases, as well as Specified Heart Arrhythmias, have a significantly higher probability of being misreported when documented via HRAs compared to standard clinical encounters. These findings suggest that policymakers could consider excluding diagnoses derived solely from chart reviews and HRAs from payment calculations to mitigate misreporting.

6. Discussion

In this work, we introduced SCaMEr, a method for quantifying misreporting in clinical settings with directional incentives, such as Medicare Advantage. While existing methods require an unmanipulated reference population, SCaMEr uses data-stitching to leverage different reporting incentives across MA and TM. To account for unobserved confounding, we integrated causal sensitivity analysis, allowing auditors to obtain upper and lower bound estimates on the misreporting rate. Compared to baselines such as CMRE that do not control for downcoding, we showed that SCaMEr produces more accurate MR estimates.

Applied to a real-world Medicare dataset, SCaMEr reveals widespread misreporting across multiple HCCs, that persists even under conservative sensitivity analysis. In addition, we analyzed the mechanisms that drive misreporting, such as chart reviews and health risk assessments, showing that diagnoses submitted under these claims are much more likely to be misreported than those that were not. These findings highlight chart reviews and HRAs as primary targets for regulatory auditing and policy interventions.

Our findings indicate that misreporting remains a pervasive challenge within the MA program. However, the identified variance in robustness across different HCCs and documentation mechanisms provides a blueprint for developing more reliable risk adjustment models. By systematically auditing and excluding features susceptible to manipulation, and replacing them with robust alternatives, we can ensure that payments are governed by genuine clinical need rather than administrative gaming.

Limitations Several limitations of our approach warrant consideration. First, SCaMEr relies on directional misreporting, requiring populations with complementary reporting biases (e.g., exclusive upcoding or downcoding). In settings where all available agents exhibit the same misreporting direction, data-stitching cannot identify the misreporting rate. We provide additional simulations that show that the method remains robust to mild, two-sided violations of this directional assumption (see Appendix C).

Second, our estimates assume conditional average treatment effect invariance across populations. This assumption may be violated by unobserved effect modifiers. For example, if a factor like Body Mass Index varies between TM and MA and modifies the diagnosis-anchor relationship, our estimates may suffer from bias.

We also assume clinical consistency across TM and MA plans. If MA and TM clinicians utilize different treatment protocols for the same condition, different causal effect estimates between the two populations could be erroneously attributed to misreporting rather than genuine clinical differences. For instance, some plans may provide treatment through inpatient care instead of outpatient prescriptions, leading to an incomplete anchor variable.

Finally, our point-wise estimates may be biased due to not controlling for relevant confounders, such as housing instability, substance use, and incarceration. Biased estimates may also occur if relevant treatments are affected by other diagnoses. Thus, we must often rely on sensitivity analysis bounds instead of point-wise estimates of the misreporting rate.

Acknowledgments

We thank the reviewers for their insightful comments. Special thanks to Ezekiel Emanuel, Michelle Qiu, Zhiyi Hu, Cecilia Ehrlichman, Marko Veljanovski, Daniel K. Shenfeld, Ravi B. Parikh, and Trenton Chang 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 Ferrell 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). This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2241144. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Rajender Agarwal, John Connolly, Shweta Gupta, and Amol S Navathe. Comparing medicare advantage and traditional medicare: A systematic review: A systematic review compares medicare advantage and traditional medicare on key metrics including preventive care visits, hospital admissions, and emergency room visits. *Health Affairs*, 40(6): 937–944, 2021.

Kelly E Anderson, G Caleb Alexander, Carolyne Ma, Sydney M Dy, and Aditi P Sen. Medicare advantage coverage restrictions for the costliest physician-administered drugs. *The American journal of managed care*, 28(7):e255, 2022.

Flavia Barsotti, Rüya Gökhan Koçer, and Fernando P Santos. Transparency, detection and imitation in strategic classification. In *IJCAI*, pages 67–73, 2022.

Trenton Chang, Lindsay Warrenburg, Sae-Hwan Park, Ravi B 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.

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*, pages 785–794, 2016.

Michael E Chernew, Andy Johnson, Paul B Masi, and Karen Stockley. Aligning the medpac and cms estimates of coding intensity: The importance of the risk model and trend. *Health Affairs Forefront*, 2026.

CMS. Centers for medicare and medicaid services justification of estimates for appropriations committees fiscal year 2026, 2025a. URL https://www.cms.gov/files/document/

fy2026-cms-congressional-justification-estimates-appropriations-committees.pdf.

CMS. Cms rolls out aggressive strategy to enhance and accelerate medicare advantage audits, 5 2025b. URL https://www.cms.gov/newsroom/press-releases/cms-rolls-out-aggressive-strategy-enhance-accelerate-medicare-advantage-audits.

Amanda Cook and Susan Averett. Do hospitals respond to changing incentive structures? evidence from medicare's 2007 drg restructuring. *Journal of Health Economics*, 73:102319, 2020.

Brian C DeBusk, Brian J Miller, Craig Samitt, and Kenny Kan. The need for holistic policy thinking in medicare. *Health Affairs Forefront*, 2024.

National Center for Toxicological Research. FDALABEL: Full-Text search of Drug product Labeling, 4 2026. URL https://www.fda.gov/science-research/bioinformatics-tools/fdalabel-full-text-search-drug-product-labeling.

Bianca K Frogner, Gerard F Anderson, Robb A Cohen, and Chad Abrams. Incorporating new research into medicare risk adjustment. *Medical care*, 49(3):295–300, 2011.

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

Niru Ghoshal-Datta, Michael E Chernew, and J Michael McWilliams. Lack of persistent coding in traditional medicare may widen the risk-score gap with medicare advantage: Article examines diagnostic coding in traditional medicare and medicare advantage. *Health Affairs*, 43(12):1638–1646, 2024.

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.

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

Guido W Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003.

Paul D Jacobs. In-home health risk assessments and chart reviews contribute to coding intensity in medicare advantage: Study examines the role of health risk assessments and chart reviews in coding intensity in medicare advantage. *Health Affairs*, 43(7):942–949, 2024.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

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

J Michael McWilliams, Gabe Weinreb, Lin Ding, Chima D Ndumele, and Jacob Wallace. Risk adjustment and promoting health equity in population-based payment: Concepts and evidence: Study examines accuracy of risk adjustment and payments in promoting health equity. *Health Affairs*, 42(1):105–114, 2023.

MedPAC. The medicare advantage program: Status report, 3 2025. URL https://www.medpac.gov/wp-content/uploads/2025/03/Mar25_Ch11_MedPAC_Report_To_Congress_SEC.pdf.

David J Meyers and Amal N Trivedi. Medicare advantage chart reviews are associated with billions in additional payments for some plans. *Medical care*, 59(2):96–100, 2021.

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.

World Health Organization et al. Anatomical therapeutic chemical (atc) classification system, 2018.

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

Jane Hyatt Thorpe, Elizabeth A. Gray, Anthem Public Policy Institute, Inc. Anthem, EBG Advisors, AHIP, Milken Institute School of Public Health, Centers for Medicare & Medicaid Services, George Washington University, and Teresa R. Cascio. Medicare Advantage and the False Claims Act. Technical report, 2019. URL https://www.elevancehealth.com/content/dam/elevance-health/articles/ppi_assets/partner-papers/GWU_MedAdv_FCA_Final_with_Contact_info.pdf.

Tyler J VanderWeele and Onyebuchi A Arah. Unmeasured confounding for general outcomes, treatments, and confounders: bias formulas for sensitivity analysis. *Epidemiology (Cambridge, Mass.)*, 22(1):42, 2011.

Christopher Weaver, Tom McGinty, Anna Wilde Mathews, Mark Maremont, and Andrew Mollica. Exclusive — insurers pocketed $50 billion from medicare for diseases no doctor treated, Jul 2024. URL https://www.wsj.com/health/healthcare/medicare-health-insurance-diagnosis-payments-b4d99a5d.

John E Wennberg and Megan M Cooper. The dartmouth atlas of health care in the united states: The center for the evaluative clinical sciences, 2022.

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *Annals of statistics*, 50(5):2587, 2022.

John P Yeatts and Devdutta G Sangvai. Hcc coding, risk adjustment, and physician income: what you need to know. *Family Practice Management*, 23(5):24–27, 2016.

Dylan Zapzalka, Trenton Chang, Lindsay Warrenburg, Sae-Hwan Park, Daniel K Shenfeld, Ravi B Parikh, Jenna Wiens, and Maggie Makar. Disentangling misreporting from genuine adaptation in strategic settings: a causal approach. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

Appendix A. Proofs for Theoretical Results

A.1. Proof for Lemma 1

Before presenting the proof for Theorem 1, we first present Lemma 1, which is identical to Lemma 1 in Zapzalka et al. (2025). This lemma demonstrates that estimating the true and nominal causal effects of X^* on Y allows us to estimate the misreporting rate.

Lemma 1 *Let Assumptions 1-5 hold. Also, let P follow any of the DAGs in Figure 1. Then for $\delta_i^*(a) \neq 0$, the MR can be expressed as:*

$$P_a(X_i^* = 0 | X_i = 1) = \frac{\tau_i^*(a) - \tau_i(a)}{\delta_i^*(a)}$$

Proof This proof for this Lemma is identical to the proof given for Lemma 1 in Zapzalka et al. (2025). For the first part of the proof, we show that we can decompose $\tau_i(a)$ as $\tau_i^*(a)$ and some bias term:

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

The third equality holds as $Y \perp X_i | X_i^*, A$ for the DAGs in Figure 1 and the fourth equality holds due to Assumption 1. To obtain the fifth equality, the first and fourth terms are collapsed into the definition of $\tau_i^*(a)$ applying the standard causal Assumptions 3-5.

For the next part of the proof, we show that the bias term can be rewritten in terms of the misreporting rate, $P_a(X_i^* = 0|X_i = 1)$:

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

The first three equalities use the definition of conditional probability and Bayes' rule to factor the misreporting rate $P_a(X_i^* = 0|X_i = 1)$ out of the integral. The final equality where the remaining integral is substituted with the definition of $\delta_i^*(a)$ holds due to standard causal Assumptions 3-5:

As the final step, we can obtain the misreporting rate by simply rearranging the terms:

$$\tau_i(a) = \tau_i^*(a) - P_a(X_i^* = 0|X_i = 1)\delta_i^*(a) \implies P_a(X_i^* = 0|X_i = 1) = \frac{\tau_i^*(a) - \tau_i(a)}{\delta_i^*(a)},$$

for $\delta_i^*(a) \neq 0$. ■

A.2. Proof for Lemma 2

Next, we give an additional Lemma used for the proof in Theorem 1, which shows that $\tau_i(\overline{a})$, $\tau_i'(\overline{a}, \underline{a})$, and $\delta_i'(\overline{a}, \underline{a})$ are identifiable.

Lemma 2 *Let Assumptions 1-5 hold. Also let $\overline{a}$ be an agent that never downcodes and $\underline{a}$ be an agent that never upcodes. Then $\tau_i(\overline{a})$, $\tau_i'(\overline{a}, \underline{a})$, and $\delta_i'(\overline{a}, \underline{a})$ are identifiable.*

Proof By definition, we know that

$$\tau_i(\overline{a}) = \int_C (\mathbb{E}_{P_{\overline{a}}}[Y_i|X_i = 1, C = c] - \mathbb{E}_{P_{\overline{a}}}[Y_i|X_i = 0, C = c]) P_{\overline{a}}(C = c|X_i = 1) dc.$$

It follows immediately that $\tau_i(\overline{a})$ is identifiable as it is a statistical estimand and all necessary variables are observed.

Also recall by definition that

$$\tau_i'(\overline{a}, \underline{a}) = \int_C (\mathbb{E}_{P_{\underline{a}}}[Y(X_i^* = 1)|C = c] - \mathbb{E}_{P_{\overline{a}}}[Y(X_i^* = 0)|C = c]) P_{\overline{a}}(C = c|X_i = 1) dc,$$

and

$$\delta_i'(\overline{a}, \underline{a}) = \int_C (\mathbb{E}_{P_{\underline{a}}}[Y(X_i^* = 1)|C = c] - \mathbb{E}_{P_{\overline{a}}}[Y(X_i^* = 0)|C = c]) P_{\overline{a}}(C = c|X_i = 0) dc.$$

Since $P_{\overline{a}}(C = c|X_i = 1)$ and $P_{\overline{a}}(C = c|X_i = 0)$ are identifiable, it only remains to show that the stitched conditional average treatment effect

$$\mathbb{E}_{P_{\underline{a}}}[Y(X_i^* = 1)|C = c] - \mathbb{E}_{P_{\overline{a}}}[Y(X_i^* = 0)|C = c]$$

is identifiable.

This follows directly from the standard causal Assumptions 3-5 which allow us to transform the statistical estimand into a causal one, as well as by the assumptions that agent $\overline{a}$ doesn't downcode and $\underline{a}$ doesn't upcode:

$$\begin{aligned}
&\mathbb{E}_{P_{\underline{a}}}[Y(X_i^* = 1)|C = c] - \mathbb{E}_{P_{\overline{a}}}[Y(X_i^* = 0)|C = c] \\
&= \mathbb{E}_{P_{\underline{a}}}[Y(X_i^* = 1)|X_i^* = 1, C = c] - \mathbb{E}_{P_{\overline{a}}}[Y(X_i^* = 0)|X_i^* = 0, C = c] \\
&= \mathbb{E}_{P_{\underline{a}}}[Y|X_i^* = 1, C = c] - \mathbb{E}_{P_{\overline{a}}}[Y|X_i^* = 0, C = c] \\
&= \mathbb{E}_{P_{\underline{a}}}[Y|X_i = 1, C = c] - \mathbb{E}_{P_{\overline{a}}}[Y|X_i = 0, C = c]
\end{aligned}$$

Thus, $\tau_i(\overline{a})$, $\tau_i'(\overline{a}, \underline{a})$, and $\delta_i'(\overline{a}, \underline{a})$ are identifiable. ■

A.3. Proof for Theorem 1

Finally, we use the results from both Lemma 1 and Lemma 2 to prove Theorem 1.

Theorem 3 (Restated Theorem 1 in main text.) *Let Assumptions 1-6 hold. Suppose $\overline{a}$ does not downcode and $\underline{a}$ does not upcode. Also, let P follow any of the DAGs in Figures 1a, 1c, or 1d. Then for $|\delta_i'(\overline{a}, \underline{a})| > 0$, the MR is identifiable and can be expressed as:*

$$P_{\overline{a}}(X_i^* = 0|X_i = 1) = \frac{\tau_i'(\overline{a}, \underline{a}) - \tau_i(\overline{a})}{\delta_i'(\overline{a}, \underline{a})}$$

Proof

Since we already know by Lemma 2 that $\tau_i(\overline{a})$, $\tau_i'(\overline{a}, \underline{a})$, and $\delta_i'(\overline{a}, \underline{a})$ are identifiable, by Lemma 1, it only remains to show that $\tau_i'(\overline{a}, \underline{a}) = \tau_i^*(\overline{a})$, and $\delta_i'(\overline{a}, \underline{a}) = \delta_i^*(\overline{a})$.

To show that $\tau_i'(\overline{a}, \underline{a}) = \tau_i^*(\overline{a})$, recall that

$$\tau_i'(\overline{a}, \underline{a}) = \int_C (\mathbb{E}_{P_{\underline{a}}}[Y(X_i^* = 1)|C = c] - \mathbb{E}_{P_{\overline{a}}}[Y(X_i^* = 0)|C = c])P_{\overline{a}}(C = c|X_i = 1)dc$$

and

$$\tau_i^*(\overline{a}) = \int_C (\mathbb{E}_{P_{\overline{a}}}[Y(X_i^* = 1)|C = c] - \mathbb{E}_{P_{\overline{a}}}[Y(X_i^* = 0)|C = c])P_{\overline{a}}(C = c|X_i = 1)dc$$

Since $\mathbb{E}_{P_{\overline{a}}}[Y(X_i^* = 1)|C = c] = \mathbb{E}_{P_{\underline{a}}}[Y(X_i^* = 1)|C = c]$ by Assumption 6, it follows that $\tau_i'(\overline{a}, \underline{a}) = \tau_i^*(\overline{a})$.

Next, to show that $\delta_i'(\overline{a}, \underline{a}) = \delta_i^*(\overline{a})$, recall that

$$\delta_i'(\overline{a}, \underline{a}) = \int_C (\mathbb{E}_{P_{\underline{a}}}[Y(X_i^* = 1)|C = c] - \mathbb{E}_{P_{\overline{a}}}[Y(X_i^* = 0)|C = c])P_{\overline{a}}(C = c|X_i = 0)dc.$$

and

$$\delta_i^*(\overline{a}) = \int_C (\mathbb{E}_{P_{\overline{a}}}[Y(X_i^* = 1)|C = c] - \mathbb{E}_{P_{\overline{a}}}[Y(X_i^* = 0)|C = c])P_{\overline{a}}(C = c|X_i^* = 0, X_i = 1)dc.$$

By Assumption 6, it only remains to show that $P_{\overline{a}}(C = c|X_i = 0) = P_{\overline{a}}(C = c|X_i^* = 0, X_i = 1)$:

$$\begin{aligned} P_{\overline{a}}(C = c|X_i = 0) &= P_{\overline{a}}(C = c|X_i = 0, X_i^* = 1)P_{\overline{a}}(X_i^* = 1|X_i = 0) \\ &\quad + P_{\overline{a}}(C = c|X_i = 0, X_i^* = 0)P_{\overline{a}}(X_i^* = 0|X_i = 0) \\ &= P_{\overline{a}}(C = c|X_i = 0, X_i^* = 0) \\ &= P_{\overline{a}}(C = c|X_i^* = 0) \\ &= P_{\overline{a}}(C = c|X_i^* = 0, X_i = 1) \end{aligned}$$

The final equality holds due to $X_i \perp C|X_i^*, A$. Thus, the proof is complete. ■

A.4. Proof of Proposition 1

In Proposition 1, we provide a conditional independence test for showing when cross-feature interference occurs.

Proposition 4 (Restated Proposition 1 in main text.) *Assume X_i^* causes Y_i, assume that A causes X_j and X_j^*, and assume that X_j^* causes X_j for all $j \in [M]$. Finally, let $X' = \{X_j : j \in [M] \setminus \{i\}\}$, assume P is faithful, and assume that there is no unmeasured confounding between all X_j^* and Y_i, i.e., $Y_i(X_j^* = 0), Y_i(X_j^* = 1) \perp X_j^*|C$ for all $j \in [M]$. Then*

$$X' \perp Y_i|X_i^*, C \implies X_i^* \text{ is the unique causal parent of } Y_i \text{ among all features } X_j^*.$$

Proof For this proof, we show that the contrapositive statement holds: if X_i^* is not a unique causal parent of Y_i among all features X_j^*, then $X' \not\perp Y_i|X_i^*, C$.

We know that if X_i^* is not a unique causal parent, then some X_j^* causes Y_i where $j \neq i$. Since X_j^* causes X_j by assumption, we get that $X_j \leftarrow X_j^* \rightarrow Y_i$. Since conditioning on X_i^* and C does not block the path between X_j and Y_i, we get that $X_j \not\perp Y_i|X_i^*, C$. Therefore, it follows that $X' \not\perp Y_i|X_i^*, C$.

■

A.5. Proof for Theorem 2

In the following Theorem, we show that by bounding the causal estimands used to estimate the MR, we can obtain upper and lower bounds on the MR in the presence of unobserved confounding.

Theorem 5 (Restated Theorem 2 in main text.) *Suppose a sensitivity analysis method provides valid bounds for $\tau'(\overline{a}, \underline{a})$, $\tau(\overline{a})$, and $\delta'(\overline{a}, \underline{a})$, i.e.,*

$$\tau'_L \leq \tau'(\overline{a}, \underline{a}) \leq \tau'_U, \quad \tau_L \leq \tau(\overline{a}) \leq \tau_U, \text{ and } \quad \delta'_L \leq \delta'(\overline{a}, \underline{a}) \leq \delta'_U.$$

For $|\delta'(\overline{a}, \underline{a})| \geq \epsilon$ with $\epsilon > 0$, let $\Delta = \{\delta'_L, \delta'_U, -\epsilon, \epsilon\} \cap [\delta'_L, \delta'_U] \cap \{\tilde{\delta}' : |\tilde{\delta}'| \geq \epsilon\}$. Then we have that the misreporting rate, $\frac{\tau'-\tau}{\delta'}$ is bounded as:

$$\min_{\substack{\tilde{\tau}' \in \{\tau'_L, \tau'_U\} \\ \tilde{\tau} \in \{\tau_L, \tau_U\} \\ \tilde{\delta}' \in \Delta}} \frac{\tilde{\tau}' - \tilde{\tau}}{\tilde{\delta}'} \leq \frac{\tau' - \tau}{\delta'} \leq \max_{\substack{\tilde{\tau}' \in \{\tau'_L, \tau'_U\} \\ \tilde{\tau} \in \{\tau_L, \tau_U\} \\ \tilde{\delta}' \in \Delta}} \frac{\tilde{\tau}' - \tilde{\tau}}{\tilde{\delta}'}. \tag{3}$$

Proof We prove that

$$\min_{\substack{\tilde{\tau}' \in \{\tau'_L, \tau'_U\} \\ \tilde{\tau} \in \{\tau_L, \tau_U\} \\ \tilde{\delta}' \in \Delta}} \frac{\tilde{\tau}' - \tilde{\tau}}{\tilde{\delta}'} \leq \frac{\tau' - \tau}{\delta'}.$$

The upper bound proceeds in an identical way.

We first consider the case where $0 \notin [\delta'_L, \delta'_U]$. Note that $\frac{\tilde{\tau}'-\tilde{\tau}}{\tilde{\delta}'}$ is continuous in $\tilde{\tau}', \tilde{\tau}$ and $\tilde{\delta}'$, which means that it attains some minimum over the intervals $[\tau'_L, \tau'_U], [\tau_L, \tau_U], [\delta'_L, \delta'_U]$ by the extreme value theorem. We need to next show that the minimum is attained *at* the boundaries of those intervals. To do so, we examine the first derivative with respect to each of the 3 values. Letting $h := \frac{\tilde{\tau}'-\tilde{\tau}}{\tilde{\delta}'}$, we have that:

$$\frac{\partial h}{\partial \tilde{\tau}'} = \frac{1}{\tilde{\delta}'}.$$

In the case where $0 \notin [\delta'_L, \delta'_U]$, we have that $\tilde{\delta}' \neq 0$. The derivative shows that the function is monotonically decreasing for $\tilde{\delta}' < 0$, in which case the function attains its minimum at the τ'_U. The opposite is true if $\tilde{\delta}' > 0$. This shows that h attains its minimum at the boundaries of $\tilde{\tau}'$.
The same argument can be applied to show that h attains its minimum at the boundaries of allowable values for $\tilde{\tau}$ and $\tilde{\delta}'$.

We next consider the case where $0 \in [\delta'_L, \delta'_U]$. Note that the assumption that $|\delta'| \geq \epsilon$ means that we can obtain more informative bounds for $\tilde{\delta}'$. Specifically, we have that $\tilde{\delta}' \in [\delta'_L, -\epsilon] \cup [\epsilon, \delta'_U]$. Using the same logic as the one presented in the case where $0 \notin [\delta'_L, \delta'_U]$, we can construct a valid lower bound $\underline{\text{MR}}_1$ for the case where $\tilde{\delta}' \in [\delta'_L, -\epsilon]$ and another valid lower bound $\underline{\text{MR}}_2$ for the case where $\tilde{\delta}' \in [\epsilon, \delta'_U]$. Taking the final lower bound to be $\min\{\underline{\text{MR}}_1, \underline{\text{MR}}_2\}$ gives our final lower bound. ■

Appendix B. Medicare Dataset

B.1. Medicare Dataset Overview

The Medicare data was provided to the authors under a data usage agreement with the Centers for Medicare and Medicaid Services. It consists of a 20% sample of all U.S. beneficiaries enrolled in either Traditional Medicare or Medicare Advantage. The data used in

our experiments only uses enrollees that had Medicare Coverage for both 2018 and 2019. For our analysis, we looked at how much MAOs were upcoding V23 HCCs. To obtain the V23-HCCs, we used an official crosswalk provided by CMS which allowed us to map ICD-10 codes in each of the beneficiaries' claims data to the corresponding HCC code. We looked at two different types of HCCs: HCCs that were used in the risk-adjustment model to determine the payment given to an insurer for each beneficiary (payment HCCs) as well as HCCs that were not included within the payment model (non-payment HCCs).

B.2. Cohort Characteristics and HCC Prevalence

Table 1 provides the beneficiary counts and sample prevalence for all Hierarchical Condition Categories evaluated in our experiments. In total, there were 5,166,191 beneficiaries in our sample.

Table 1: Count of Analyzed HCCs in the 20% Medicare Sample.

HCC	**Count (N)**	**Prevalence (%)**
HCC 1	15,899	0.31%
HCC 18	640,550	12.40%
HCC 25	3,707,419	71.76%
HCC 38	2,399,911	46.45%
HCC 57	12,117	0.23%
HCC 78	84,754	1.64%
HCC 95	3,679,222	71.22%
HCC 96	882,621	17.09%

B.3. Obtaining Anchor Variables

For the anchor variables used in our analysis, e.g., downstream causal variables for each HCC, we use a binary variable indicating whether or not the beneficiary obtained a relevant prescription for the HCC within the year 2019. For this, we used prescriptions that were reported from enrollees' Part D data. To obtain which prescriptions are relevant for each HCC, we used an FDA database of drug labels to obtain a set of relevant drugs for each HCC (for Toxicological Research, 2026). These were determined by searching whether or not a prescription in the FDA database contained a relevant keyword in the "Indications and Usage" section of the drug label. For instance, to determine if a drug is used to treat HIV, we searched for the keyword "HIV" within the drug labels. The keywords used to obtain these prescriptions are shown in Table 2 below. Each of the keywords was determined to be clinically relevant as determined by a clinician.

Appendix C. Additional Experiments and Results

C.1. p-values for Figure 3 experiments

In Table 3 below, we give the p-values that were calculated using our conditional independence test for detecting cross-feature interference. For these experiments, we used KCIT

Table 2: HCCs used in our analysis, along with the keywords used to search for relevant prescriptions for each HCC.

HCC #	HCC Name	Keywords
HCC1	HIV	"HIV"
HCC18	Diabetes with Chronic Complications	"Diabetic Nephropathy" or "Diabetic Kidney Disease"
HCC25	Disorders of Lipid Metabolism	"Hypercholesterolemia" or "Familial Hypercholesterolemia" or "Primary Hyperlipidemia" or "Heterozygous Familial Hypercholesterolemia"
HCC38	Other Gastrointestinal Disorders	"Gastroesophageal Reflux Disease" or "Barrett's Esophagus"
HCC57	Schizophrenia	"Schizophrenia"
HCC78	Parkinson's/Huntington's Disease	"Parkinson's Disease" or "Parkinson Disease"
HCC95	Hypertension	"Hypertension"
HCC96	Specified Heart Arrhythmias	"Atrial Fibrillation"

Zhang et al. (2012) with a Gaussian kernel and a sample size of 10000. For these experiments, we only used the 8 total HCCs that were in the plot, and not all V23 HCCs that are included in the CMS risk-adjustment model.

Table 3: p-values for the cross-feature interference conditional independence tests for each of the HCCs in Figure 3.

HCC	p-value
HCC1	0.069
HCC18	0.080
HCC25	0.000
HCC38	0.006
HCC57	0.170
HCC78	0.174
HCC95	0.000
HCC96	0.000

C.2. p-values synthetic anchor selection experiments

In Table 4, we give the p-values for each of the 10 simulations we performed for our synthetic anchor selection experiments, and Table 5 shows the downcoding, upcoding, and genuine adaptation rates used in the simulations. The tests all gave p-values above 0.05, which aligns with the results expected from the conditional independence test. For these simulations, the conditional independence test was only done for the subset of the features that did not causally affect the anchor such that $X' \perp Y|A, C$. Therefore, we expect a p-value of greater than 0.05 for each simulation.

Table 4: p-values for the synthetic anchor selection experiments. A total of 10 simulations were done.

Simulation #	p-value	Accurate Result
1	0.469866	True
2	0.469629	True
3	0.509115	True
4	0.589766	True
5	0.301700	True
6	0.501420	True
7	0.392114	True
8	0.574213	True
9	0.434885	True
10	0.522808	True

Table 5: Simulation setup for semi-synthetic cross-feature interference conditional independence tests. Describes the downcoding rates, upcoding rates, and the amount of genuine adaptation for each feature used in the simulation.

Feature	Downcoding Rate	Upcoding Rate	Genuine Adaptation
1	0.1	0.04	0.3
2	0.2	0.1	0.05
3	0.3	0.2	0.06
4	0.0	0.3	0.17
5	0.05	0.0	0.03
6	0.06	0.05	0.2
7	0.07	0.06	0.1
8	0.04	0.4	0.2
9	0.03	0.03	0.3
10	0.04	0.6	0.0

C.3. Sensitivity to violations of directional misreporting

As with any statistical method, violations of the underlying assumptions can lead to biased estimates. We conducted additional simulations with varying degrees of directional assumption violations to test the robustness of SCaMEr. As shown in Tables 6-8, the method remains robust under mild violations, with performance degrading as violations increase. Notably, the results demonstrate that two-sided violations of directional misreporting (i.e., TM upcoding and MA downcoding occurring simultaneously) tend to result in less error than one-sided violations.

Table 6: MSE (± std. of the squared errors) of the estimated MR when TM upcodes and MA downcodes. The true MR is 0.2, and the TM upcoding and MA downcoding rates are given in the "Violation" column.

Violation	SCaMEr	CMRE	NDEE	OCSVM
0	0.0000 ± 0.0001	0.0074 ± 0.0033	0.1407 ± 0.0025	0.0341 ± 0.0015
0.02	0.0001 ± 0.0002	0.0005 ± 0.0009	0.1257 ± 0.0025	0.0344 ± 0.0010
0.04	0.0001 ± 0.0003	0.0009 ± 0.0010	0.1104 ± 0.0024	0.0345 ± 0.0009
0.06	0.0003 ± 0.0004	0.0054 ± 0.0022	0.0945 ± 0.0025	0.0344 ± 0.0011
0.08	0.0007 ± 0.0008	0.0114 ± 0.0033	0.0787 ± 0.0020	0.0347 ± 0.0008
0.1	0.0011 ± 0.0010	0.0206 ± 0.0046	0.0628 ± 0.0021	0.0346 ± 0.0008

Table 7: MSE (± std. of the squared errors) of the estimated MR when MA downcodes. The true MR is 0.2, and the MA downcoding rate is given in the "Violation" column.

Violation	SCaMEr	CMRE	NDEE	OCSVM
0	0.0000 ± 0.0001	0.0074 ± 0.0033	0.1407 ± 0.0025	0.0341 ± 0.0015
0.02	0.0001 ± 0.0003	0.0000 ± 0.0004	0.1333 ± 0.0025	0.0340 ± 0.0015
0.04	0.0010 ± 0.0007	0.0051 ± 0.0021	0.1255 ± 0.0025	0.0341 ± 0.0014
0.06	0.0025 ± 0.0010	0.0173 ± 0.0036	0.1170 ± 0.0026	0.0339 ± 0.0014
0.08	0.0044 ± 0.0016	0.0329 ± 0.0050	0.1082 ± 0.0022	0.0340 ± 0.0013
0.1	0.0077 ± 0.0021	0.0534 ± 0.0064	0.0988 ± 0.0024	0.0340 ± 0.0016

Appendix D. Misreporting Estimators

D.1. Additional Estimands

In addition to the misreporting rate defined in Definition 1, we note that our approach could also extend to two other estimands introduced in Zapzalka et al. (2025):

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

Table 8: MSE (± std. of the squared errors) of the estimated MR when TM upcodes. The true MR is 0.2, and the TM upcoding rate is given in the "Violation" column.

Violation	SCaMEr	CMRE	NDEE	OCSVM
0	0.0000 ± 0.0001	0.0074 ± 0.0033	0.1407 ± 0.0025	0.0341 ± 0.0015
0.02	0.0005 ± 0.0004	0.0122 ± 0.0039	0.1330 ± 0.0025	0.0344 ± 0.0010
0.04	0.0015 ± 0.0009	0.0175 ± 0.0053	0.1252 ± 0.0024	0.0346 ± 0.0009
0.06	0.0034 ± 0.0011	0.0251 ± 0.0056	0.1173 ± 0.0025	0.0345 ± 0.0011
0.08	0.0068 ± 0.0018	0.0366 ± 0.0080	0.1094 ± 0.0021	0.0348 ± 0.0008
0.1	0.0110 ± 0.0022	0.0468 ± 0.0094	0.1015 ± 0.0023	0.0348 ± 0.0007

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

How our analysis extends to these estimands follows directly from the analysis done in Appendix C in Zapzalka et al. (2025).

D.2. Baselines

Similar to Zapzalka et al. (2025), we use the following estimators as baselines to estimate the misreporting rates.

D.2.1. NDEE

The NDEE baseline attempts to estimate a misreporting rate by measuring the direct effect of A on X

$$\frac{1}{P_{\overline{a}}(X=1)} \int_C (\mathbb{E}_{P_{\overline{a}}}[X|C=c] - \mathbb{E}_{P_{\underline{a}}}[X|C=c]) P_{\overline{a}}(C=c) dc. \tag{4}$$

D.2.2. OCSVM

The One-Class SVM follows a similar setup to that of Zapzalka et al. (2025), where the One-Class SVM model, denoted as $g(y, c)$, is trained to identify outliers (e.g., misreported instances) by using the variables Y and C. It is trained only on datapoints from an agent $\underline{a}$ where $X = 1$, and it predicts a misreporting rate over N_1 samples of an agent $\overline{a}$ where $X = 1$:

$$\hat{\text{MR}} = \frac{1}{N_1} \sum_{j: j \in \mathcal{D}_{\overline{a}1}} g(y_j, c_j).$$

Here, $\mathcal{D}_{\overline{a}1}$ is a dataset for agent $\overline{a}$ where $X = 1$.