Predictions as Behaviors: Steering Machine Learning Models Towards High-Level Principles

by

Trenton Chang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in the University of Michigan
2026

Doctoral Committee:

Associate Professor Jenna Wiens, Chair
Assistant Professor Maggie Makar
Associate Professor Michael W. Sjoding
Research Scientist Adith Swaminathan, Netflix, Inc.

Trenton Chang
ctrenton@umich.edu
ORCID iD: 0000-0003-4679-5841

© Trenton Chang 2026

ACKNOWLEDGEMENTS

First, I thank my committee members for their invaluable feedback on my dissertation, and for their investment in my growth as a researcher. Specifically, I thank my advisor Jenna Wiens. I cannot imagine a better teacher for this stage of my research career. There is no institution, no lab, no environment that would have made a better scientist out of me. Under her advising I have learned so much about how to conduct independent research and communicate effectively across fields—and, I will always hear her encouraging me to think a little more critically.

I would also like to thank my other committee members, Michael Sjoding, Maggie Makar, and Adith Swaminathan, for supporting me on my various projects, listening to my struggles, and offering their empathy throughout the Ph.D. Each of them was integral in kick-starting a separate chapter of my thesis, be it through collaborations initiated by Jenna, or through internships. It is no coincidence that these three individuals form the rest of my thesis committee today.

Various academic collaborators have also invested countless time into my work: Amanda Kowalski, Daniel Shenfeld, Divya Shanmugam, Dylan Zapzalka, Keith E. Kocher, Lindsay Warrenburg, Mark Nuppnau, Ravi B. Parikh, Sae-Hwan Park, Thomas S. Valley, Tobias Schnabel, Ying He, Zach Brown, and the Augmented Learning & Reasoning Group at Microsoft Research Redmond, not to mention countless others I briefly crossed paths with. I am also grateful to have been part of the ML4H, MLHC, and CHIL communities. I would like to especially thank Stephanie Shepard, for wearing so many hats: helping with grant questions, room reservations, paper editing, presentation advice, and everything else.

Next, I thank the friends that supported me through this journey. First, I thank my labmates in the MLD3 lab both past and present—mentors, mentees, friends. I have learned a lot from each and every one of them. Second, I thank the friends and colleagues from Microsoft Research for the good memories. Third, I thank the other two-thirds of my jazz trio, Harry Rubin-Falcone and Eli Ochshorn, for reminding me of the joy of music. A special thanks to Sarah Jabbour, Serafina Kamp, Daniel Geng, Ricardo Cannizzaro (stay stochastic!), Emily Bao, Dhia Rzig, Katie Hough, John Wall, Joe Lou, Zheng Yan, and Bjorn Rodriguez: thanks for all the memes.

I thank my wonderful partner, Jung Min Lee. I am grateful for her tremendously positive effect on my sanity throughout the Ph.D., for her constant encouragement and understanding, and for introducing me to so many of her (our?) favorite shows. And of course, I thank my parents, Tom Chang and Hsiu-Hsiang Liu for all advice professional and unprofessional through the years.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF FIGURES vi

LIST OF TABLES xv

LIST OF APPENDICES xix

LIST OF ACRONYMS xx

ABSTRACT xxi

CHAPTER

1 Introduction 1

1.1 Challenges and Opportunities 2
1.2 Contributions 4

2 Background 6

2.1 Bias and fairness in supervised learning 6
2.1.1 Supervised learning 6
2.1.2 Measuring fairness in supervised learning 8
2.2 Large language models 9
2.2.1 Overview of conditional generative models 9
2.2.2 Training large language models 9
2.2.3 Controlling large language models 10
2.3 Strategic adaptation 11
2.3.1 Example: The CMS-HCC Risk Adjustment Model 11
2.3.2 The strategic classification framework 12

3 Disparate Censorship: A Case Study of Biases at Training Time 13

3.1 Racial differences in laboratory testing as a potential mechanism for bias in AI . . 14
3.1.1 Introduction 14
3.1.2 Methods 15
3.1.3 Results 17
3.1.4 Evidence of disparate censorship 21

3.2 Disparate Censorship & Undertesting: A Source of Label Bias in Clinical Machine Learning . . . 25
3.2.1 Introduction . . . 25
3.2.2 Problem Setup & Definitions . . . 27
3.2.3 Theoretical Analysis of Disparate Censorship & Undertesting . . . 30
3.2.4 Empirical Analysis of Disparate Censorship & Undertesting . . . 34
3.2.5 Experiments & Results . . . 35
3.2.6 Potential impacts of disparate censorship . . . 39
3.3 From Biased Selective Labels to Pseudo-Labels: An Expectation-Maximization Framework for Learning from Biased Decisions . . . 42
3.3.1 Introduction . . . 42
3.3.2 Preliminaries: Disparate Censorship . . . 42
3.3.3 Methodology . . . 44
3.3.4 Experimental Setup . . . 49
3.3.5 Experiments & Discussion . . . 51
3.3.6 Related Work . . . 55
3.3.7 Conclusion . . . 56
3.4 Concluding remarks on disparate censorship: equitable ML models, and beyond . 58

4 A Course Correction in Steerability Evaluation: Goal-conditioned Model Behavior at Inference Time . . . 59

4.1 A Course Correction in Steerability Evaluation: Revealing Miscalibration and Side Effects in LLMs . . . 60
4.1.1 Introduction . . . 60
4.2 A steerability measurement framework . . . 61
4.2.1 Designing a steerability metric . . . 62
4.2.2 Measuring steerability in practice . . . 63
4.3 Experimental setup . . . 64
4.3.1 Steerability probe construction . . . 64
4.3.2 Candidate steerability interventions . . . 65
4.3.3 LLM inference setup . . . 67
4.4 Empirical results . . . 67
4.4.1 Large language models are not steerable . . . 68
4.4.2 Inference-time steering is costly . . . 70
4.4.3 RL yields progress towards steerable models . . . 71
4.5 Discussion & Conclusion . . . 72

5 Gaming and Strategic Adaptation: Facilitating Oversight over Deployment-time Interactions . . . 74

5.1 Who's Gaming the System? A Causally-Motivated Approach for Detecting Strategic Adaptation . . . 75
5.1.1 Introduction . . . 75
5.1.2 Related works . . . 76
5.1.3 Background & Problem Setup . . . 77
5.1.4 Theoretical analysis: finding agents most likely to game . . . 79

5.1.5 Empirical results & discussion . . . 83
5.1.6 Conclusion . . . 88
5.2 Streamlining Oversight in Medicare Advantage: A New Approach to Measuring Coding Intensity via Contract Rankings . . . 89
5.2.1 Introduction . . . 89
5.2.2 Methods . . . 90
5.2.3 Limitations . . . 93
5.2.4 Study Results . . . 94
5.2.5 Discussion . . . 97
5.2.6 Conclusion . . . 99
5.3 Concluding remarks on gaming . . . 102

6 Conclusion . . . 103

6.1 Summary of contributions . . . 103
6.2 Future extensions . . . 104
6.3 Closing remarks . . . 105

APPENDICES . . . 106

BIBLIOGRAPHY . . . 253

LIST OF FIGURES

FIGURE

1.1 Overview of three layers of alignment under study in this thesis: (1) default behavior of the model, which encompasses concerns such as fairness, (2) goal-conditioned model behavior, which concern the ability of the model to satisfy user requests, and (3) strategic behavior, which concern how the underlying incentive and regulatory environment shapes user-model interactions. 2

3.1 Cohort diagram of patient groups selected for study at BIDMC (top) and U-M (bottom). Note that at BIDMC, there were no ED visits in BIDMC associated with individuals of unknown sex. Numbers may not sum to 100.0% due to rounding. 17

3.2 Differences in testing rates between White and Black patients, with and without matching on age, sex, chief complaint, and ED triage score. Forest plots show difference in testing rate across White vs. Black patients (White % - Black %), with no matching adjustment (top), and matched on age, sex, chief complaint, and ED triage score (bottom). Differences at BIDMC are shown in red; differences at U-M are shown in blue. Actual values shown in Tables A.5 & A.6 (results with no matching adjustment), or Tables A.3 & A.4 (matched analysis). 19

3.3 Differences in admission rates between White and Black patients matched on age, sex, chief complaint, and ED triage score. Bar plots show difference in rate of admission to the hospital among ED visits in study cohort grouped by site. Differences at BIDMC are shown in red; differences at U-M are shown in blue. "**" denotes statistical significance at 5% level. Actual values are shown in Tables A.3 and A.4. 21

3.4 Stylized undertesting patterns. In both examples, the red group is undertested (higher risk r required for a test). Left: undertesting is harmful due to the disproportionate rates of missed diagnoses. Right: the absence of missed diagnoses means that undertesting has no disproportionate impact in terms of missed positives. In this chapter, we focus on the scenario on the left. 28

3.5 Causal DAG used in our analysis. (Un)shaded variables are (un)observed. Patient risk-stratification models aim to predict y given x, but only $\tilde{y}$ is observed. Disparate censorship arises when a affects t. Also, a may affect $\mathbf{x}$ and y (dashed arrows). 29

3.6 Decomposition of overall AUC into within-group AUCs (cyan) and xAUCs (magenta). Zero ranking performance gap between groups a requires parity for both within-group AUCs (cyan arrow) and xAUCs (magenta arrow). 29

3.7 Example decision (black) and censorship boundaries (blue, orange) in 2D. In our noise model, we assume that the clinician "test decision" (t) and true "condition status" (y) decision boundaries are based on a threshold with the same functional form (*i.e.* "parallel"). 32
3.8 Example of disparate censorship in 2D with groups $a = 0$ (in blue) and $a = 1$ (in orange), and decision boundary in black. All blue positives ($a = 0$, +) below the blue dashed line are censored; likewise for orange positives ($a = 1$, +). 35
3.9 ΔAUC (left) and ΔxAUC (right) with 95% empirical CIs for the model trained on y (gray) and the model trained on $\tilde{y}$ (green). As the degree of undertesting experienced by the high-risk group $a = 1$ ($\tau_1 - \tau_0$) increases (dashed red), with a sharply increasing amount of overall missed positives (solid red), ΔAUC, ΔxAUC rises. 36
3.10 ΔAUC (left) and ΔxAUC (right) with 95% empirical CIs for the model trained on y (gray) and the model trained on $\tilde{y}$ (green). Even as the degree of undertesting experienced by the low-risk group $a = 0$ ($\tau_0 - \tau_1$) increases, such that the missed positive rate approaches 1 (dashed red), all metrics stay near oracle performance while the overall missed positive rate plateaus (red). 37
3.11 Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (ϕ; x-axis) and censorship rate in $a = 1$ ($P_1(t = 0)$, y-axis); 4 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger gaps are in dark red. Performance gaps widen as conditional shift or disparate censorship intensify. 38
3.12 Decision tree for identifying when disparate censorship may negatively impact model performance gaps (assuming perfect separation). In summary, if either the marginal risk distribution differs (1) and the high-risk group is undertested (2), or the conditional risk distribution differs (3) and the decision and censorship boundaries are non-parallel (4), performance gaps may emerge (ΔAUC, ΔxAUC > 0). 39
3.13 Comparison of ROC gap (left) and AUC (right) of selected models at $q_y = 0.5, k = 1, q_t = 2$. Each point represents a different s_Y. Our method (DCEM, magenta) mitigates bias while maintaining competitive AUC compared to baselines, with a tighter range and improved empirical worst-case for both metrics. "-": median, "$\triangle$": worst-case ROC gap, "$\triangledown$": worst-case AUC. 51
3.14 Relative frequencies of ROC gaps for DCEM vs. tested-only models at similar AUC (increasing to the right), pooling models across all k, q_y, q_t tested. Dashed lines = mean ROC gap by model. DCEM improves bias mitigation among models with similar AUC. 52
3.15 ROC gaps (left) and AUC (right) of baselines and DCEM on sepsis classification task at $q_t = 1.5, k = 4$. Each dot represents a different s_T. Our method (DCEM, magenta) maintains competitive or better bias mitigation and discriminative performance compared to baselines. "-": median, "$\triangle$": worst-case ROC gap, "$\triangledown$": worst-case AUC. 53

4.1 Steerability metrics in 2D goal-space (reading level & text length). A user aims to rewrite text ($\mathbf{z}_0$) according to intent $\mathbf{z}^*$, expressed via a prompt (`Make this harder to read...`). The steering error (red dotted line) is the gap between the user's intent $\mathbf{z}^*$ (blue) and the LLM's output $\hat{\mathbf{z}}$ (red). Miscalibration (miscal.) and orthogonality (ortho.) decompose steering error into components parallel and orthogonal to steering error respectively. 61
4.2 Median (IQR) of steering error (left), miscalibration (middle), and orthogonality, Llama3 family. Steering error does not improve with model size (left), but miscalibration does (middle). Orthogonality drops slightly (right), but remains skewed. . . . 66
4.3 Vector flow of goal-space movement (blue), Llama3.3-70B, in requests to change reading difficulty but not formality. Horizontal movement is desired, but not vertical movement. Source texts in red. 67
4.4 Median and IQR steerability, Llama3.3-70B, in correlated (darker) vs. anti-correlated (lighter) requests for change in reading difficulty and formality. Llama3.3-70B struggles more with anti-correlated changes. 68
4.5 Median and IQR of steering error (left), miscalibration (middle), and orthogonality (right) of Llama3.1-8B across prompting strategies. More detailed prompts and removal of the negative prompt marginally improve miscalibration over the default. However, side-effects remain severe. 68
4.6 Median and IQR for best-of-$\{4, 8, \ldots, 128\}$ approaches on Llama3.1-8B, with a direct + negative prompt. Increasing N consistently improves steerability, but improvements scale slowly. 69

5.1 **Left:** Two agents with gaming deterrence parameters $\lambda_1 = 30$ (purple) and $\lambda_2 = 50$ (blue) maximize utility (reward R - cost c) with respect to diagnosis rate. Gaming costs *increase* in $\lambda_{(\cdot)}$, and lower an agent's optimal diagnosis rate (stars). **Center:** Agents' observed decisions reflect utility-maximizing behavior. **Right:** A decision-maker computes a payout based on agent decisions. 77
5.2 **Left:** Causal graph for gaming detection with confounders $\mathbf{x}$, agent indicator p, ground truth diagnosis d^*, and agent decision d. **Right:** Toy dataset with observed factual outcomes $d_i(p)$ and $d_i(p')$. "?" denotes missing counterfactual outcomes. . . . 82
5.3 Causally-motivated gaming detection. **Left:** First, we impute counterfactual decisions for each agent. **Middle:** The imputed counterfactuals yield average treatment effects (ATEs) across *pairs* of agents. **Right:** Using ATE estimates to rank agents yields a ranking of the gaming parameter λ_p. We show one direction of comparison across agents without loss of generality. In practice, we impute decisions for both directions of comparison and average (given blue agent's observations, impute purple agent's decisions). 82
5.4 Pseudocode for causally-motivated gaming detection. Causal effect estimators take pairs of agents (p, p') and their observations $\mathbf{x}^{(\cdot)}$ as inputs, and output the mean difference in predicted decision rate. 83
5.5 Mean top-5 sensitivity (left) and DCG (center) across # of agents audited, and top-5 sensitivity with 7 audits (right) at mean range 0.9, with $\pm\sigma$ error. Causal methods improve over non-causal baselines. $\triangledown$: naïve baselines. $\circ$: anomaly detectors. $\times$: causal effect estimators. 86

5.6 Area under the sensitivity curve (AUSC) for causal vs. non-causal methods across levels of confounding, with $\pm\sigma$ error. As confounding increases, a payout-only ranking degrades. Anomaly detection performance does not vary across confounding strength, maintaining slightly better than random rankings. Causal methods generally maintain higher mean AUSC than baselines across confounding levels. $\triangledown$: naïve baselines. $\circ$: anomaly detectors. $\times$: causal effect estimators. 87
5.7 AUSC of S+IPW (causal) vs. KNN (non-causal) across confounding strength with $\pm\sigma$ error. The advantage of S+IPW over KNN decreases as confounding diminishes. . 87
5.8 Overview of the proposed contract ranking pipeline. From left to right: for each diagnosis, we compute the per-contract, per-market prevalence. We adjust for differences in beneficiary health using post-stratification weights and apply shrinkage towards the market-level mean, yielding a contract ranking for the market. These rankings are aggregated across diagnoses and markets in a weighted Bradley-Terry model to produce a final, national ranking. 91
5.9 Cohort diagram with exclusion criteria for TM-to-MA beneficiaries analyzed. Starting with a switchers cohort, we focus on non-dual beneficiaries to isolate... that did not move states, were not on hospice care and survived to the end of the study period (Dec. 2019), stayed on a single MA contract, and did not reside in a monopolistic county (only one MA contract available). 95
5.10 Local, shrinkage-based and adjusted diagnosis rates for contracts (names anonymized) in Los Angeles County, CA for diabetes with (left) and without complications (right). Red indicates high national rank (top 50%); blue indicates low national rank (bottom 50%), with increased saturation indicating more extreme (high/low) ranks. While county-level signals are noisy, we aggregate such patterns across HCCs and across counties to produce the national ranking. Contract identifiers are anonymized using alphabetical labels (A, B, ..., Z, AA, AB, ...). 96
5.11 Median rank and IQR (width of the bars) across bootstraps (n=1000) of coding intensity rank for top 20-ranked contracts (sorted by median rank in ascending order), with number of switchers in each contract. A higher rank (further left) indicates more intense coding. 97
5.12 Share of contracts by primary plan type (HMO vs. PPO) by decile (1 = highest coding intensity). The share of HMOs relative to PPOs decreases along the coding intensity ranking. 98
5.13 Binscatter plot of relevant treatments with respect to rankings computed using our approach over **a)** diabetes with complications (HCC18), **b)** various psychiatric conditions (HCC58), **c)** arrhythmias/vascular disease (HCC96, 107, and 108), **d)** COPD (HCC111), and **e)** Parkinson's disease (HCC78). Transparent dots = treatment rate for a contract, where the size of the dot corresponds to the number of beneficiaries with the HCC in the contract. Solid dots = per-decile treatment rates, with an enrollment-weighted trendline (dashed) summarizing per-decile treatment rates. 100

5.14 Graphs representing the fraction of misreported diagnoses audited as a function of the c-index (accuracy of a ranking) for different auditing intensities (10%, 30%, 50%, 70%, and 90%) and contract-level diagnosis rates. **a)** displays the fraction of recovered diagnoses when the mean diagnosis rate is 0.1 and the spread of the diagnosis rate is 0.06; **b)** displays the fraction of recovered diagnoses when the mean diagnosis rate is 0.3 and the spread of the diagnosis rate is 0.02. 101

A.1 Sensitivity analysis of matching strategy showing differences in testing rates between White and Black patients matched on age, sex, and chief complaint only. 109
A.2 Differences in rate of admission to hospital between White and Black patients matched on age, sex, and chief complaint. 110
A.3 Differences in testing rates between White and Black patients matched on age, biological sex, chief complaint, and ED triage score, stratified by male (top) and female (bottom) patients. 111
A.4 Subgroup analysis of testing differences by admission status (discharge vs. admission) between White and Black patients matched on age, sex, chief complaint, and ED triage score. 112
A.5 Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (x-axis) and censorship rate in $a = 1$ (y-axis); 2 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger performance gaps are in dark red. 128
A.6 Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (ϕ; x-axis) and censorship rate in $a = 1$ ($P_1(t = 0)$, y-axis); 4 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger performance gaps are in dark red. 129
A.7 Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (ϕ; x-axis) and censorship rate in $a = 1$ ($P_1(t = 0)$, y-axis); 6 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger performance gaps are in dark red. 129
A.8 Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (ϕ; x-axis) and censorship rate in $a = 1$ ($P_1(t = 0)$, y-axis); 8 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger performance gaps are in dark red. 130
A.9 Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (ϕ; x-axis) and censorship rate in $a = 1$ ($P_1(t = 0)$, y-axis); 10 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger performance gaps are in dark red. 130
A.10 ΔAUC (left) and ΔxAUC (right) with 95% empirical CIs for the model trained on y (gray) and on $\tilde{y}$ (green). As σ^2 increases to the right, increasing overlap between the distributions, the number of missed positives increases, and ΔAUC, ΔxAUC decrease. 131
A.11 ΔAUC (left) and ΔxAUC (right) with 95% empirical CIs for the model trained on y (gray) and on $\tilde{y}$ (green). As $\Delta\mu$ increases to the right, the AUC and xAUC gaps *widen*. Note that $P_0(\tilde{y} = 0 \mid y = 1) = 0$—there are no missed positives in group $a = 0$. The number of missed positives first increases, then decreases, since sufficiently high $\Delta\mu$ means that untested patients' risk largely lies in one tail of the risk distribution. 131

A.12 Contour plot of $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ with respect to $\hat{t}^{(i)}$ (x-axis) and $Q(y^{(i)})$ (y-axis). $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ scales with $Q(y^{(i)})$ but decreases in $\hat{t}^{(i)}$. 152
A.13 Cohort diagram for our Sepsis-3 cohort ($N = 5301$). *: Further details are provided in [133]. ˆ: Our cohort size differs slightly from that reported in [133] due to an apparent pre-processing error in defining Sepsis-3 [260]; our reported cohort size applies the relevant correction. 160
A.14 Sensitivity analysis of causal effect estimators with respect to AUC and ROC gap (min, max across s_T in parentheses) for $q_y = 0.5, k = 1, q_t = 2$ across varying levels of overlap. Causally-motivated methods are shown in green, while DCEM is shown in magenta. Empirically, DCEM improves robustness to overlap violations. 167
A.15 Calibration plot of $\hat{t}$ for (from left to right) $\tau \in \{0.1, 1, 10\}$. While $\hat{t}$ is well-calibrated for $\tau = 1$, changing τ in either direction (< 1 vs. > 1) induces miscalibration error. . . 169
A.16 Relative frequencies of AUC for DCEM vs. SELF at similar ROC gaps, pooling models across all k, q_y, q_t tested. Dashed lines = mean AUC by model. DCEM improves AUC among models with similar ROC gaps when the ROC gap is below 0.04. 170
A.17 ROC gaps (left) and AUC (right) of selected models on sepsis classification task as weighting of systolic blood pressure (BP) and respiratory (resp.) rate (s_T) for testing changes ("0.0"/left: consider systolic BP only; "1.0"/right: consider resp. rate only) at $q_t = 1.5, k = 4$. If a feature is "more salient," it is weighted higher than the other in the testing decision function s_T. 171
A.18 ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/3, q_t = 1/2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 172
A.19 ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/3, q_t = 1$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 172
A.20 ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/3, q_t = 2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 173
A.21 ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/2, q_t = 1/2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 173
A.22 ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/2, q_t = 1$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 173
A.23 ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/2, q_t = 2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 174
A.24 ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1, q_t = 1/2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 174
A.25 ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1, q_t = 1$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 174
A.26 ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1, q_t = 2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 175
A.27 ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 2, q_t = 1/2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 175
A.28 ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 2, q_t = 1$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 175
A.29 ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 2, q_t = 2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 176

A.30 ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 3, q_t = 1/2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 176
A.31 ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 3, q_t = 1$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 176
A.32 ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 3, q_t = 2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 177
A.33 ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 1/4, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 177
A.34 ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 1/3, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 177
A.35 ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 1/2, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 178
A.36 ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 1, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 178
A.37 ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 2, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 178
A.38 ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 3, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 179
A.39 ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 4, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 179
A.40 ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 5, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC. 179

B.1 Median and IQR steerability metrics for (from left to right) Llama3 (blue), GPT (yellow), o1/o3-mini (orange), Deepseek (red), and Qwen3 (gray) family models. 195
B.2 Median and IQR steerability metrics for (from left to right) Llama3.3-70B (blue), GPT-4.1 (yellow), Deepseek-80B (red), and Qwen3-32B (gray), stratified by requests for correlated (darker) vs. anti-correlated (anti-correlated) changes to reading difficulty and formality. Across all model families, LLMs struggle to satisfy in anti-correlated requests more so than correlated requests. 196
B.3 Median and IQR steerability metrics for various prompting strategies, Llama3.1-8B (from left to right): underspecified, direct, direct with negative prompt, chain-of-thought, chain-of-thoughts with negative prompt, instruction-only, direct with instructions, and direct with instructions and negative prompt. 197
B.4 Median and IQR steerability metrics for a probe sampled with (left) and without (right) uniform reweighting. Steerability metrics remain similar across probes, but reweighting remains a principled approach to ensure that differences in steerability metrics are not dominated by common/easy goals. 198
B.5 Llama3.1-8B flow diagram on steerability-tuning evaluation set prior to RL on instructions where reading difficulty is specified, but formality is not. 198
B.6 Llama3.1-8B flow diagram on steerability-tuning evaluation set prior to RL on instructions where formality is specified, but reading difficulty is not. 199
B.7 Llama3.1-8B flow diagram on steerability-tuning evaluation set after RL on instructions where reading difficulty is specified, but formality is not. 199

B.8 Llama3.1-8B flow diagram on steerability-tuning evaluation set after RL on instructions where formality is specified, but reading difficulty is not. 199
B.9 Llama3.3-70B flow diagram, (reading difficulty, formality) subspace. 204
B.10 Llama3.3-70B flow diagram, (reading difficulty, textual diversity) subspace. 205
B.11 Llama3.3-70B flow diagram, (reading difficulty, text length) subspace. 205
B.12 Llama3.3-70B flow diagram, (formality, reading difficulty) subspace. 205
B.13 Llama3.3-70B flow diagram, (formality, textual diversity) subspace. 206
B.14 Llama3.3-70B flow diagram, (formality, textual length) subspace. 206
B.15 Llama3.3-70B flow diagram, (textual diversity, reading difficulty) subspace. 206
B.16 Llama3.3-70B flow diagram, (textual diversity, formality) subspace. 207
B.17 Llama3.3-70B flow diagram, (textual diversity, text length) subspace. 207
B.18 Llama3.3-70B flow diagram, (text length, reading difficulty) subspace. 207
B.19 Llama3.3-70B flow diagram, (text length, formality) subspace. 208
B.20 Llama3.3-70B flow diagram, (text length, textual diversity) subspace. 208

C.1 **Top:** Causal graph for gaming detection with agent indicator p, ground truth diagnosis d^*, and agent decision d. **Bottom:** Causal graph for gaming detection with added partially observed confounders $\mathbf{x}$. 215
C.2 Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.0, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator. 227
C.3 Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.1, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator. 227
C.4 Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.2, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator. 227
C.5 Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.3, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator. 228
C.6 Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.4, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator. 228
C.7 Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.5, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator. 228
C.8 Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.6, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator. 229
C.9 Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.7, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator. 229
C.10 Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.8, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator. 229

C.11 Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.9, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator. 230
C.12 Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 1.0, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator. 230
C.13 Area under the sensitivity curve (AUSC) for all methods tested across levels of confounding (mean range $R_\mu \leq 0.5$). $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator. 230
C.14 Area under the sensitivity curve (AUSC) for all methods tested across levels of confounding (mean range $R_\mu > 0.5$). $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator. 231
C.15 Sensitivity analysis of all causal methods tested, mean range 0.0. $\triangledown$: naïve baseline. $\circ$: causal effect estimator. 232
C.16 Sensitivity analysis of all causal methods tested, mean range 0.1. $\triangledown$: naïve baseline. $\circ$: causal effect estimator. 232
C.17 Sensitivity analysis of all causal methods tested, mean range 0.2. $\triangledown$: naïve baseline. $\circ$: causal effect estimator. 233
C.18 Sensitivity analysis of all causal methods tested, mean range 0.3. $\triangledown$: naïve baseline. $\circ$: causal effect estimator. 233
C.19 Sensitivity analysis of all causal methods tested, mean range 0.4. $\triangledown$: naïve baseline. $\circ$: causal effect estimator. 233
C.20 Sensitivity analysis of all causal methods tested, mean range 0.5. $\triangledown$: naïve baseline. $\circ$: causal effect estimator. 234
C.21 Sensitivity analysis of all causal methods tested, mean range 0.6. $\triangledown$: naïve baseline. $\circ$: causal effect estimator. 234
C.22 Sensitivity analysis of all causal methods tested, mean range 0.7. $\triangledown$: naïve baseline. $\circ$: causal effect estimator. 234
C.23 Sensitivity analysis of all causal methods tested, mean range 0.8. $\triangledown$: naïve baseline. $\circ$: causal effect estimator. 235
C.24 Sensitivity analysis of all causal methods tested, mean range 0.9. $\triangledown$: naïve baseline. $\circ$: causal effect estimator. 235
C.25 Sensitivity analysis of all causal methods tested, mean range 1.0. $\triangledown$: naïve baseline. $\circ$: causal effect estimator. 235
C.26 Contracts ranking 1st to 90th (inclusive) in switchers' cohort, (2018 TM to 2019 MA). 246
C.27 Contracts ranking 91st to 180th (inclusive) in switchers' cohort, (2018 TM to 2019 MA). 247
C.28 Contracts ranking 181st to 270th (inclusive) in switchers' cohort, (2018 TM to 2019 MA). 248
C.29 Contracts ranking 271st to 360th (inclusive) in switchers' cohort, (2018 TM to 2019 MA). 249
C.30 Contracts ranking 361st to 427th (inclusive) in switchers' cohort, (2018 TM to 2019 MA). 250

LIST OF TABLES

TABLE

3.1 Cohort demographics for BIDMC among all ED visits for the unmatched cohort. . . . 18
3.2 Cohort demographics for U-M among all ED visits for the unmatched cohort. 20

4.1 Comparison of single-turn LLM evaluation strategies by scoring/loss function (ℓ) and a representative evaluation dataset ($P(\mathbf{z}^*, \mathbf{z}_0)$) in terms of LLM output $\hat{\mathbf{z}}$ and user goal $\mathbf{z}^*$. 62
4.2 Main results for RL, with an ablation study of the reward model. Mean (std. dev.) of steerability metrics across evaluation probe ($N = 1,024$: 64 held-out source texts; 16 goals each). 70
4.3 Mean (std. dev.) of (from left to right) orthogonality for pre- vs. post-RL model on correlated (top; e.g., increase both dimensions) vs. anti-correlated requests (middle; e.g., change dimensions in opposite directions). RL shrinks the gap in side-effects (bottom) between correlated and anti-correlated requests, despite only supervised via 1D instructions. 70
4.4 Mean (std. dev.) sentence-level BLEU (original vs. rewrite) by dataset, pre- & post-RL. 70

5.1 Summary statistics for switchers' cohort ($n = 128{,}455$ beneficiaries). 94

A.1 Case-normalized free-text chief complaint for 25 most common chief complaints in the unmatched cohort at BIDMC among White and Black patients. 107
A.2 Case-normalized free-text chief complaint for 25 most common chief complaints in the unmatched cohort at U-M among White and Black patients. 108
A.3 Full results for differences in testing rates at BIDMC between White and Black patients matched on age, sex, chief complaint, and ED triage score. 113
A.4 Full results for differences in testing rates at U-M between White and Black patients matched on age, sex, chief complaint, and ED triage score. 113
A.5 Testing rates at BIDMC, White and Black pateitns. 113
A.6 Testing rates at U-M, Black and White patients. 114
A.7 Cohort demographics and subsequent hospital admission rate for BIDMC among all ED visits after matching on age, biological sex, chief complaint, and ED triage score. . 114
A.8 Cohort demographics and subsequent hospital admission rate for U-M among all ED visits after matching on age, biological sex, chief complaint, and ED triage score. . . . 115
A.9 Differences in testing rates at BIDMC between White and Black patients matched on age, sex, and chief complaint, all patients. 115
A.10 Differences in testing rates at U-M between White and Black patients matched on age, sex, and chief complaint, all patients. 115

A.11 Differences in testing rates at BIDMC between White and Black patients matched on age, sex, chief complaint, and ED triage score among all patients, by biological sex. . 116
A.12 Differences in testing rates at U-M between White and Black patients matched on age, sex, chief complaint, and ED triage score among all patients, by biological sex. 116
A.13 Difference in testing rates at BIDMC between White and Black patients in ED visits resulting in discharge, matched on age, sex, chief complaint, and ED triage score. . . . 117
A.14 Difference in testing rates at U-M between White and Black patients in ED visits resulting in discharge, matched on age, sex, chief complaint, and ED triage score. . . . 117
A.15 Difference in testing rates at BIDMC between White and Black patients in ED visits resulting in admission, matched on age, sex, chief complaint, and ED triage score. . . 117
A.16 Difference in testing rates at U-M between White and Black patients in ED visits resulting in admission, matched on age, sex, chief complaint, and ED triage score. . . 118
A.17 Simulation settings for each of the three distributional settings studied. Note that $\sigma_0^2 = \sigma_1^2 = 0.1$ for all settings. Scoring function $s(\mathbf{x})$ is defined in Eq. A.17. 124
A.18 Disparate censorship in common laboratory/diagnostic tests in White vs. Black/African-American patients, MIMIC-IV v1.0, with testing rates by group, z-statistics, and p-values. 127
A.19 Group-wise AUC and ΔAUC under marginal distributional distance for all values of τ_0, τ_1. 133
A.20 xAUC metrics and ΔxAUC under marginal distributional distance for all values of τ_0, τ_1. 134
A.21 Group-wise AUC and ΔAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 2$. 135
A.22 xAUC and ΔxAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 2$. 136
A.23 Group-wise AUC and ΔAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 4$. 137
A.24 Group-wise xAUC and ΔxAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 4$. 138
A.25 Group-wise AUC and ΔAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 6$. 139
A.26 Group-wise xAUC and ΔxAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 6$. 140
A.27 Group-wise AUC and ΔAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 8$. 141
A.28 Group-wise xAUC and ΔxAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 8$. 142
A.29 Group-wise AUC and ΔAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 10$. 143
A.30 Group-wise xAUC and ΔxAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 10$. 144
A.31 AUC, xAUC with empirical 95% CIs at varying settings of $\Delta\mu$, $\tau_0 = 5, \tau_1 = 6.6$. . . . 145
A.32 AUC, xAUC with empirical 95% CIs at varying settings of σ^2, $\tau_0 = 5, \tau_1 = 6.6$. 145
A.33 Sensitivity analysis of DCEM components with respect to AUC and ROC gap (min, max across s_T in parentheses) for $q_y = 0.5, k = 1, q_t = 2$. Maximum (minimum) median AUC (ROC gap) in bold. 166

A.34 Sensitivity analysis of causal effect estimators for estimating $P(Y \mid X)$ compared to DCEM with respect to AUC and ROC gap (min, max across s_T in parentheses) for $q_y = 0.5, k = 1, q_t = 2$. Maximum (minimum) median AUC (ROC gap) in bold. . . . 167
A.35 Sensitivity analysis of softmax temperature scaling parameter (T) with respect to DCEM AUC and ROC gap (min, max across s_T in parentheses) for $q_y = 0.5, k = 1, q_t = 2$. Maximum (minimum) median AUC (ROC gap) in bold. . . . 169
A.36 Sensitivity analysis of E-step initialization with respect to DCEM AUC and ROC gap (min, max across s_Y in parentheses) for $q_y = 0.5, k = 1, q_t = 2$. Maximum (minimum) median AUC (ROC gap) in bold. . . . 170

B.1 Kendall's τ of an LLM-based evaluation of each text dimension compared to the goal-space mapping function used. Evaluation performed on rewrites by Llama3.1-8B (direct + neg. prompt). We also provide the pairwise agreement rate ($(\tau + 1)/2$). . . . 183
B.2 Values for each goal-dimension corresponding to $[0, 1]$ in normalized goal-space. . . . 183
B.3 Model endpoint and version information for all models evaluated. For HuggingFace models, the first seven digits of the model commit hash on the HuggingFace model hub is provided. . . . 191
B.4 Example showing copy-pasting behavior in the pre-RL base model un-learned after RL, sourced from CNN/DailyMail validation split. [...] added to indicate mid-sentence truncation. FK: Flesch-Kincaid (reading difficulty). HD: Heylighen-Dewaele (formality). Note that metrics are for the entire source text and may not match the excerpt provided. **Instruction** (immediately before source text): "Please rewrite the following, but make it much more informal. You MUST not change anything else about the other parts of the text, even if it makes the rewritten text sound unnatural or otherwise awkward. Respond with only the rewritten text and do not explain your response." . . . 209
B.5 Example showing the model demonstrating the ability to disentangle goals post-RL, sourced from Reddit TIFU. FK: Flesch-Kincaid (reading difficulty). HD: Heylighen-Dewaele (formality). Note that metrics are for the entire source text and may not match the excerpt provided. **Instruction** (immediately before source text): "Please rewrite the following, but make it more formal, and easier to read. You MUST not change anything else about the other parts of the text, even if it makes the rewritten text sound unnatural or otherwise awkward. Respond with only the rewritten text and do not explain your response." . . . 210
B.6 Counts of grounded rewrites in main steerability probe for Llama3.1-8B under different prompting strategies, with number of rewrites flagged as un-grounded by the LLM, and number of LLM-flagged rewrites overruled. NP: negative prompt. Inst.: instruction(s). . . . 211
B.7 Counts of grounded rewrites in main steerability probe under different best-of-N strategies. Note that only the lowest steering-error response for each was evaluated. . . . 211
B.8 Counts of grounded rewrites in main steerability probe under different models. . . . 212

C.1 Features chosen for analysis derived from the provider of service file, with mean and standard deviation across states. S.d.: standard deviation. OP PT/SP: outpatient physical therapy and speech pathology. ICF/ID: intermediate care facility for individuals with intellectual disability. *: invalid (1/0 or 0/0) ratios dropped. 223
C.2 Features chosen for analysis derived from NANDA, with mean and standard deviation across states. All chosen summary statistics are for providers with > 0 U.S. dollars in sales. S.d.: standard deviation. Sq. mi.: square mile. P/O/ST: physical, occupational, and speech therapists. IDD: intellectual and developmental disabilities. MHSA: mental health and substance abuse. 225
C.3 Eligible HCCs (prevalence >0.5%) in switchers' cohort, sorted in descending order of prevalence. 243
C.4 HCC to treatment (ATC drug categories or RBCS procedure categories) mappings used for treatment gap analysis. 245
C.5 Correlation matrix (Kendall's τ) between different rankings with empirical 95% confidence intervals across bootstraps. 246
C.6 Raw rates of relevant treatments by contract decile for diabetes with complications (HCC18). ATC A10: drugs for diabetes. 250
C.7 Raw rates of relevant treatments by contract decile for major depressive, bipolar, and paranoid disorders (HCC58). ATC N06A: antidepressants; RBCS EB: evaluation and management, behavioral health. 251
C.8 Raw rates of relevant treatments by contract decile for specified heart arrhythmias, vascular disease with complications, and vascular disease (HCC96, 107, and 108). ATC B01A: antithrombotic agents; C01B: antiarrhythmics (class I and III); C07A: beta-blocking agents; C04A: peripheral vasodilators; RBCS PC: cardiovascular procedures. 251
C.9 Raw rates of relevant treatments by contract decile for chronic obstructive pulmonary disease (HCC111). ATC R03: drugs for obstructive airway diseases; RBCS TF/TP: pulmonary function tests. 252
C.10 Raw rates of relevant treatments by contract decile for Parkinson's and Huntington's diseases (HCC78). ATC N04: anti-parkinson drugs. 252

LIST OF APPENDICES

A Appendix for Disparate Censorship: A Case Study of Biases at Training Time 106

Supplementary material for "Racial differences in laboratory testing as a potential mechanism for bias in AI" 106
Supplementary material for "Disparate Censorship and Undertesting: A Source of Label Bias in Clinical Machine Learning" 119
Supplementary material for "From Biased Selective Labels to Pseudo-Labels: An Expectation-Maximization Framework for Learning from Biased Decisions" . . . 146

B Appendix for A Course Correction in Steerability Evaluation: Goal-conditioned Model Behavior at Inference Time 180

Supplementary material for A Course Correction in Steerability Evaluation: Revealing Miscalibration and Side Effects in LLMs 180

C Appendix for Gaming and Strategic Adaptation: Facilitating Oversight over Deployment-time Interactions 213

Supplementary material for "Who's Gaming the System? A Causally-Motivated Approach to Detecting Strategic Adaptation" 213
Supplementary material for "Streamlining Oversight in Medicare Advantage: A New Approach to Measuring Coding Intensity via Contract Rankings " 240

LIST OF ACRONYMS

AUC Area under receiver operating characteristic curve

AI Artificial intelligence

CMS-HCC Centers for Medicare & Medicaid Services–Hierarchical Category Condition

EHR Electronic Health Record

HCC Hierarchical Category Condition

IPW Inverse propensity weighting

IQR Interquartile range

LLM Large language model

NANDA National Neighborhood Data Atlas

MA Medicare Advantage

ML Machine learning

RL Reinforcement learning

ROC Receiver operating characteristic

ABSTRACT

Machine learning (ML) systems are used in settings where they interact with people and institutions at scale. However, ML pipeline design often focuses on a narrow loop: data collection, operationalizing an objective, optimization, and release of the resulting model. While this process can produce effective prediction machines, deployed ML systems do not merely emit predictions: they shape workflows, influence decisions, and affect human behavior. Misalignment arises when good prediction under pre-specified metrics fails to translate into good behavior in context, and can take many forms. The data collection process may be biased, such that the ML model is trained only on filtered views of various communities. Metrics may also fail to capture what model users want, resulting in undesirable behavior when steered towards specific goals. Post-deployment, users may modify their own behavior to exploit the system's assumptions.

This dissertation argues that alignment is not a fixed property of a trained artifact, but continual, spanning data collection, model development, and post-deployment monitoring. Concretely, we develop tools for detecting misalignment in training data (disparate censorship), in model outputs under goal-directed steering (steerability probes), and in agent behavior when interacting with deployed systems (gaming scores and rankings). We also explore partial mitigation strategies, including changes to training objectives to correct biased supervision (for disparate censorship) or post-training methods to reduce behavioral side effects in LLMs (for steerability).

Biases in laboratory testing motivate our study of model training data. Many clinical ML models predict outcomes defined by the presence of laboratory test results (*e.g.*, sepsis), a setup that implicitly assumes the absence of a test indicates the absence of disease. When testing patterns differ across subgroups, this introduces disparate censorship, which can yield inequitable model outputs. In an observational study, we identify racial differences in testing rates and show such disparities lead to performance gaps under standard ML training. Naively deployed, these models could reinforce inequitable testing decisions in the delivery of clinical care. We develop an expectation-maximization-based method that reduces bias with minimal loss in predictive performance.

Next, we study outputs of generative ML models when steered toward user goals. Large language models (LLMs) demonstrate the ability to follow natural language instructions, yet the extent to which they can satisfy precise constraints during open-ended generation remains underexplored.

We study steerability in text rewriting, and introduce a framework that defines steerability as goal-conditioned instruction following with minimal unintended side effects. Our framework enables the construction of benchmarks families tailored to specific goals. Applying it to text rewriting, we find that LLMs are consistently un-steerable, exhibiting systematic side effects. We show that reinforcement learning (RL) methods can partially mitigate these effects.

Finally, we examine agent behavior when interacting with deployed systems. We consider the CMS-HCC risk-adjustment model used in Medicare Advantage, which translates diagnosis codes into payments, generating an incentive for insurers to overreport diagnoses. Yet identifying overreporting is challenging without ground truth data. Although identifying overreporting is difficult in the absence of ground truth, we show that insurers can be meaningfully ranked by their propensity to overreport using tools from causal inference. Applying this framework to national Medicare claims data, we uncover heterogeneity in coding practices across insurers suggestive of strategic behavior. We propose a validation strategy that relates these rankings to partially observed indicators of gaming, such as follow-up treatment patterns.

CHAPTER 1

Introduction

Machine learning (ML) models are increasingly deployed as decision-making components in high-stakes settings such as healthcare and public policy [139, 224]. Standard ML development pipelines typically cast learning as the optimization of a fixed objective, such as predictive accuracy in supervised learning or likelihood-based losses in self-supervised pretraining, on some fixed dataset. Model quality is then assessed primarily through aggregate performance metrics aligned with that objective. This standard paradigm has produced highly capable predictive models. However, it provides only a limited characterization of model behavior in practice. Optimizing a single objective under fixed evaluation conditions leaves open many degrees of freedom in how a model represents information, how it responds to user goals or conditioning, and how it behaves when embedded in environments with incentives, oversight, and other agents. Ultimately, good performance under the "standard" evaluation pipeline does not guarantee good performance in real-world contexts. This gap gives rise to misalignment, which may manifest as systematic performance disparities across subpopulations, unintended behavior under goal conditioning, or undesirable behavior by model users [94, 6, 103].

We argue that alignment should be understood as a property of model behavior across multiple modes of interaction, rather than as a single scalar property of a trained model. First, there is the model's default behavior: how it encodes information and produces outputs in the absence of explicit conditioning beyond the training objective, including potential systematic performance disparities. Second is the model's goal-conditioned behavior: how these defaults change when the model is prompted, steered, or optimized toward specific objectives, capturing notions such as controllability, steerability, and side effects. Finally, we consider strategic behavior in response to such models, such as model manipulation or gaming. This captures how existing institutional incentives and oversight mechanisms incentive certain types of user-model interactions. Ultimately, ensuring alignment at all layers of the ML pipeline is key to the responsible deployment of ML systems.

Concretely, this thesis develops tools for detection: in the data used to train models (disparate censorship; Chapter 3), in the outputs of models when steered toward user goals (steerability probes;

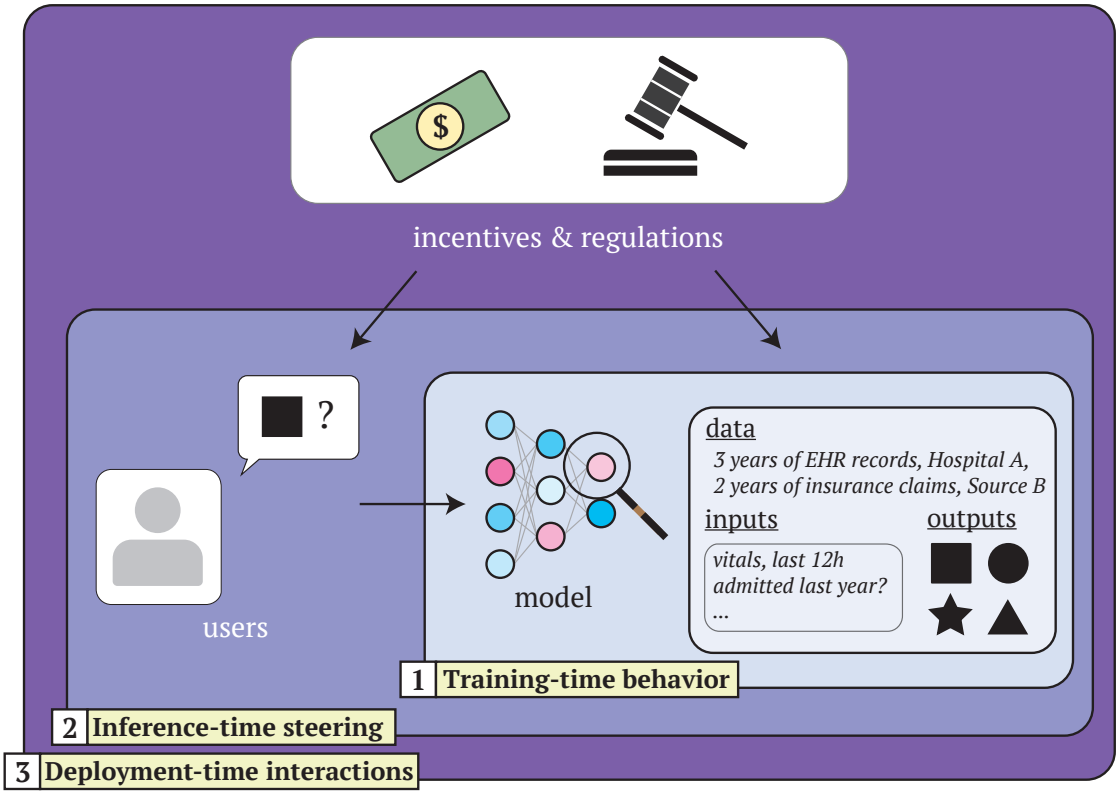


Figure 1.1: Overview of three layers of alignment under study in this thesis: (1) default behavior of the model, which encompasses concerns such as fairness, (2) goal-conditioned model behavior, which concern the ability of the model to satisfy user requests, and (3) strategic behavior, which concern how the underlying incentive and regulatory environment shapes user-model interactions.

Chapter 4), and in the behavior of agents interacting with deployed systems (gaming scores and rankings; Chapter 5). It also explores tools for partial mitigation, whether by modifying the training objective and procedure to correct biased supervision (for disparate censorship), by post-training and evaluation methods to improve controllability and reduce side effects in LLMs (for steerability). While these scenarios are inspired by real-world applications of different ML models in diverse settings, they each provide a window into forms of misalignment that may emerge in practice. Understanding the unique interactions between a model, a user, and a broader system under the proposed taxonomy provides a basis for characterizing alignment.

1.1 Challenges and Opportunities

Here, we discuss challenges and opportunities associated with each scenario under study in this thesis.

Examining data used to train models. We study disparate censorship, which emerges when systematic biases in label collection occur. For example, many ML models are trained to predict outcomes whose definitions depend on laboratory test results (e.g., sepsis depends on a lactate measurement). However, the absence of a test does not always imply the absence of disease. In practice, many models proposed for clinical risk stratification ignore this issue, training models directly on laboratory test results while treating untested individuals as negative [115, 130, 186, 4, 139]. Differences in laboratory testing patterns across subgroups cause systematic biases in label collection, resulting in differential label noise. Without intervention, ML models trained on such data are essentially supervised using potentially incorrect data, where noise in the data varies across subgroups, leading to potential bias in model outputs. However, the extent to which disparate censorship exists in real-world clinical data is underexplored, as is the extent to which disparate censorship would impact the equitability of clinical risk-stratification models when present in data and approaches for addressing this potential disparity.

Examining model outputs when steered towards user goals. While discriminative models are generally evaluated based on predictive performance, conditional generative models can produce open-ended outputs, often guided by user inputs. Such models should be steerable: appropriately responsive to user inputs. For example, a steerable text-rewriting model should be able to produce responses at varying levels of simplicity, formality, or other dimensions. While ample work has focused on steering models to elicit single behaviors of interest, whether by prompting [293, 172], or by intervening on intermediate model outputs [209, 277], steering a model along one dimension may produce unwanted shifts in other dimensions, or *side effects* [6]. There remains a need for a framework for evaluating steerability that accounts for such multidimensional aspects of model behavior.

Agent behavior in response to deployed ML systems. While the ability to control ML models to align with high-level principles and values is ideal, a model whose outputs are overly amenable to steering towards user goals could be exploited via *gaming*. One model for gaming is *strategic adaptation*, in which a user can misreport their features in order to secure a more favorable outcome from a model [103]. A suspected instance of strategic adaptation occurs in the Medicare Advantage public healthcare program in the U.S., in which health plans are reimbursed based on enrollee risk scores derived from diagnosis codes [224]. Insurers have strong financial incentives to inflate these codes to secure higher payments. From a technical standpoint, strategic adaptation and the design of manipulation-proof models are well-studied [103, 70, 170]. However, such methods have been restricted to settings in which agents have known or shared costs and capacities for gaming, or settings where we observe non-strategic behavior for each agent. Gaming in Medicare Advantage is

particularly pernicious, not merely due to the lack of ground truth diagnosis labels, but due to the unknown cost of gaming to insurers, incompleteness of government auditing, and geographically disparate offerings of insurers across U.S. counties. Practical approaches for combating strategic adaptation, including strategies for prioritizing audits, remain underexplored in such settings.

1.2 Contributions

We make the following contributions towards understanding and improving the alignment of ML models with respect to:

- **Equity in clinical risk-stratification models.** In **§3.1**, we conduct a 1:1 matched observational study of racial differences in the order rate of common laboratory tests in emergency rooms at two large academic hospitals. We find tests where both White and Black patients are significantly more likely to be tested, providing evidence of disparate censorship. This contribution was published in PLoS Global Public Health in 2024. We formalize disparate censorship in **§3.2**, with theory showing that performance gaps can only be avoided under narrow conditions, and a simulation study showing that performance gaps between subgroups exhibiting different simulated levels of testing biases emerge otherwise. This contribution was published at the Machine Learning for Healthcare Conference in 2022. In **§3.3**, we build on our theory by developing an expectation-maximization approach to train models to mitigate performance gaps between subgroups while maintaining discriminative performance, leveraging a joint generative model of the relationship between testing biases, covariates, and the outcome of interest. Theoretically, we show that the resulting objective can help compensate for biases in testing. In a simulation study, and a semi-synthetic sepsis risk-stratification task, we empirically show that the proposed approach enjoys favorable tradeoffs between bias mitigation and discriminative performance (i.e., AUROC) compared to related baselines in noisy-label, semi-supervised, and positive-unlabeled learning [178, 171, 199, 21]. This contribution was published at the International Conference on Machine Learning in 2024.

- **Steerability in conditional generative models.** In **§ 4.1** we develop a measurement framework formalizing steerability as a multi-dimensional measurement problem in a space of user intents, and develop an evaluation framework for steerability. We test our framework in a text-rewriting task, and show that various LLMs induce side effects in response to text rewriting instructions, which persist despite interventions in prompt engineering and test-time scaling. A reinforcement fine-tuning (RFT) approach from verifiable rewards [255] shows promise in reducing side effects, but remains a partial solution. This contribution resulted in a

workshop presentation at NeurIPS 2024 (SafeGen AI) and an oral presentation at AAAI 2026 (Alignment Track).

- **Non-manipulability in public health insurance risk-adjustment models.** In **§5.1**, we develop a utility-based model of insurer behavior where each insurer has a different "risk tolerance" for gaming, and show that ranking insurers is equivalent to a causal effect estimation problem. We define a utility model in which each agent has a differing risk tolerance for gaming. We prove theoretically that, exactly identifying an agent's risk tolerance is impossible, but a meaningful ranking is identifiable and equivalent to a treatment effect estimation problem. In **§5.2**, we apply causally-motivated gaming detection to Medicare Advantage claims data in a switchers' cohort of 128,455 Medicare Advantage beneficiaries, identifying contracts that report diagnoses at the highest rates relative to their peers. To address challenges with comparing insurers across disparate geographical areas and conditions, we develop and analyze a method for aggregating adjusted county/market-level diagnosis rate estimates. In our ranking, contracts ranked as higher-risk for gaming tend to provide relevant treatments at a lower rate, suggesting potential gaming. Further simulation studies indicate our ranking is most robust when diagnosis rates are low. These contributions resulted in a publication at NeurIPS in 2024, with a second journal manuscript in preparation.

In sum, these studies highlight that alignment issues rarely stem from a single point in the ML pipeline: they typically reflect interactions between data, model objectives, and the surrounding environment. A common theme is that considering factors beyond the usual training objective can be valuable for understanding model alignment. First, the disparate censorship setting shows how examining data collection and labeling practices offers insight into intrinsic model behavior that standard discriminative training does not explicitly capture. Although our example comes from healthcare, similar questions arise in any domain where labels depend on selective measurement or missingness. Second, our study of steerability shows that looking beyond task performance to how models respond along multiple dimensions can reveal how different aspects of model behavior interact. Even when a model achieves seemingly strong benchmark accuracy, understanding its behavior vis-a-vis steerability across a range of user intents may provide a more holistic picture of performance. Third, examining the incentives faced by users highlights aspects of interaction that standard modeling pipelines do not typically incorporate. While we study misuse of a risk-adjustment model in health insurance, similar considerations apply in settings where deployed systems create uneven opportunities or incentives for manipulation. Together, these themes point to a view of alignment as a multi layer problem in the ML pipeline.

CHAPTER 2

Background

In this chapter, we discuss the main machine learning settings studied in this thesis. In §2.1, to provide background on our study of *disparate censorship* (Chapter 3), or inequitable labeling across data subgroups, we give an overview of the standard supervised learning training and evaluation, and briefly discuss some fairness metrics. In §2.2, to provide background on our study of the steerability of LLMs (Chapter 4), we briefly overview conditional generative models and standard methods for training and controlling LLMs. Lastly, in §2.3, we provide exposition on the U.S. government's Medicare risk-adjustment model under study and the strategic classification framework which we extend in Chapter 5. We use bold lowercase to denote vectors (e.g., $\mathbf{x}$) and regular lowercase for scalars (e.g., x), with upper-cased constants (e.g., T) and random variables (e.g., Y) where specified.

2.1 Bias and fairness in supervised learning

In Chapter 3, we study disparate censorship, a problem setting in which biases in label collection (*e.g.*, laboratory test ordering) propagate to bias. We present a formal definition of supervised learning and bias metrics here.

2.1.1 Supervised learning

Without loss of generality, we study the problem of *risk-stratification*, a subset of supervised learning. Here, we consider the training and evaluation stages.

Training. Formally, consider a model $f : \mathbf{x} \mapsto y$, which maps some features $\mathbf{x} \in \mathcal{X}$ to a real-valued "score" $y \in \mathcal{Y}$. We take $\mathcal{Y} = [0, 1]$ without loss of generality. Such a model f is generally fitted using a dataset of the form $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ using binary labels $y^{(i)} \in \{0, 1\}$, where f is a model *class* (*e.g.*, a neural network, or a tree-based model). Fitting f proceeds by solving the

optimization problem

$$\min_f \sum_{i=1}^{N} \ell(\hat{y}^{(i)}, y^{(i)}) \tag{2.1}$$

where $\hat{y}^{(i)} \triangleq f(x^{(i)})$; *i.e.*, the model's output, for some loss function $\ell : [0,1] \times \{0,1\} \to \mathbb{R}$ that captures the discrepancy or error between the model's outputs and the labels. In binary classification, where $y^{(i)} \in \{0,1\}$, a common choice of ℓ is the binary *cross-entropy loss*:

$$\ell(\hat{y}^{(i)}, y^{(i)}) \triangleq -y^{(i)} \log \hat{y}^{(i)} - (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}). \tag{2.2}$$

We use ℓ to refer to cross-entropy elsewhere except as specified.

Evaluation metrics. There are multiple options for evaluating the performance of a supervised learning model. We focus on the *area under the receiver operating characteristic curve* (AUC), which is a measure of *discriminative performance*: informally, how well the model f is able to "separate" positive from negative examples.

Formally, the receiver operating characteristic curve is defined parametrically as

$$\text{ROC}(f,t) = \text{TPR}(\text{FPR}^{-1}(f,t)) \tag{2.3}$$

and can be interpreted as the true positive rate obtained at the threshold such that the false positive rate of the model f is t for some $t \in [0,1]$. The ROC curve always encompasses the points $(0,0)$ and $(1,1)$.

Then, the AUC can be defined as

$$\text{AUC}(f) = \int_0^1 \text{ROC}(f,t)\ dt. \tag{2.4}$$

Intuitively, this metric aggregates the ability of the model to discriminate positive from negative examples across all thresholds (as swept via $\text{FPR}^{-1}(f,t)$). Intuitively, A perfect AUC is equal to one, while a random AUC is equal to 0.5. It can be shown that the AUC is also equivalent to the probability that the model f correctly ranks a randomly selected positive example *above* a randomly selected negative example; *i.e.*,

$$\text{AUC}(f) = \mathop{\mathbb{E}}_{\substack{\mathbf{x} \sim \mathcal{X}|y=1 \\ \mathbf{x} \sim \mathcal{X}|y=0}} [\mathbf{1}[f(\mathbf{x}) > f(\mathbf{x}')]]. \tag{2.5}$$

2.1.2 Measuring fairness in supervised learning

Certain metrics for evaluating f aim at evaluating the "fairness" of model outputs f under some criterion across some pre-defined subgroups of the population, rather than discriminative performance. In general, fairness metrics in supervised learning require some property of model outputs $f(\mathbf{x})$ to be invariant across subgroups under specific conditions. These conditions are often expressed as independence constraints, which we overview here.

Subgroup-based definitions of fairness. Without loss of generality, we consider the three main classes of fairness definitions in classification problems, also known as statistical non-discrimination criteria [16]. Let Y, A, and $\hat{Y}$ be random variables representing the ground-truth label, the subgroup variable A, and the model prediction, respectively.

- **Independence** ($\hat{Y} \perp\!\!\!\perp A$). Intuitively, independence requires that the model predictions $\hat{Y}$ be completely independent of the subgroup variable A. Thus, the rate of positive (or negative) predictions must be equal across all groups. Such a metric may be reasonable in cases where absolutely no correlation is expected between the subgroup variable and the outcome of interest.

- **Separation** ($\hat{Y} \perp\!\!\!\perp A \mid Y$). Separation requires that, conditioned on the ground truth, that model predictions are identical. For example, the *equalized odds* metric requires that false positive rates be identical across groups [104]. As a relaxation, a model that satisfies separation outputs equivalent predictions for two individuals with equivalent ground-truth "risk" belonging to different subgroups. Such a definition may be reasonable if it is desirable to treat subgroups differently only if justified by evidentiary factors (summarized by Y).

- **Sufficiency** ($Y \perp\!\!\!\perp A \mid \hat{Y}$). Sufficiency requires that, conditioned on model predictions, the distribution of the ground truth Y is identical across subgroups. In that sense, sufficiency is a form of calibration. Sufficiency may be an appropriate criterion for threshold-based decision making using the model f, to ensure that populations selected/not selected by the threshold are similar across subgroups.

AUC-based extensions of fairness metrics. In this thesis, we anchor to AUC-based measures of discriminative performance, an intuitive choice for the clinical risk-stratification tasks under consideration. The first such metric is equality of AUCs. This is a separation-based metric: if AUCs are equal, then, on average across all thresholds $t \in [0, 1]$, the true positive rates across subgroups are equal, similarly to equalized odds.

However, the aggregation across all thresholds $t \in [0, 1]$ may hide cases where, despite equal AUC across subgroups, the resultant model predictions differ dramatically across subgroups at specific thresholds. To address this limitation, variations of fairness criteria, such as the xAUC [138], which is a special case of discriminative performance among the population in different subgroups, or area between ROC curves [86], are similarly separation-based metrics suited to clinical risk-stratification.

2.2 Large language models

In Chapter 4, we study the steerability of large language models, an important class of conditional generative models. We present the formulation of generative models, as well as a brief overview of how generative models are trained and controlled.

2.2.1 Overview of conditional generative models

A generative model produces samples $\mathbf{x}$ from some distribution $P(\mathbf{x})$ supported on some $\mathcal{X}$. Large language models are an important class of *conditional* generative models, where the samples $\mathbf{x}$ are produced from some distribution $P(\mathbf{x} \mid z)$, where z represents a "prompt," which contains some description of the "intended" output. In practice, large language models are trained to produce natural language outputs in response to natural language queries (*i.e.*, the conditioning variables). In other words, z can be expressed in natural language, from which sampling of natural language responses $\mathbf{x}$ follows.

2.2.2 Training large language models

Training LLMs broadly proceeds through three main stages: (i) pre-training, (ii) supervised fine-tuning, (iii) and reinforcement learning. Stages (ii) and (iii) are often together referred to as "post-training." We provide a brief overview of each stage.

Pre-training. During pre-training, LLMs are trained to perform next-token prediction, a self-supervised training objective. The goal of pre-training is to leverage massive self-supervision to learn a strong "base model," from which desirable behaviors can be elicited via post-training. Formally, given a dataset of sequences $(x_1, \ldots, x_T) \sim \mathcal{D}$ of length T (without loss of generality), we optimize

$$\min_{\text{LLM}} \mathop{\mathbb{E}}_{(x_1,\ldots,x_T)\sim\mathcal{D}} \left[\sum_{t=1}^{T} \log \text{LLM}(x_{t+1} \mid x_1, \ldots, x_t) \right], \tag{2.6}$$

where we have formulated an LLM as a sequence model $\text{LLM} : x_1, \ldots, x_t \mapsto x_{t+1}$.

Supervised fine-tuning. Following pre-training, supervised fine-tuning proceeds. The goal of supervised fine-tuning is to endow the base model with the capability to complete tasks or address queries from a user. In general, supervised fine-tuning proceeds with a dataset of question-answer pairs $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N}$, where $\mathbf{x}$ and $\mathbf{y}$ are both sequences of text. Optimization proceeds via

$$\min_{\text{LLM}} \mathop{\mathbb{E}}_{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \sim \mathcal{D}} \log \text{LLM}(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)}). \tag{2.7}$$

Intuitively, supervised fine-tuning aims to ensure that model outputs match factual or other "ground truth" answers. Instruction tuning is an important subset of supervised fine-tuning that formulates question-answer pairs to match potential user queries.

2.2.3 Controlling large language models

Given a pre-trained model, various techniques have emerged for controlling the behavior of LLMs at inference time.

Residual "steering." For open-weight models, where a user can access the underlying weights of the model (*e.g.*, Llama and Qwen series models [72, 271]), one can intervene on model activations to elicit desired behaviors. Activation steering techniques generally select layer(s) of an LLM, and aim to learn a "direction" that encodes the behavior of interest. More formally, let $a_L(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d$ be a d-dimensional activation vector for layer L on prompt $\mathbf{x}$ and completion $\mathbf{y}$, and let $\mathbf{v}$ be the steering vector that encodes a particular behavior. Then, steering is performed by intervening on model activations as follows:

$$a_L(\mathbf{x}, \mathbf{y}) \leftarrow a_L(\mathbf{x}, \mathbf{y}) + \alpha \mathbf{v}, \tag{2.8}$$

where $\alpha \in \mathbb{R}$ is a parameter that controls the strength of the steering intervention.

Multiple methods for learning $\mathbf{v}$ have been proposed. Here, we discuss the "mean difference" approach, a common method for learning $\mathbf{v}$ [209]. Let $\mathcal{D}_+$ and $\mathcal{D}_-$ be datasets of prompts and completions $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N}$ such that $\mathcal{D}_+$ contains completions $\mathbf{y}_+^{(i)}$ that exhibit the behavior of interest conditional on $\mathbf{x}^{(i)}$, while $\mathcal{D}_-$ contains completions $\mathbf{y}_-^{(i)}$ that do *not* exhibit the behavior of interest given the *same* set of prompts $\mathbf{x}^{(i)}$. Then, steering vector $\mathbf{v}$ is given by

$$\mathbf{v} = \frac{1}{|\mathcal{D}_+|} \sum_{\mathbf{x}^{(i)}, \mathbf{y}_+^{(i)} \sim \mathcal{D}_+} a_L(\mathbf{x}^{(i)}, \mathbf{y}_+^{(i)}) - \frac{1}{|\mathcal{D}_-|} \sum_{\mathbf{x}^{(i)}, \mathbf{y}_-^{(i)} \sim \mathcal{D}_-} a_L(\mathbf{x}^{(i)}, \mathbf{y}_-^{(i)}); \tag{2.9}$$

i.e., the difference between the mean activation of responses exhibiting the behavior of interest, and

the mean of responses that do not exhibit the behavior of interest. A similar method extracts the mean positive activation (first term in Eq. 2.9), subtracts the mean activation from the activations for all positive responses, and leverages PCA on the resultant matrix of modified activations, extracting the first principal component as the steering vector $\mathbf{v}$.

The vast majority of steering approaches thus far are designed to elicit single, well-defined behaviors of interest. Steering beyond binary behaviors or jointly controlling multiple behaviors remains difficult, and underexplored in most evaluations of steerability. Furthermore, steering vectors have faced criticism for their inability to generalize beyond the dataset upon which they were developed [33, 267].

Steering by prompting. Large language models also exhibit the ability to follow instructions. Thus, *prompt engineering*, or the design of methods to phrase instructions for a certain intent, has also emerged as a technique for steering model behaviors. Common techniques include *chain of thought* (CoT) prompting, in which the model is prompted to explain its reasoning step by step [293], as well as *few-shot prompting*, in which models are prompted with multiple representative question-answer examples [36]. Recent methods also aim to automatically optimize prompts [245].

2.3 Strategic adaptation

In Chapter 5, we study *gaming*, in which individuals or entities facing a decision made by a machine learning model may aim to *misreport* their features to secure more favorable outcomes.

2.3.1 Example: The CMS-HCC Risk Adjustment Model

The CMS-HCC risk adjustment model is operated by the U.S. government to calculate capitation payments to insurance companies called Medicare Advantage organizations (MAOs) contracted with the U.S. Medicare public health insurance program [224]. The model outputs annual individual-level risk scores, with the aim to produce estimates of relative healthcare cost such that insurance companies are reimbursed in a manner consistent with the health of their enrollees. For exposition, we write a simplified version of the general structure of the CMS-HCC model as follows:

$$y = \boldsymbol{\theta}^\top \mathbf{d} + \boldsymbol{\beta}^\top \mathbf{x}, \tag{2.10}$$

where y is a risk score or "multiplier" intended to reflect the ratio between the beneficiary's cost and that of the *average* beneficiary, $\mathbf{d} \in \{0,1\}^{N_d}$ is a vector of N_d indicator variables that reflect which diagnoses are reported in a patient's record for that year, and $\mathbf{x}$ is a vector of beneficiary demographics (*e.g.*, age, biological sex). $\boldsymbol{\theta} \succeq 0$ and $\boldsymbol{\beta} \succeq 0$ are appropriately shaped coefficient

vectors. For example, if $y = 1.5$, then that beneficiary is expected to cost the MAO approximately 1.5x the average beneficiary in terms of healthcare expenditure in the following year. The payout received by the MAO is proportional to y.

If MAOs are able to increase y without incurring additional healthcare expenditures for their beneficiaries, the MAO could potentially increase their profits. One such method for doing so is known as *favorable selection*, in which MAOs aim to encourage beneficiaries with higher predicted costs are higher than their actual costs to enroll in their plans. Another method, which is the focus of our study, is *coding intensity*. Since MAOs exert some control of the reporting of $\mathbf{d}$, if MAOs can surreptitiously overreport diagnoses $\mathbf{d}$, such behavior can increase MAO profits as well. We subsequently formalize coding intensity under the strategic classification framework.

2.3.2 The strategic classification framework

In strategic classification [103], we assume a model $f : \mathbf{x} \mapsto y$, where $\mathbf{x} \in \mathcal{X}$ are some features and $y \in \mathbb{R}$ is an outcome of interest. Without loss of generality, we assume that higher values of y are more *desirable*. The inputs to the model $\mathbf{x}$ are generated by some *agent* via a function $\Delta : \mathbf{x}^* \mapsto \mathbf{x}$, which encodes how an agent may change model inputs $\mathbf{x}$ from the ground truth values $\mathbf{x}^*$ in order to secure higher values of y. More formally, in strategic classification, an agent $\mathbf{x}$ aims to maximize a utility function of the form

$$\max R(\mathbf{x}; f) - c(\mathbf{x}, \mathbf{x}^*), \tag{2.11}$$

where $R : \mathbf{x} \mapsto \mathbb{R}$ is some *reward* or utility provided by the model f given input $\mathbf{x}$, and c is some *cost* incurred by the agent for perturbing their features from $\mathbf{x}^*$ to $\mathbf{x}$.

In cases where we assume R is concave (but not necessarily strictly) and c is strictly convex, then Eq. 2.11 has a unique maximizer in $\mathbf{x}$. With known R and c, we can infer the optimal action x given ground truth x^* if c is invertible on the domain of x. The CMS-HCC model represents a more difficult setting in which x^* is not known, and c is not known: only the first term $R(x; f)$ may be reliably computed by calculating the captitation payment via the CMS-HCC model. We apply a version of the strategic classification formulation to the CMS-HCC model in Chapter 5.

CHAPTER 3

Disparate Censorship: A Case Study of Biases at Training Time

This chapter examines intrinsic factors of model alignment by studying how unequal opportunities for label collection can lead to inequitable model behavior. First, we provide empirical evidence for disparate censorship by identifying differences in laboratory testing rates between White and Black patients via an observational study in Section 3.1. We then develop and analyze a theoretical framework that models this setting in Section 3.2.3, framing the problem as a special case of biased label collection, which we refer to as *disparate censorship*. Using techniques from analyzing noisy-label learning, we prove that gaps in discriminative performance between subgroups with differential testing or label collection are only avoidable under restrictive conditions. A simulation study shows that discriminative performance gaps quickly emerge under disparate censorship when the theoretical conditions for avoiding performance gaps are violated. We build on our analysis of disparate censorship in Section 3.3 by demonstrating that an expectation-maximization algorithm can mitigate performance gaps that emerge under disparate censorship, with minimal impacts to discriminative performance. Empirically, the proposed approach achieves lower discriminative performance gaps than baselines from noisy-label, semi-supervised, and positive-unlabeled learning, and maintains competitive discriminative performance.

3.1 Racial differences in laboratory testing as a potential mechanism for bias in AI

3.1.1 Introduction

Racial differences in the delivery of clinical care disproportionately harm racial minorities. For example, multiple studies have found lower rates of colon cancer screenings and increased all-cause mortality among Black patients [23]. AI models could amplify biases present in the data, further hindering health equity. In particular, AI models may rely on spurious correlations or confounded relationships to predict outcomes of interest [88, 61]. For example, AI models trained on data that reflect racial health disparities (e.g., differences in healthcare spending) under-recommend Black patients for high-risk care management programs [203]. Without accounting for spurious correlations between race and the clinical outcome of interest, AI models may encode and amplify racial inequities in the delivery of clinical care.

Racial differences in laboratory testing rates are a potentially widespread source of bias in AI models for healthcare settings. Many AI models are developed to predict clinical outcomes in which the pragmatic definition depends on laboratory test results (e.g., sepsis [236, 260]), such that untested patients are assumed to have a normal result. This assumption is often operationalized by assigning untested patients a negative label during model training [4, 115, 139, 108, 130, 186]. However, under this assumption, if laboratory testing rates differ across race, AI models may exhibit disparate performance across racial subgroups [42]. Such models could inappropriately underestimate risk for patients in racial groups less likely to receive laboratory tests, potentially amplifying inequities in clinical care. While previous work [202, 215] has examined racial differences in specific laboratory tests in patient subpopulations, to the best of our knowledge, a large-scale study of racial differences in laboratory testing has not been performed.

We performed a retrospective cohort study of laboratory testing differences between Black and White patients visiting the emergency department (ED) at two large U.S. teaching hospitals. We analyzed the ED since it features a wide range of populations and clinical conditions with ample diagnostic data. Furthermore, patients are not pre-selected for care and generally do not choose their providers. ED patient cohorts are also commonly used in training AI models for predicting clinical outcomes [4, 139, 268, 120, 269]. We quantified testing rates within Black and White patients across seven common laboratory tests with and without matching on age, biological sex, chief complaint, and ED triage score.

3.1.2 Methods

3.1.2.1 Study Population

The two study cohorts included all patients seen in the emergency department (ED) between 2011-2019 at Beth Israel Deaconess Medical Center (BIDMC), Boston, MA via the publicly-available MIMIC-IV dataset (Medical Information Mart for Intensive Care, version 2.0), and between 2015-2022 at Michigan Medicine (U-M), in Ann Arbor, MI [132]. Both institutions are teaching hospitals, and both datasets comprise electronic health records that include demographic and clinical data collected during hospital visit(s), with patient and ED visit identifiers for linking data tables. MIMIC-IV is also one of the most widely used clinical datasets for AI model training. We restrict the analysis to White and Black patients, the two most common racial groups at each institution. We exclude psychiatric admissions due to low rates of laboratory testing. Data for the BIDMC was downloaded as authorized by the data usage agreement on August 15, 2022, and accessed for this study between August 15, 2022, and December 8, 2023. Data for the U-M cohort was accessed for this study between April 13, 2023, and September 21, 2023.

3.1.2.2 Extracting Patient and ED Visit Characteristics

We collected race, age, biological sex, free-text chief complaint, and ED triage scores from the electronic health record based on information gathered in each ED visit. Race was provided as a text string based on information collected during patient registration. At BIDMC, race labels sometimes included nationality. Thus, we defined patients as White if their race was of the form "WHITE" or "WHITE–[nationality]," and likewise for Black patients. Furthermore, BIDMC (but not U-M) records "HISPANIC/LATINO" as a distinct racial category. Thus, we were unable to include White and Black Hispanic/Latino patients in our cohort. At U-M, patients were defined as White if their race was "Caucasian," and Black if their race was "African American." For both institutions, patients with unknown/missing race were excluded from the analysis. At BIDMC and U-M, biological sex was provided via fields called "Gender" and "GenderName," respectively. Since very few records contained values other than "Male" or "Female" (BIDMC: 0 [0.0%], U-M: 14 [0.0%]), we interpreted this variable as biological sex. To extract chief complaint information, the free-text chief complaint was case-normalized and sorted at the word level to merge duplicate complaints (e.g., "abdominal pain, chest pain" vs. "chest pain, abdominal pain"). ED triage score is an integer from 1 (most urgent) to 5 (least urgent), representing a patient's medical needs, typically assigned by an ED triage nurse at intake using pre-specified, structured criteria. ED triage score represents the priority for which ED patients should be treated and is a proxy for illness severity.

3.1.2.3 Outcome Measure

We measured laboratory testing rates for the following common laboratory tests in Black and White patients: complete blood count (CBC) with and without differential, basic or comprehensive metabolic panel, arterial blood gas (ABG), blood cultures, troponin T, brain natriuretic peptide (BNP), and d-dimer.

3.1.2.4 Statistical Analysis

Prior to matching, we compared demographics across White versus Black patients within ED visits in each cohort. We measured the distributions of biological sex (percent female), hospital admission rate (including observational status), the median and interquartile range of age, the ED triage score distribution, and laboratory testing rates. Laboratory testing rate was defined as the percentage of ED visits in which a test was administered. We then performed a 1:1 exact match across White and Black patients on age, biological sex, chief complaint, and ED triage score at the ED visit level within each site to control for variables that potentially explain racial differences in testing rate. We limited the set of matching variables since some candidate matching variables could encode systematic biases (e.g., healthcare utilization). Individuals with no match were excluded. Since ED triage score could be assessed differently in Black versus White patients for the same medical problem, we repeated the matched analysis without ED triage score as a sensitivity analysis. [246] Within matched cohorts, we compared laboratory testing rates between White and Black patients and computed testing differences disaggregated by biological sex and admission status (whether an ED visit resulted in discharge or admission to the hospital).

For hypothesis testing, we used two-sided two-sample z-tests for proportions [79] for differences in rates across race, the Mann-Whitney U-test [184] for differences in age, and the chi-squared test for independence [217] for ED triage score (df=5), treating unknown missing or triage scores as a separate category. We also report the mean and standard deviation for ED triage scores within White and Black patients (no hypothesis tested), since the chi-squared test is not directional. We assessed statistical significance at a 5% level and applied a Bonferroni correction for multiple hypotheses [73] for the primary analysis (N=44). In secondary analyses, we used correction factors N=14 for the subgroup analyses and N=28 for the analysis disaggregated by biological sex. No correction is applied for the sensitivity analysis since it is exploratory. We also compute parametric 95% confidence intervals for all differences in proportions [79]. Full methodological details are provided in the supplementary material.

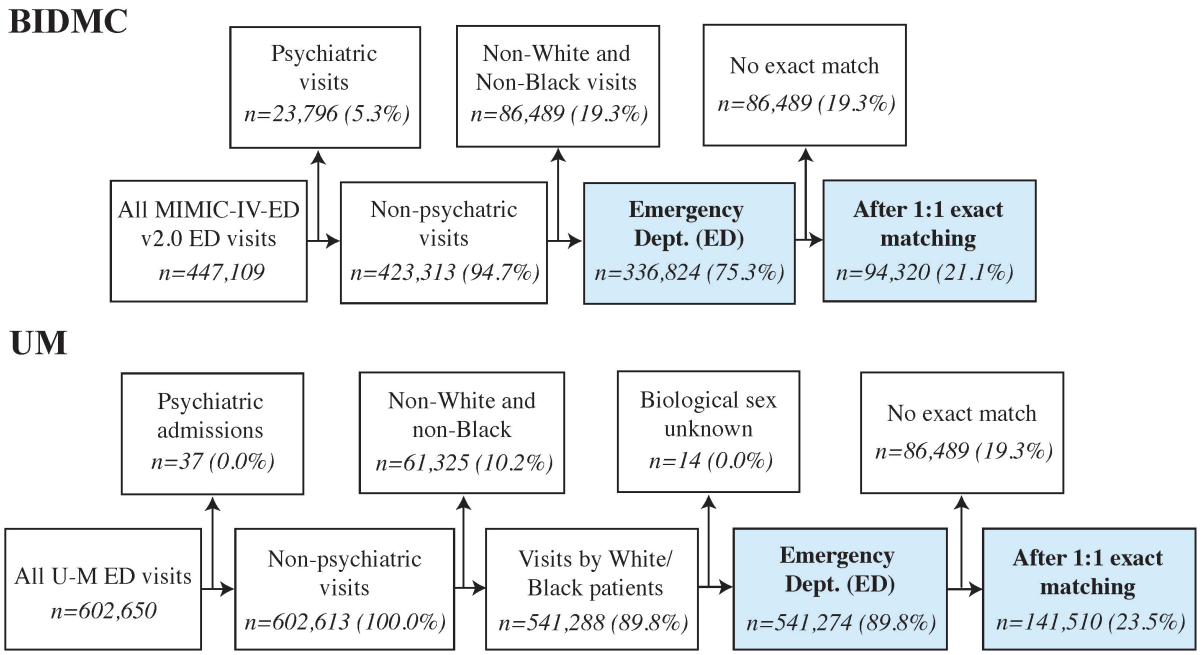


Figure 3.1: Cohort diagram of patient groups selected for study at BIDMC (top) and U-M (bottom). Note that at BIDMC, there were no ED visits in BIDMC associated with individuals of unknown sex. Numbers may not sum to 100.0% due to rounding.

3.1.2.5 Ethics Statement

This analysis was deemed exempt by the U-M IRB (IRBMED) under Exemption 4(iii) at 45 CFR 46.104(d) in the U.S. Code of Federal Regulations, which is applied for secondary research for which consent is not required. Research involving only information collection and analysis as in our study is covered under Exemption 4(iii) of the relevant statute. The exemption included a Waiver of HIPAA Authorization. This study was determined by the IRB to conform with applicable regulations and policies.

3.1.3 Results

After applying exclusion criteria (Fig. 3.1), we analyzed 336,824 and 541,310 ED visits at BIDMC and U-M, respectively (Table 3.1 & 3.2). At BIDMC and U-M, Black patients accounted for 92,437 (27.4%) and 102,313 (18.9%) ED visits, respectively. In the unmatched cohort, White patients

Table 3.1: Cohort demographics for BIDMC among all ED visits for the unmatched cohort.

Variable	White (n=244,387)	Black (n=92,437)	P value
Median age, years (IQR)[a]	55 (35 - 71)	46 (30 - 61)	<0.001
Sex (% Female)[b]	127,213 (52.0)	57,271 (62.0)	<0.001
ED triage score (%)[c]			<0.001
1	15,271 (6.3)	3,847 (4.2)	
2	86,063 (35.2)	24,021 (26.0)	
3	126,126 (51.6)	54,620 (59.1)	
4	13,106 (5.4)	8,796 (9.5)	
5	520 (0.2)	325 (0.4)	
Missing	3,301 (1.4)	828 (0.9)	
Mean ED triage score (SD)[d]	2.57 (0.70)	2.76 (0.69)	N/A
Admitted to hospital (%)	124,310 (50.9)	33,210 (35.9)	<0.001

[a] Interquartile range
[b] Male vs. female were the only two options in the MIMIC-IV dataset.
[c] Percentages may not sum to 100.0% due to rounding.
[d] Missing ED triage scores excluded.

were significantly older than Black patients (BIDMC: 55 vs. 46 years, difference: 9 years, p<.001; U-M: 52 vs. 43 years, difference: 9 years, p<.001) and were less often female (BIDMC: 52.0% vs. 62.0%, 95% CI: -10.5% to -9.3, p<.001, U-M: 53.1% vs. 57.0%, 95% CI: -4.5% to -3.5%, p<.001). The ED triage score distribution also differed significantly across White and Black patients (p ¡ .001). On average, White patients were assessed as more ill than Black patients (i.e., lower ED triage score; BIDMC: 2.6 vs. 2.8, difference: -0.2; U-M: 2.7 vs. 2.8, difference: -0.1). At BIDMC and U-M, 157,520 (46.8%) and 196,918 (36.4%) patients were admitted to the hospital following the ED visit, respectively. White patients were admitted at significantly higher rates than Black patients (BIDMC: 51.0% vs. 36.0%, difference 15.0%, 95% CI: 14.3% to 15.5%, p<.001; U-M: 37.9% vs. 29.8%, difference 8.1%, 95% CI: 7.6% to 8.6%, p<.001). We also report the 25 most common chief complaints by race at each institution (Table A.1 for BIDMC; Table A.2 for U-M, Supplement).

The unmatched analysis also found that White patients were significantly more likely than Black patients to receive all laboratory tests in our analysis except d-dimer (Fig. 3.2, top). For example, White patients were significantly more likely than Black patients to receive a CBC (BIDMC: 69.6% vs. 60.5%, difference 9.1%, 95% CI: 8.5% to 9.8%, p<.001). Fig. 3.2 (top) shows similar findings for metabolic panels, blood cultures, ABG, troponin tests, and BNP (i.e., all other tests analyzed other than d-dimer). Trends were similar at U-M. See Tables A.5 and A.6 (Supplement) for the testing rates shown in Fig. 3.2 (top).

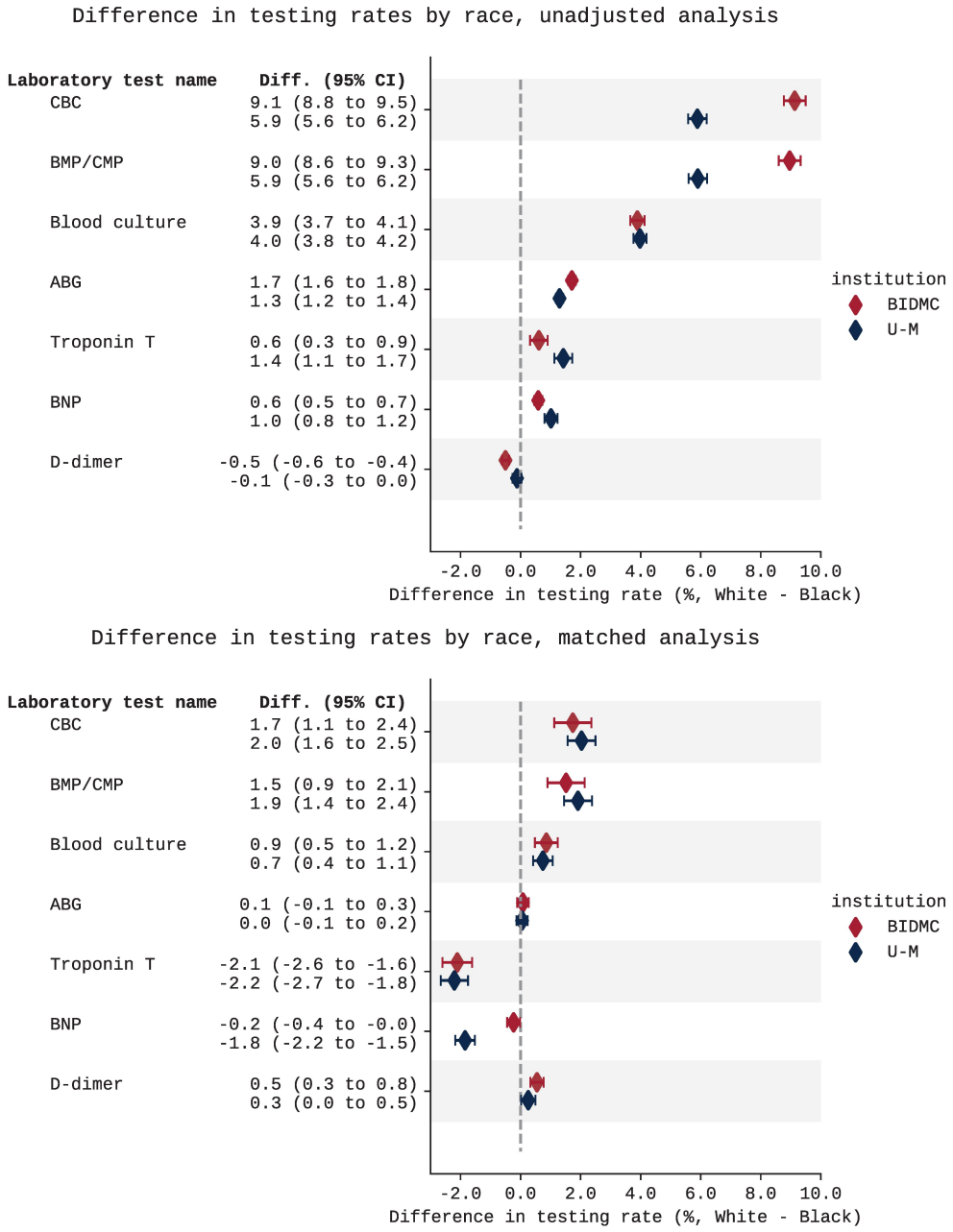


Figure 3.2: Differences in testing rates between White and Black patients, with and without matching on age, sex, chief complaint, and ED triage score. Forest plots show difference in testing rate across White vs. Black patients (White % - Black %), with no matching adjustment (top), and matched on age, sex, chief complaint, and ED triage score (bottom). Differences at BIDMC are shown in red; differences at U-M are shown in blue. Actual values shown in Tables A.5 & A.6 (results with no matching adjustment), or Tables A.3 & A.4 (matched analysis).

Table 3.2: Cohort demographics for U-M among all ED visits for the unmatched cohort.

Variable	White (n=438,997)	Black (n=102,313)	P value
Median age, years (IQR)[a]	52 (34 - 67)	43 (30 - 58)	<0.001
Sex (% Female)	233,022 (53.1)	58,344 (57.0)	<0.001
ED triage score			<0.001
1	5,630 (1.3)	796 (0.8)	
2	177,355 (40.4)	35,016 (34.2)	
3	202,633 (46.1)	49,977 (48.9)	
4	38,815 (8.8)	13,123 (12.8)	
5	3,659 (0.8)	1,018 (1.0)	
Missing	10,913 (2.5)	2,339 (2.3)	
Mean ED triage score (SD)[b]	2.67 (0.69)	2.79 (0.71)	N/A
Admitted to hospital (%)	166,405 (37.9)	30,513 (29.8)	<0.001

[a] Interquartile range

[b] Percentages may not sum to 100.0% due to rounding.

Exact 1:1 matching on age, sex, ED triage score, and chief complaint yielded 94,320 and 141,510 ED visits at BIDMC and U-M, respectively. At BIDMC, this cohort included 51.0% and 19.3% of ED visits by Black and White patients, respectively. At U-M, this cohort included 69.2% and 16.1% of ED visits by Black and White patients, respectively. Matched cohort demographics can be found in Tables A.7 and A.8 (Supplement). Matching narrowed differences in testing rates (Fig. 3.2). White patients remained significantly more likely than Black patients to receive a CBC (BIDMC: 63.0% vs. 61.3%, difference 1.7%, 95% CI: 1.1% to 2.4%, p<.001), metabolic panel (BIDMC: 63.2% vs. 61.7%, difference 1.5%, 95% CI: 0.9% to 2.1%, p<.001), or blood culture (BIDMC: 10.3% vs. 9.5%, difference 0.9%, 95% CI: 0.5% to 1.2%, p<.001), while White patients were significantly less likely than Black patients to receive a troponin test (BIDMC: 17.2% vs. 19.3%, difference -2.1%, 95% CI: -2.6% to -1.6%, p<.001). Differences in laboratory testing were not statistically significant at both institutions for BNP, d-dimer, or ABG. Trends were similar at U-M (Fig. 3.2, bottom). See Tables A.3 and A.4 (Supplement) for the testing rates shown in Fig. 3.2 (bottom). White patients remained significantly more likely than Black patients to be admitted to the hospital after an ED visit at both institutions (BIDMC: 37.7% vs. 35.0%, difference 2.7%, 95% CI: 2.0% to 3.3%, p<.001; Fig. 3.3, left). Trends were similar at U-M (Fig. 3.3, right).

As a sensitivity analysis, we repeated the matched analysis without ED triage scores. Results were similar. White patients were significantly more likely than Black patients to receive a CBC, metabolic panel, and blood culture, and significantly less likely than Black patients to receive a troponin test (Figure A.1 and Tables A.9 & A.10, Supplement). White patients also remained

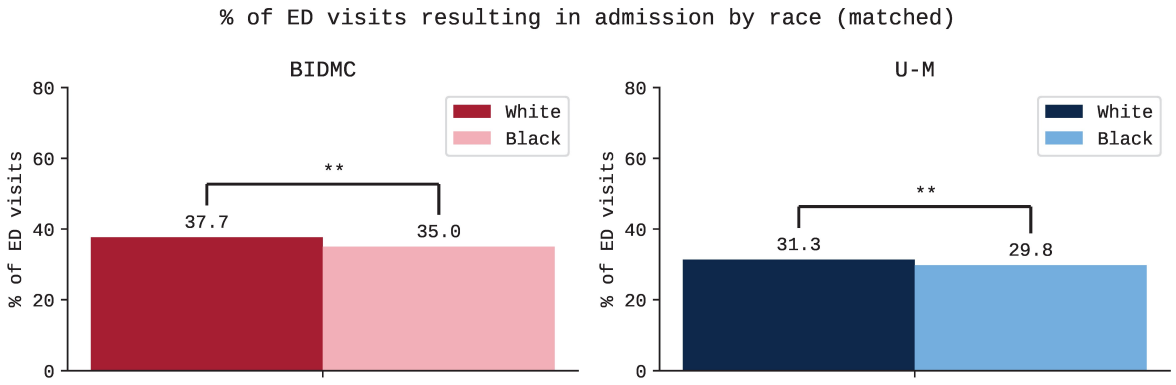


Figure 3.3: Differences in admission rates between White and Black patients matched on age, sex, chief complaint, and ED triage score. Bar plots show difference in rate of admission to the hospital among ED visits in study cohort grouped by site. Differences at BIDMC are shown in red; differences at U-M are shown in blue. "**" denotes statistical significance at 5% level. Actual values are shown in Tables A.3 and A.4.

more likely to be admitted to the hospital (Fig. A.2, Supplement). The analysis disaggregated by biological sex found similar trends: within male and female patients, White patients were significantly more likely than Black patients to receive a CBC or metabolic panel, and significantly less likely than Black patients to receive a troponin test (Figure A.3 and Tables A.11 & A.12, Supplement). Disaggregating by admission status narrowed some testing differences in visits resulting in discharge (Tables A.13 and A.14, Supplement) and visits resulting in admission (Tables A.15 and A.16, Supplement). However, given the significant differences in admission rate across race, narrower testing differences do not necessarily signal less systematic bias.

3.1.4 Evidence of disparate censorship

Before discussing our findings, as an exercise in reflexivity, we as authors acknowledge how our positionality influenced this project. A subset of the authors serve as clinicians in the ED and the ICU at one of the hospitals in our analyses. Another subset of our authors are machine learning

researchers. Several of the authors have prior work on health disparities research. In addition, one author had clinical training at a hospital that cared for a predominantly Black patient population. The authors identify as a White North American woman, White North American men, Asian Indian American man, Asian Taiwanese American man, and a female first-generation immigrant to the United States from Asia. None of the authors have lived experiences with anti-Black racism. It was our prior research in addition to the education received around implicit bias and racial disparities in access to clinical care that inspired, in part, the initial hypothesis. We acknowledge that, as an observational study, our work has some extractive elements, since the authors do not interact with the communities from which data are sourced. Furthermore, our work is unlikely to immediately and directly remediate inequities in the delivery of clinical care, despite our alignment with equitable goals. However, using the power of our positionality as computer scientists and clinicians at a research institution, we aim to amplify cautionary messages about AI with respect to marginalized populations.

We now summarize the findings. Across six of seven laboratory tests, in the unmatched cohort, Black patients had significantly lower testing rates than White patients. After matching to account for differences in clinical characteristics, Black patients remained significantly less likely to receive a CBC or metabolic panel, while White patients were significantly less likely to receive a troponin or BNP test. Hospital admission partially explained racial differences in laboratory testing rates, with Black patients significantly less likely to be admitted to the hospital. Given similar findings across two institutions, these trends may hold more broadly.

Although this study is unable to examine the reasons for laboratory test orders, testing differences could impact AI models trained to predict outcomes defined using laboratory test results. For example, Black patients were found to be significantly less likely to receive blood cultures. They may subsequently be less likely to be diagnosed with serious infections. AI models trained to predict this outcome may underestimate risk in Black patients since models are generally trained to optimize a performance metric (e.g., accuracy) without explicitly accounting for confounding (i.e., between race and the outcome of interest). Specifically, such models could ultimately "learn" that Black patients are less likely to experience serious infections without "accounting" for lower admission rates and lower testing probability.

While quantifying differences in AI model performance across race under biased laboratory testing is not possible without "ground truth" diagnosis labels, studies in similar problem settings have consistently demonstrated the potential for harm. Past work has documented disparate performance in AI models with respect to race in healthcare applications. [203, 126] Furthermore, a simulation study of disparities in testing across population subgroups suggests ample potential for performance gaps to emerge under biased laboratory testing patterns, such as those found in this study. [42] Ultimately, acting on predictions of AI models trained on data exhibiting racial

differences in laboratory testing could reinforce undertesting in marginalized patient populations. Furthermore, since laboratory tests are often used as a source of signal for a variety of downstream diagnoses, even these seemingly small differences could have widespread impacts on the equitable delivery of clinical care.

In general, machine learning researchers do not adjust for confounding prior to training ML models. Potential solutions for addressing selection bias are well-studied but may have limited practical utility for correcting biases in laboratory testing. For example, in the econometric and epidemiology literature, methods such as reweighting or matching are standard in observational studies to adjust for selection bias/survey sampling bias. [110, 143] These approaches are most useful when there is no unobserved confounding, which is difficult to guarantee in practice. One may also consider removing race as an input to the model, but such approaches are insufficient if other variables are correlated with race (i.e., proxy discrimination [58, 226]). In certain situations, it may be appropriate to train AI models on a sub-cohort where the spurious correlation is weaker (e.g., admitted or tested patients). However, this technique may discard most available training data, leading to lower model performance. Furthermore, the resulting model would not apply beyond the sub-cohort used for training (e.g., predicting serious infections in patients without a blood culture, using data from patients who received a blood culture). [120, 230]

Thus, we require novel approaches to mitigate the potential impacts of racial testing differences on AI models. We highlight causally-motivated methods as a promising direction, since they could be used to model mechanisms of bias in laboratory testing and facilitate "counterfactual fairness" analyses (e.g., would a patient with these characteristics be tested regardless of their race). [216] One may also consider extending techniques from semi-supervised/weakly-supervised learning (i.e., treating untested patients as unlabeled instead of negative), many of which assume that label missingness occurs randomly, motivating future methods that account for potentially biased laboratory testing patterns. [299] In parallel, it is imperative that we examine how racial testing differences emerge, to inform policies that minimize biases in laboratory testing, reducing the prevalence of spurious correlations in clinical data. Ultimately, due to the potential for AI to amplify existing harms, one should carefully consider how to adjust for racial differences in laboratory testing in AI models. A "color-blind" approach that ignores correlations between race and testing rates is unlikely to mitigate harm. [295]

Our study has several limitations. First, race is not an objective/unambiguous construct, which is reflected in our data. Some racial categories in BIDMC could overlap with patients who identify as either White or Black (e.g., "Brazilian," "Hispanic/Latino") or mixed-race individuals. We excluded such patients from our study. Our definition of "White" and "Black" captures the majority (79.6%) of all ED visits in BIDMC, mitigating the impact of such ambiguity. There may also be heterogeneity in testing differences within White and Black patients, which this study is unable to

capture. [195] Second, to simplify the analysis, we restricted the study to White and Black patients. We recognize that there may be testing differences across other racial and ethnic groups. Past work highlights the disproportionate burden of racial disparities in clinical care on Hispanic/Latino, American Indian/Alaska Native, and Native Hawaiian/Pacific Islander populations in the U.S. [164]

Third, matching on additional variables (e.g., comorbidities, additional markers of illness severity, patient history) could further explain correlations between race and testing rates. Such variables may also be subject to bias. Furthermore, exact-matching on additional variables could induce statistical bias if many units are dropped, while relaxations to exact matching may yield incomparable matched cohorts. [148] Fourth, our cohorts are drawn from U.S. academic teaching hospitals, which may not accurately reflect standard care across the U.S. Fifth, this study was unable to determine the appropriateness of testing rates. Although some of our findings align with previous work on the disproportionate impact of undertreatment/underdiagnosis on Black patients in various clinical contexts, the differences we observed could represent either under- or over-testing. [23, 215, 118] Further study of racial differences in testing is warranted. Lastly, due to differences in data format, the laboratory test extraction pipelines necessarily differed across institutions. This could result in over/undercounting of tests. These limitations have minor impacts since our main outcome measure is testing differences rather than testing rates.

Black patients were observed to be significantly less likely than White patients to receive multiple laboratory tests across two large academic hospitals. Racial differences in testing narrowed or reversed after matching to account for differences in clinical characteristics. Hospital admission also partially explained these gaps. However, many racial testing differences remained significant. The findings highlight a need to better understand how racial differences in ED laboratory testing emerge, since the patterns observed in this study could represent either over- or under-testing. Furthermore, racial testing differences may amplify bias in AI models, since such models may spuriously associate race with laboratory test outcomes. Standard methods for addressing confounding may be insufficient to mitigate spurious correlations.

3.2 Disparate Censorship & Undertesting: A Source of Label Bias in Clinical Machine Learning

3.2.1 Introduction

Medical applications are increasingly considering the usage of machine learning (ML) models. However, researchers have found that ML models may perform disproportionately poorly on marginalized groups [37, 203, 221]. Biases in training data resulting from spurious correlations between the inputs and outputs have received much attention (*i.e.*, "shortcuts" [88, 126]). However, biases can also arise in the labels used for model training. [203] highlighted one such type of *label bias* in equating healthcare need with healthcare cost led to downstream inequities in clinical care. We hypothesize that label bias may arise from other sources as well. Specifically, many researchers rely on assumptions about labels that could exacerbate pre-existing disparities in healthcare delivery.

For the purposes of ML model training, patient outcomes are often defined based on laboratory/diagnostic test results extracted from the electronic health record (EHR; e.g. [236, 250, 114]), since clinical chart review on large patient databases can be prohibitively costly. In doing so, many researchers assign "negative" labels to untested patients (the *negativity* assumption in positive-unlabeled learning [20]). For example, many sepsis prediction models derive labels from laboratory test-based definitions, such that untested patients are negative by definition [4, 115, 80, 235]. Beyond sepsis, this is also the case when building models to predict healthcare-associated infections [204, 272, 108]. Researchers typically justify this assumption, since without it, a model trained on *only* patients who were tested may only apply to the small fraction of tested patients, limiting its utility.

This common approach to labeling patient outcomes in ML model development can have harmful downstream effects when patient groups are tested at different rates. We refer to this setting as *disparate censorship*, which serves as the focus of our analyses. Examples of disparate censorship in clinical care include lower rates of colon cancer screening for Black patients [68], or biases in cardiac evaluations for women [248]. When disparate censorship occurs in patients of equivalent risk from different groups, one group is *undertested* relative to the other. Undertesting can result in higher rates of missed diagnoses/positives in certain patient group(s) as compared to other group(s), leading to disproportionate harm. In practice, disparate censorship and undertesting can be caused by pre-existing healthcare disparities such as clinician biases [248, 59], different levels of healthcare access or consent [263], or different test performance across groups [84].

While the immediate harm of undertesting and missed diagnoses is clear [182, 23], ML has the potential to amplify this harm. Patient risk-stratification models that do not account for the potential impact of disparate censorship and undertesting may underestimate the risk of the

condition of interest. This could reinforce a harmful feedback loop during ML model deployment, in which models reinforce biased "do not test" decisions.

In this chapter, we characterize when disparate censorship and undertesting can result in ML model performance gaps across patient subgroups. We study three different settings. In the first setting, both the covariates and the drivers of the outcome/disease are drawn from the same distributions for all patient subgroups of interest. In the second setting, the *marginal distribution* of covariates varies across groups. For example, social determinants of health such as education, socioeconomic status, and healthcare access may differ across race and gender [261]. Racial disparities in biomarkers indicating COVID-19 severity have also been documented [225]. In the third setting, the *conditional probability* distribution of the outcome given the covariates may vary across subgroups. For example, the symptoms most indicative of coronary heart disease may differ between female and male patients [175].

First, if the marginal and conditional risk distributions across groups are identical (first setting), neither disparate censorship nor undertesting result in performance gaps. However, in many healthcare settings, it is unlikely that the distribution of covariates and drivers of outcome/disease are identical. When differences in the marginal and conditional risk distributions arise, we show theoretically and validate empirically that disparate censorship can contribute to model performance gaps via certain patterns of undertesting. Then, we identify disparate censorship in clinical data and suggest practical approaches for identifying when disparate censorship may disproportionately negatively impact one group more than another. We encourage the ML for healthcare community to further explore methods for detecting and mitigating the negative effects of disparate censorship and undertesting.

3.2.1.1 Generalizable Insights about Machine Learning in the Context of Healthcare

This chapter highlights and analyzes how disparate censorship and undertesting can result in performance gaps in risk-stratification models when the underlying data generation processes differ across patient subgroups. Our contributions are as follows:

- We introduce "disparate censorship", in which patients are tested at different rates across groups, and formalize how undertesting, or testing disparities in patients with equivalent risk, can lead to gaps in ML model performance across patient subgroups.
- We show that undertesting can lead to model performance gaps across subgroups when certain differences in the marginal and conditional distributions of risk emerge.
- We validate our theory, demonstrating empirically how disparate censorship and undertesting lead to performance gaps across subgroups via a simulation study.

- We identify instances of disparate censorship in clinical data (MIMIC-IV) and suggest practical approaches for mitigating negative impacts.

3.2.2 Problem Setup & Definitions

In this section, we propose a causal model of disparate censorship (Section 3.2.2.1), and outline three settings in which we study the impacts of disparate censorship (Section 3.2.2.2).

3.2.2.1 Causal Model of Disparate Censorship

We formalize disparate censorship using a causal directed acyclic graph (DAG; Figure 3.5). We define the problem using five variables: $a \sim P(A)$ (binary),[1] representing a patient subgroup (*e.g.*, race, biological sex, hospital location), $\mathbf{x} \sim P(X, A)$ (continuous), representing feature vectors/covariates for an individual patient, $t \sim P(T \mid A, X)$ (binary), representing whether a patient was tested for a condition of interest, $y \sim P(Y \mid X, A)$ (binary, unobserved), representing whether a patient has the condition of interest, and $\tilde{y} \sim P(\tilde{Y} \mid Y, T)$ (binary), representing whether the patient tested positive for the condition. We assume X, Y may or may not depend on A. Throughout the paper, we notate subgroup-level versions of various distributions as $P_a(\cdot)$ for $a \in \{0, 1\}$. For example, $P_0(\mathbf{x})$ denotes "the distribution of covariates for patients in group $a = 0$".

The DAG (Figure 3.5) encodes two additional assumptions. First, we assume clinician decisions to test, *i.e.*, T, depend on both A and X. In other words, the level of testing disparity depends on both the values of subgroup A and covariates X. Second, we assume no unobserved confounding between X and Y, such that X contains all direct causes of Y. Later, we will relax our modeling assumptions by allowing either X or Y to causally depend on A (indicated by the dashed blue arrows).[2]

Suppose that we aim to train a patient risk-stratification model. We frame this as a supervised ML task, in which one aims to learn a mapping $s : \mathcal{X} \mapsto \mathbb{R}$, where $\mathcal{X}$ is the support of $\mathbf{x}$, and predicted values of y are generated by thresholding $s(\mathbf{x})$. In our setting, we assume that the true label, y, is unobserved. Instead, we observe $\tilde{y}$ (*i.e.*, a test result), a (potentially) noisy proxy for y. The model s is trained on a dataset $\mathcal{D} = \{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^n$ by minimizing some loss function $\mathcal{L} : (x_i, \tilde{y}_i, s) \to \mathbb{R}$ (*e.g.*, regularized binary cross-entropy loss) that aims to make model predictions $s(\mathbf{x}_i)$ close to the observed labels $\tilde{y}_i$.

We assume that the dataset contains disjoint subgroups $\mathcal{D}_a$, where $a \in \{0, 1\}$. Let $t \in \{0, 1\}$ be a variable indicating whether a patient was tested (0 = no, 1 = yes). To simplify, we assume

[1] In practice, a is often categorical, for which our analysis generalizes, but we restrict a to be binary for simplicity. We leave analyses of non-categorical/overlapping a as future work.

[2] Although A is an exogenous variable, we caution against general interpretations of social categories as inherent/static characteristics.

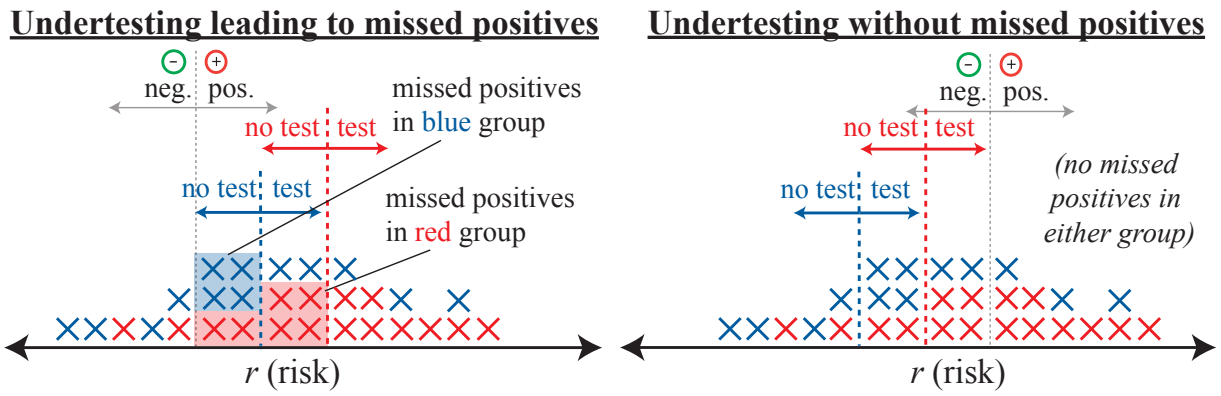


Figure 3.4: Stylized undertesting patterns. In both examples, the red group is undertested (higher risk r required for a test). Left: undertesting is harmful due to the disproportionate rates of missed diagnoses. Right: the absence of missed diagnoses means that undertesting has no disproportionate impact in terms of missed positives. In this chapter, we focus on the scenario on the left.

perfectly accurate tests.[3] We define disparate censorship as follows:

Definition 1 (Disparate censorship.)**.** *Let $P_a(t)$ be the probability that a patient in group a was tested for a condition of interest y. Disparate censorship occurs if $P_0(t) \neq P_1(t)$.*

Under disparate censorship, the true label y is censored/unobserved at different rates in each patient subgroup. Consequently, if a clinician decides to not test a patient for condition y (*i.e.*, $t = 0$), then y is censored, so $\tilde{y} = 0$. Conversely, if $t = 1$, then $\tilde{y} = y$—we observe whether the patient has condition/outcome y.

Similarly, undertesting can be defined as follows:

Definition 2 (Undertesting.)**.** *Define r as a random variable representing the probability of condition y given covariates $\mathbf{x}$; i.e., $r \propto P(y \mid \mathbf{x})$. Let $P_a(t|r)$ denote the probability that a patient in group a with risk r received a test. Without loss of generality, we say that group $A = 1$ is undertested relative to group $A = 0$ if $\int_r \max(0, [P_0(t|r) - P_1(t|r)])dr > 0$.*

This definition captures all *positive* testing gaps between two groups across all levels of risk. Furthermore, this definition demonstrates how undertesting and disparate censorship differ: the absence of disparate censorship does not guarantee the absence of undertesting. However, undertesting can arise due to disparate censorship. In our work, we focus on undertesting such that one group suffers disproportionately high rates of missed diagnoses/positives compared to the other group (Figure 3.4, left). In this stylized example, undertesting as defined raises fewer concerns if it occurs in patients without the condition of interest, since the missed positive rate is already zero (Figure 3.4, right).

[3] In practice, tests may not be perfectly accurate, and may have different sensitivities in each group. Our analysis remains the same, since our argument is based on label noise rates. In such cases, we can define T as a variable indicating whether a patient was tested *and* the correctness of the test result, from which the results follow identically.

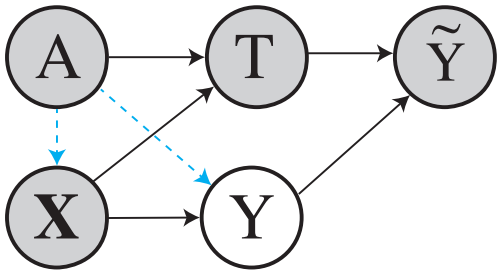


Figure 3.5: Causal DAG used in our analysis. (Un)shaded variables are (un)observed. Patient risk-stratification models aim to predict y given x, but only $\tilde{y}$ is observed. Disparate censorship arises when a affects t. Also, a may affect $\mathbf{x}$ and y (dashed arrows).

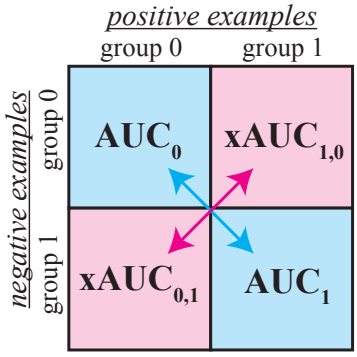


Figure 3.6: Decomposition of overall AUC into within-group AUCs (cyan) and xAUCs (magenta). Zero ranking performance gap between groups a requires parity for both within-group AUCs (cyan arrow) and xAUCs (magenta arrow).

3.2.2.2 Data Generating Processes

We analyze undertesting and disparate censorship in three settings defined by different assumptions for the data generating processes. Within each setting, we assume the same distributional conditions hold at training and inference time.

Setting 1: No difference in marginal and conditional distributions. Formally, $P_0(\mathbf{x}) \overset{d}{=} P_1(\mathbf{x}), P_0(y \mid \mathbf{x}) \overset{d}{=} P_1(y \mid \mathbf{x})$.[4] This means that A has no effect on X or Y. In this setting, both population subgroups have the same distribution of features/covariates, and the same probability of disease given specific covariate values. Note that A is independent of X: not only is A not included in X, but no component of X is affected by A. This is an idealized setting in which there are no differences between subgroups in (1) the distribution of covariates and (2) their relationship to the outcome of interest. However, this is unlikely in practice: even when A is not used as a model input, proxies for A can appear in X such that $P_0(\mathbf{x}) \overset{d}{\neq} P_1(\mathbf{x})$, as discussed in [286, 125].[5] We provide examples of such proxies in the next setting.

Setting 2: Difference in the marginal distribution only. In this setting, $P_0(y \mid \mathbf{x}) \overset{d}{=} P_1(y \mid \mathbf{x})$, but $P_0(\mathbf{x}) \overset{d}{\neq} P_1(\mathbf{x})$. This means that patients with the same covariates are at equal risk for the disease, but the covariate distribution may vary across subgroups a. In causal terms, variable A is a cause of/associated with components of X; the distribution of the covariates of interest varies

[4] $\overset{d}{=}$: Equal in distribution.

[5] $\overset{d}{\neq}$: not equal in distribution.

between patient subgroups. Examples of this setting arise in the study of social determinants of health [261], disparities in hypertension rates [160], or multicenter datasets [25]. In addition, differences in healthcare utilization (alluded to in [225]) and/or consent [263] may also give rise to covariate differences.

Setting 3: Difference in the conditional distribution only. Formally, $P_0(\mathbf{x}) \overset{d}{=} P_1(\mathbf{x})$, but $P_0(y \mid \mathbf{x}) \overset{d}{\neq} P_1(y \mid \mathbf{x})$. In causal terms, this means that variable A is a cause of/associated with Y. In this setting, we make a simplifying assumption that covariate distributions are identical across a. However, patients with the same covariates that belong to different subgroups may have different probabilities of disease. For example, female patients with acute myocardial infarction (heart attack) may present differently than male patients [175].

If both the marginal risk distribution (Setting 2) and the conditional risk distribution (Setting 3) differ (*i.e.*, $P_0(\mathbf{x}) \overset{d}{\neq} P_1(\mathbf{x})$ and $P_0(y \mid \mathbf{x}) \overset{d}{\neq} P_1(y \mid \mathbf{x})$, then negative impacts from either setting would apply as well.

3.2.3 Theoretical Analysis of Disparate Censorship & Undertesting

We develop a theoretical framework for analyzing disparate censorship. We characterize settings under which disparate censorship and undertesting are unlikely to lead to performance gaps across patient subgroups. We focus on ranking performance gaps: differences in within-group area under the receiver operating characteristic curve (AUC) and cross-AUC (xAUC, [138]).

3.2.3.1 Measuring Ranking Performance Gaps

To quantify performance gaps, we focus on ranking metrics, which are frequently used for evaluating clinical risk-stratification models. In this setting, clinicians aim to identify the top-k patients to treat, where k may be determined by resource constraints. To evaluate ranking performance gaps, we report two metrics: (1) the gap in within-group AUC and (2) the gap in cross-AUC (xAUC, [138]).

As intuition, recall that the AUC is the probability that a randomly chosen positive example, $\mathbf{x}_i$, has a greater risk score, than a randomly chosen negative example, $\mathbf{x}_j$ (*i.e.*, $P(s(\mathbf{x}_i) > s(\mathbf{x}_j))$). Then the within-group AUC for group a, written as AUC_a, is the probability that two patients from group a (one positive, one negative) were correctly ranked (Figure 3.6, cyan). The xAUC has a similar interpretation: $\text{xAUC}_{a,a'}$ is the probability that a random positive patient in group a was ranked above a negative patient from group a' (Figure 3.6, magenta). Both within-group AUC and xAUC are key to explaining ranking performance gaps: while within-group AUC only captures misranking error between patients in one group, xAUC quantifies misranking error between groups.

Ideally, group ranking performance is equal across groups (Figure 3.6, cyan arrow), while cross-group ranking performance should be symmetric (Figure 3.6, magenta arrow). This yields the following performance gap metrics—ΔAUC and ΔxAUC (lower is better):

$$\Delta\text{AUC} \triangleq |\text{AUC}_1 - \text{AUC}_0| \tag{3.1}$$

$$\Delta\text{xAUC} \triangleq |\text{xAUC}_{0,1} - \text{xAUC}_{1,0}| \tag{3.2}$$

Assuming *perfect separation*, all patient pairs can be perfectly ranked, and the optimal overall AUC is 1. In such cases, the optimal $\text{AUC}_a, \text{xAUC}_{a,a'}$ values are also 1, so $\Delta\text{AUC} = \Delta\text{xAUC} = 0$ (no performance gap across groups). For details, see Appendix A.2.3.

3.2.3.2 Impact of distributional differences on performance gaps

Here, we prove conditions that lead to zero ΔAUC, ΔxAUC in the presence of disparate censorship and/or undertesting for the three distributional settings studied.

3.2.3.3 No difference in marginal or conditional risk distributions

In this idealized case (*i.e.*, Setting 1), a model would converge to zero performance gap in expectation, even when trained on data exhibiting disparate censorship. In fact, Setting 1 is a special case in which even undertesting, or different testing rates for patients of equal risk, would not result in performance gaps. This is because the two groups are indistinguishable in X, Y, since A is independent of both X and Y, so any risk-stratification model will have identical performance across groups. Thus, ΔAUC, ΔxAUC converge to zero.

3.2.3.4 Difference in marginal distribution only

In Setting 2, the marginal distribution of the covariates varies across groups. We show that undertesting the low-risk group leads to zero ΔAUC, ΔxAUC.

First, we consider a censorship model in which clinicians apply thresholds to clinical risk estimates to make testing/treatment decisions (the "threshold approach" in clinical decision-making [214]). In covariate space, we call such testing thresholds "censorship boundaries." Suppose that clinician decisions to test ($P(T \mid X, A)$) and the true condition indicators ($P(Y \mid X)$) can be written as $\mathbb{1}[s(\mathbf{x}) > (\cdot)]$ for some "scoring function" $s : \mathcal{X} \to \mathbb{R}$. An example is provided in Figure 3.7: note that the censorship boundaries are "parallel" to the true decision boundary. Concretely, we assume that there exists some $\tau_a \in \mathbb{R}$ for $a \in \{0, 1\}$, which act as a group-wise censorship thresholds, such that $t = \mathbb{1}[s(\mathbf{x}) > \tau_a \vee p = 1]$, where $p \sim Bernoulli(c)$ for a suitably

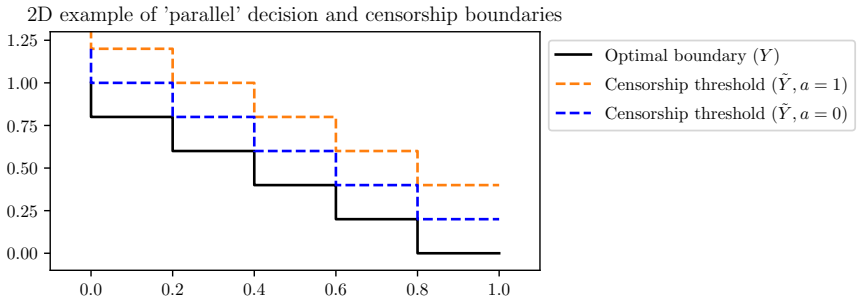


Figure 3.7: Example decision (black) and censorship boundaries (blue, orange) in 2D. In our noise model, we assume that the clinician "test decision" (t) and true "condition status" (y) decision boundaries are based on a threshold with the same functional form (*i.e.* "parallel").

small $c \in (0, 1]$.[6] Furthermore, we assume there exists some $b \in \mathbb{R}$, which acts as a decision boundary, such that $y = \mathbb{1}[s(\mathbf{x}) > b]$. That is; if a patient in group a has risk $s(\mathbf{x})$ greater than b, their true outcome y is 1. The patient is tested for the condition if $s(\mathbf{x}) > \tau_a$; otherwise, they are tested with probability c.

We show that, in Setting 2, when training a model using $\mathbf{x}$, when undertesting results in missed positives, either undertesting the low-risk group or the absence of undertesting (given that $a = 0$ is the low-risk group, $\tau_0 \geq \tau_1$) yields zero performance gap. When undertesting does not result in any missed positives, no performance gap is expected.

Theorem 3 (Zero performance gap under different marginal distributions.)**.** *Suppose that the causal graph in Figure 3.5 is a correctly-specified structural model, A causally affects X, and A is not associated with Y. Assume that Y is perfectly separable in $\mathbf{x}$, and $z \triangleq s(\mathbf{x})$ is distributed for each group as $z \mid a = 0 \sim \mathcal{N}(\mu_0, \sigma^2), z \mid a = 1 \sim \mathcal{N}(\mu_1, \sigma^2)$. Without loss of generality, suppose that $\mu_1 \geq \mu_0$. Then if $\tau_1 \leq \tau_0$ or no positives are missed, the ΔAUC, ΔxAUC of a model using $\mathbf{x}$ as features converges to 0 for suitable values of τ_1.*[7]

This result depends on τ_0, τ_1 directly instead of the absolute difference in testing rates. First, when undertesting does not result in missed positives, disparate censorship and undertesting may not lead to performance gaps. Furthermore, when patients in the low-risk group $a = 0$ undertested with respect to $a = 1$ (or no undertesting is present), the performance gap converges to zero. This is a special case of the "boundary-consistent noise model" [191]. As applied to our setting, this noise model requires that the probability of a missed positive (testing) only decreases (increases) as the risk of condition y increases.

[6] The constraint $c > 0$ ensures that no one is tested with zero probability, or else there may be no signal to learn for sufficiently high censorship thresholds.

[7] In practice, there is a broad range of suitable values for τ_1; see Appendix A.2.1 for details.

3.2.3.5 Difference in conditional distribution only

In Setting 3, the probability of condition y differs across patient groups given features $\mathbf{x}$. We show that, for the case of linear group-wise decision boundaries and censorship boundaries, disparate censorship does not result in a performance gap if the decision boundary for each group is parallel to the censorship boundary (*i.e.,* the two differ by a scalar offset).

Theorem 4 (Zero performance gap under different conditionals.)**.** *Suppose that the causal graph in Figure 3.5 is a correctly-specified structural model, and A causally affects Y. Assume that Y is perfectly separable in $(\mathbf{x}, a)$. For each a: suppose that censorship/testing decisions for group a are expressible as $t = \mathbb{1}[\boldsymbol{\theta}^\top \mathbf{x} + \beta > 0 \vee p = 1]$ for some $\boldsymbol{\theta} \in \mathbb{R}^d, \beta \in \mathbb{R}$, $p \sim Bernoulli(c)$, and all $\mathbf{x} \in \mathcal{X}$. Furthermore, assume that $s_a : \mathcal{X} \times \{0, 1\} \to \mathbb{R}$ has functional form $\boldsymbol{\theta}_a^\top \mathbf{x} + b_a$ for each group a, where $y = \mathbb{1}[s_a(\mathbf{x}) > 0]$ for $\boldsymbol{\theta}_a \in \mathbb{R}^d, b_a \in \mathbb{R}$. If there exists $\delta \in \mathbb{R}, \delta > 0$ such that $\boldsymbol{\theta}_a = \delta\boldsymbol{\theta}$, the $\Delta AUC, \Delta xAUC$ of a model with features $(\mathbf{x}, a)$ converges to 0.*

This result states that a model using both $\mathbf{x}$ *and* a as features may achieve a zero performance gap if, within each group, the corresponding decision and censorship boundaries are "parallel" with one other. We refer to this as the "parallel boundaries" assumption. For intuition, we reason about each group separately. We proceed as if the censorship boundary lies "above" the decision boundary, such that some positives are censored, or else there are no missed positives and the result is immediate. A formal proof can be found in Appendix A.2.1.

The "parallel boundaries" assumption is a special case in which undertesting leading to disproportionate missed positives in one group does not affect ranking. This is because any censorship threshold parallel to the decision boundary preserves relative ordering of patient risk within each group, because higher-risk positives are always less likely to be missed. Thus, if the "parallel boundaries" assumption holds, the resulting pattern of missed positives does not affect ranking. We can generalize this argument to non-linear decision boundaries if the decision and testing boundary parameters are expressible in an appropriate reproducing kernel Hilbert space. However, the "parallel boundaries" assumption is restrictive. We later explore violations of this assumption in our simulation study.

Our theoretical results show that disparate censorship and undertesting do not always lead to performance gaps. If there are no differences across groups in the marginal or conditional distribution of covariates (Setting 1), no gap is expected, even if undertesting is present. When the marginal distribution of covariates differs by group (Setting 2), zero gap is possible if the lower-risk group is undertested (*i.e.*, for $\mu_1 \geq \mu_0$, we must have $\tau_0 \geq \tau_1$). Lastly, when the conditional distribution of covariates differs by group (Setting 3), zero gap is possible if within each group, the censorship and true condition boundaries are "parallel," differing only in an offset

term. However, these conditions are restrictive. In the next section, we empirically demonstrate how performance gaps emerge across distributional settings when differences in the marginal or conditional distributions across groups emerge.

3.2.4 Empirical Analysis of Disparate Censorship & Undertesting

We empirically investigate the impacts of disparate censorship via a simulation study. In this section, we describe the data generation processes and ML modeling details. Then, we present our findings on ranking performance gaps under disparate censorship. In summary, when undertesting leads to missed positives, performance gaps arise when the higher-risk group is undertested (Setting 2) or if testing standards vs. the condition decision boundary increasingly violate the "parallel boundaries" assumption (Setting 3).

3.2.4.1 Simulation data generating process

We simulate a setting in which zero performance gap is theoretically possible when all patients are tested (*i.e.*, perfect separation). Following the theoretical setup, we consider the case of binary $a \in \{0, 1\}$ (without loss of generality). We generate an equal number of "patients" with $a = 0$ and $a = 1$, and simulate 10 covariates $\mathbf{x}$ for each patient by randomly sampling a multivariate Gaussian. We generate true labels y and testing decisions t by applying a non-linear decision boundary s to covariates $\mathbf{x}$ (see Figure 3.8 for a 2D example), where s may depend on $\mathbf{x}$ and a. Concretely, $y = 1$ if $s(\mathbf{x}) \geq 5$, and $y = 0$ otherwise. Similarly, we use one parameter τ_a per group to determine testing decisions: $t = 1$ if $s(\mathbf{x}) \geq \tau_a$, and $t = 0$ otherwise with probability $1 - c$ for some small $c > 0$.[8] Note that values of $\tau_a < 5$ have no effect on censorship, since all $\mathbf{x}$ such that $s(\mathbf{x}) < 5$ are negative by definition and cannot be censored. We then generate observed labels $\tilde{y}$ using y and t, flipping y from 1 to 0 if $t = 0$. Full simulation details are provided in Appendix A.2.2.

By design, this problem is perfectly separable, so an overall AUC of 1 is feasible. Thus, when the conditions of Theorem 3 or Theorem 4 are met, in Settings 2 and 3, respectively, ΔAUC, ΔxAUC converge to zero in expectation.

Simulating testing disparities. Following the theory, we represent decision thresholds for testing each group as τ_0, τ_1 to induce undertesting. Intuitively, τ_0, τ_1 represent different clinical standards for testing. Under the threshold model, in Setting 2, the level of undertesting (as in Definition 2) in group $A = 1$ relative to group $A = 0$ is $(1 - c)(\tau_1 - \tau_0)$. In Setting 3, the level of undertesting depends on both $\tau_1 - \tau_0$ and the similarity between s_0, s_1, so a general form for undertesting does not exist.

[8] See Footnote 6; c ensures that this problem is learnable.

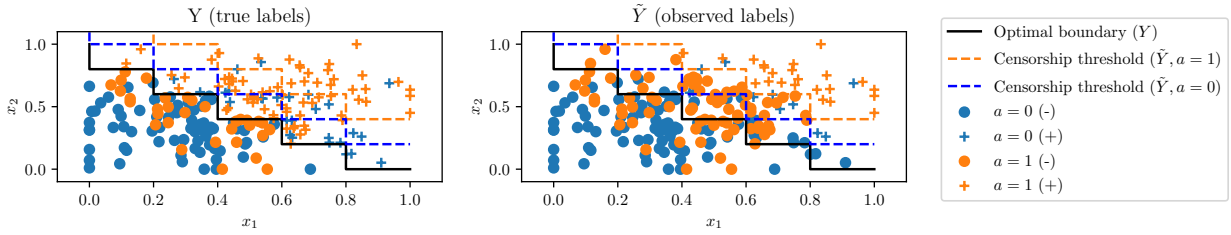


Figure 3.8: Example of disparate censorship in 2D with groups $a = 0$ (in blue) and $a = 1$ (in orange), and decision boundary in black. All blue positives ($a = 0$, +) below the blue dashed line are censored; likewise for orange positives ($a = 1$, +).

Simulating distributional differences. To simulate $P_0(\mathbf{x}) \overset{d}{\neq} P_1(\mathbf{x})$ (Setting 2), we generate covariates $\mathbf{x}$ from multivariate Gaussians with different means, but identical covariance matrices. For $P_0(y \mid \mathbf{x}) \overset{d}{\neq} P_1(y \mid \mathbf{x})$ (Setting 3), we rotate the decision boundary by ϕ degrees, where $\phi \in [0, 360)$.

3.2.4.2 Model Setup for Evaluating Impact of Disparate Censorship

Throughout our experiments, we consider two probabilistic models. The first is trained on simulated true condition labels y, which serves as an upper bound on performance. The second is trained on observed condition labels $\tilde{y}$. This model represents the realistic setting in which test results are used as labels for model training. All models are evaluated with respect to the simulated condition labels y. Since our focus is on model performance gaps rather than the impact of model specification, we describe modeling details in Appendix A.2.4.

3.2.5 Experiments & Results

We examine the impacts of disparate censorship and undertesting under certain distributional differences between groups. We investigate:

- How do disparate censorship and undertesting impact model performance gaps when there are differences in the marginal distribution only? (Setting 2, Section 3.2.5.1)
- How do disparate censorship and undertesting impact model performance gaps when there are differences in the conditional distribution only? (Setting 3, Section 3.2.5.2)

Setting 1 (no marginal or conditional distributional difference) is a special case of Setting 2 and 3 where the levels of marginal or conditional distributional difference are both zero.

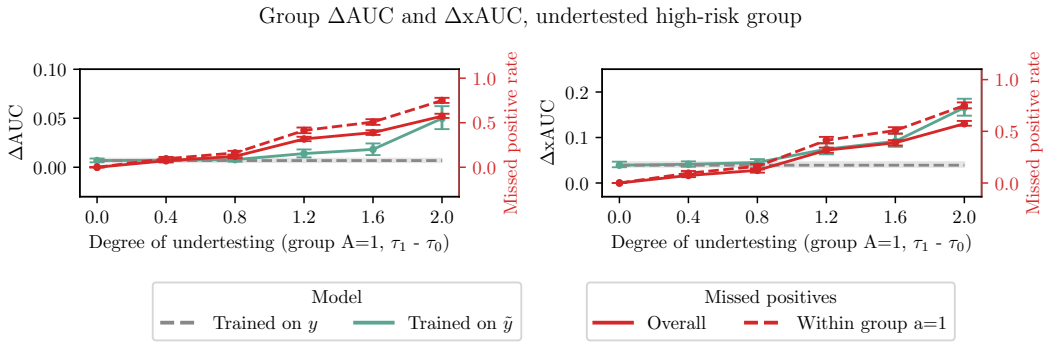


Figure 3.9: ΔAUC (left) and ΔxAUC (right) with 95% empirical CIs for the model trained on y (gray) and the model trained on $\tilde{y}$ (green). As the degree of undertesting experienced by the high-risk group $a = 1$ ($\tau_1 - \tau_0$) increases (dashed red), with a sharply increasing amount of overall missed positives (solid red), ΔAUC, ΔxAUC rises.

3.2.5.1 Undertesting may lead to performance gaps when marginal distributions of covariates differ

In this subsection, we assess the impact of disparate censorship on model performance gaps across groups when the marginal distributions differ across groups. In line with our theoretical results, we find that performance gaps between groups arise when the high-risk group is increasingly undertested.

Undertesting the high-risk patient group results in large model performance gaps. In this experiment, we define group $a = 1$ to be the "high-risk" group. Since censorship occurs at values of τ_0, τ_1 greater than or equal to 5, we set $\tau_0 = 5$ and then vary the amount of undertesting for group $a = 1$ by choosing $\tau_1 \in \{5, 5.4, 5.8, 6.2, 6.6, 7\}$. The degree of undertesting experienced by the high-risk group $a = 1$ corresponds to $\tau_1 - \tau_0$. In this setting, patients in group $a = 0$ are fully tested (no missed positives), while patients in group $a = 1$ are increasingly undertested, leading to more missed positives in group $a = 1$.

At $\tau_0 = 5, \tau_1 = 5$, there is no undertesting; furthermore, there are no missed positives in either group, so the performance gap is near zero, and matches the performance of a model trained using true labels y. In practice, convergence to ΔAUC, ΔxAUC $= 0$ is data intensive in high dimensions due to the prevalence of sparsely-sampled regions near the decision boundary (*i.e.*, the curse of dimensionality). Thus, a small performance gap *independent of the degree of undertesting* is expected, as demonstrated by the constant performance gap for the model trained on y. However, as τ_1 increases, both the rate of missed positives in group $a = 1$ and overall missed positive rate rise sharply. Consistent with Theorem 3, ΔAUC, ΔxAUC increase when the high-risk group is undertested such that missed positives result (Figure 3.9).We also examine increasing τ_0 from 5,

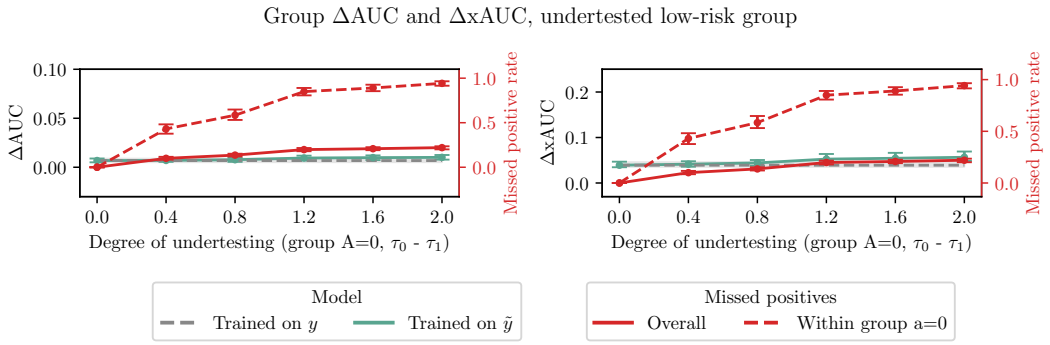


Figure 3.10: ΔAUC (left) and ΔxAUC (right) with 95% empirical CIs for the model trained on y (gray) and the model trained on $\tilde{y}$ (green). Even as the degree of undertesting experienced by the low-risk group $a = 0$ ($\tau_0 - \tau_1$) increases, such that the missed positive rate approaches 1 (dashed red), all metrics stay near oracle performance while the overall missed positive rate plateaus (red).

inducing missed positives in both groups; the correlation between the performance gap and the value of $\tau_1 - \tau_0$ still holds (Appendix A.2.5).

Undertesting the low-risk group results in small performance gaps. Reversing the direction of undertesting changes the impact of disparate censorship. We simulate undertesting in group $a = 0$ by setting $\tau_1 = 5$, and selecting $\tau_0 \in \{5, 5.4, 5.8, 6.2, 6.6, 7\}$. The level of undertesting experienced by group $a = 0$ corresponds to $\tau_0 - \tau_1$ (reversed with respect to the previous experiment). In this setting, group $a = 0$ is increasingly undertested, resulting in more missed positives, while group $a = 1$ is fully tested (no missed positives). As $\mu_1 < \mu_0$, and $\tau_1 \leq \tau_0$, Theorem 3 suggests that the performance gap converges to 0, regardless of the level of disparate censorship—even as the missed positive rate grows.

In contrast to the previous experiment, the number of overall positives in the low-risk group is lower relative to that of the high-risk group, limiting the overall missed positive rate. Thus, even when the low-risk group is severely undertested (high missed positive rate), training with labels $\tilde{y}$ versus y does not significantly affect ΔAUC and ΔxAUC. Indeed, Figure 3.10 shows that the AUC gap is at most 0.01, while the xAUC ranges from 0.04 to 0.06.

In summary, when the marginal distributions of risk vary by group, the harm (in terms of ΔAUC, ΔxAUC) from undertesting the high-risk group is greater than the harm from undertesting the low-risk group. Intuitively, this is because the high-risk group comprises the majority of positive patients, such that undertesting them yields more missed positives. On the other hand, the lower number of positive patients in the low-risk group limits the potential for undertesting to cause missed positives, yielding a smaller performance gap. This result highlights a need to understand whether testing disparities disproportionately impact higher- or lower-risk patient groups. We

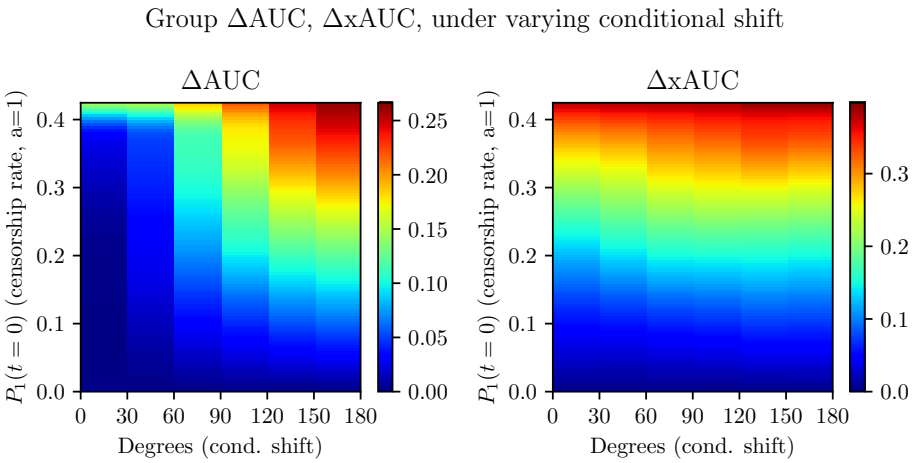


Figure 3.11: Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (ϕ; x-axis) and censorship rate in $a = 1$ ($P_1(t = 0)$, y-axis); 4 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger gaps are in dark red. Performance gaps widen as conditional shift or disparate censorship intensify.

also examined increasing the distributional distance between groups (with τ_0, τ_1 constant); while increasing distribution distance also widens observed performance gaps, overall trends are similar (Appendix A.2.5).

3.2.5.2 Differences in conditional risk distribution also lead to performance gaps under clinician bias

To simulate Setting 3, we vary the level of rotation ϕ from $\{0, 30, 60, 90, 120, 150, 180\}$ applied to d' dimensions of the decision boundary (see Table A.17 for details), and report performance gaps at $\tau_0 = 5$ and varying $\tau_1 \in \{5, 5.4, 5.8, 6.2, 6.6, 7\}$ to induce testing disparities.

We show results for $d' = 4$, but trends are similar for other $d' \in \{2, 4, 6, 8, 10\}$ (Appendix A.2.5). We plot median ΔAUC, ΔxAUC in terms of testing disparity $P_0(t = 1) - P_1(t = 1)$ (equal to $P_1(t = 0)$, since $\tau_0 = 5$), and the level of conditional shift ϕ as a heatmap. As $d' > 0, \phi \neq 0$ violates Theorem 4 assumptions, performance gaps could emerge.

Consistent with Theorem 4, increasing ϕ leads to larger ΔAUC, ΔxAUC (Figure 3.11). As the decision boundary rotates further, more true positives at varying levels of risk are rotated beneath the censorship boundary—increasing the number of missed positives, and ΔAUC, ΔxAUC increase. Thus, even with no marginal distributional differences between groups, standard ML model training using $\tilde{y}$ may result in significant disparities in model performance when conditional distributional differences are present. This result highlights the importance of recognizing disparities in conditional risk distributions, *i.e.*, when the mechanism of condition y varies by group.

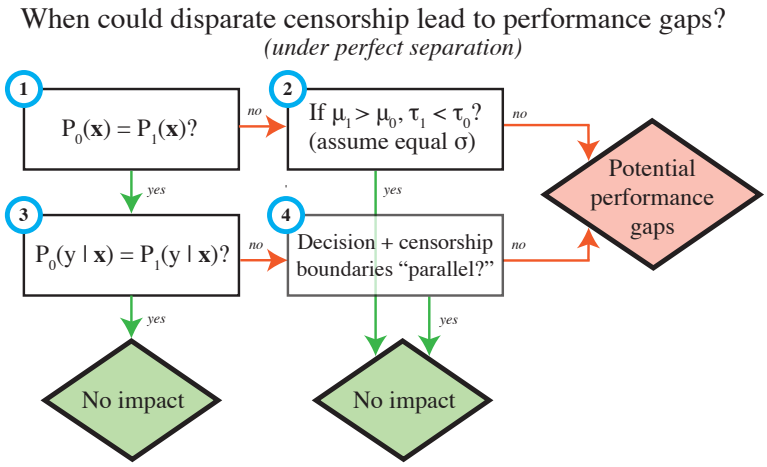


Figure 3.12: Decision tree for identifying when disparate censorship may negatively impact model performance gaps (assuming perfect separation). In summary, if either the marginal risk distribution differs (1) and the high-risk group is undertested (2), or the conditional risk distribution differs (3) and the decision and censorship boundaries are non-parallel (4), performance gaps may emerge (ΔAUC, ΔxAUC > 0).

3.2.6 Potential impacts of disparate censorship

We investigate the impact of disparate censorship and undertesting across different settings. Recall that disparate censorship can lead to harmful *undertesting* if differences in testing lead to disproportionately higher rates of missed positives in certain groups. While missed positives has immediate negative clinical implications, it can also lead to additional harm if these tests are used in defining labels for training ML models. We theoretically show that when the marginal or conditional risk distributions differ, performance gaps may arise if certain patient subgroups are undertested. These findings are further supported through a simulation study. Our findings raise awareness to undertesting and its impact on training labels—which can potentially drive model performance gaps between groups.

Although systematic biases in any part of the ML pipeline can negatively impact model performance gaps, we focus on biases in clinical data caused by disparities in diagnostic/laboratory test orders. Specifically, we identify an understudied type of label bias caused by disparities in the delivery of or access to clinical care. Previous work on label bias in ML for healthcare studied label misspecification: [203, 221] find that using outcomes such as healthcare costs or certain risk scales as proxies for patient need may disproportionately harm Black patients. They suggest that training models on redefined outcomes that align better with patient needs may mitigate the negative impacts.

However, such solutions may be less applicable to *disparate censorship*. In our setting, the

labels correspond to the outcome of interest, but are (partially) observed at different rates for each group. We found such instances of disparate censorship with respect to race in common diagnostic/laboratory tests in the MIMIC-IV dataset, with significantly lower test rates for Black patients as compared to White patients in multiple tests. This finding is particularly concerning due to the wide usage of MIMIC-IV in clinical ML.

Addressing the adverse effects of disparate censorship and undertesting requires thoroughly understanding one's problem setting. Disparate censorship and undertesting raise few concerns in terms of model performance gaps when marginal and conditional risk distributions are identical across patient groups, but this is unlikely to hold in practice. While we discuss methods for identifying when disparate censorship and undertesting may result in performance gaps, there remain gaps in algorithmic approaches for measuring conditional distributional differences. Addressing algorithmic gaps may require domain knowledge from the clinical literature and health disparities research and/or data with complete observations (*i.e.*, no missed positives).

Once disparate censorship and undertesting are identified, noisy-label and censored ML methods represent a possible direction for mitigating negative impacts [131, 49, 24, 289]. Beyond ML methods, modeling techniques in the presence of censored/missing data include inverse probability weighting-based methods in the epidemiology/causal inference literature [116] or the Heckman correction in the quantitative social sciences [111].

In addition to computational solutions, the harmful effects of disparate censorship and undertesting can be minimized by mitigating clinician biases and reducing disparities in covariates across groups. Mitigating clinician biases targets undertesting, towards ensuring that patients with equal risk are equally likely to be tested. Reducing disparities in covariates addresses Setting 2, mitigating model performance gaps that arise due to undertesting the high-risk group. While some covariate disparities may be due to physiological differences (*i.e.*, pediatric vs. adult patients), others emerge due to disparities in healthcare access or structural inequality, disadvantaging various groups such as Black and Hispanic/Latino patients [34], immigrant communities [194], and Black gender minorities intersectionally [169]. These studies further suggest that minimizing covariate differences requires addressing underlying structural inequities in healthcare access and delivery.

The main limitations of our work lie in our theoretical assumptions. First, our simulation design implicitly treats testing as diagnosis. While testing is often a prerequisite to diagnosis, diagnostic decisions may be updated over time, which standard ML development may not capture. For Settings 2 and 3, our theoretical results predicting convergence to zero performance gap require clinician testing and condition status thresholds to be expressed via a hard threshold with the same functional form. For Setting 2 in particular, we apply normality assumptions on risk score distributions and patient covariates. These assumptions are necessary to make the theory tractable. In many cases, we know that clinicians order tests/interventions on the basis of symptoms or clinical suspicion (as

studied in [68, 248]). Thus, while we expect our assumptions to partially hold, it is unclear to what extent they hold in practice.

Nevertheless, our work provides a foundation for a deeper exploration of the impacts of disparate censorship and undertesting. We suggest and identify a plausible mechanism of dataset bias, and demonstrate conditions under which model performance gaps (ΔAUC, ΔxAUC) may arise. Our findings motivate diligence in understanding and mitigating health disparities, and raise warnings about the responsible deployment of ML systems in healthcare. Ultimately, we believe that a combination of computational tools alongside social policy and public health interventions will provide a path to recognize and address the impacts of disparate censorship.

3.3 From Biased Selective Labels to Pseudo-Labels: An Expectation-Maximization Framework for Learning from Biased Decisions

3.3.1 Introduction

Selective labels occur when a decision-making process determines access to ground truth [162]. We study a practical case of selective labels: disparate censorship [42]. Disparate censorship introduces two challenges: different labeling biases across subgroups and the assumption that unlabeled individuals have a *negative* label. For example, in healthcare, labels may depend on laboratory test results only available in some patients.

Past work has trained ML models to predict outcomes based on laboratory test results (e.g., sepsis [250, 236]). In this setting, patients with no test result are defined as negative [108, 272, 130, 186, 4, 139]. However, laboratory testing decisions may be biased. For example, women are undertested and underdiagnosed for cardiovascular disease [19, 248]. ML models trained on such data may recommend women less often for diagnostic testing than men, reinforcing inequity.

To address this bias, one option is to train only on tested individuals. Such an approach may discard a large subset of the data and may not generalize to untested patients. Another option is semi-supervised approaches that do not assume untested patients are negative, such as label propagation [311, 167] or filtering [171, 199], or noisy-label learning methods [30, 290, 312]. However, such methods do not leverage causal models of label bias, a potential source of additional information. We aim to develop an approach that leverages all available signal while accounting for labeling biases.

Inspired by causal models of selective labeling [161, 42, 98], we propose a simple method for mitigating bias when training models under disparate censorship: Disparate Censorship Expectation-Maximization (DCEM; Fig. 5.1). First, we show that DCEM regularizes model estimates to counterbalance disparate censorship. We validate DCEM in a simulation study and a sepsis classification task on clinical data. We find that our method mitigates bias (area between ROC curves) while maintaining competitive discriminative performance (AUC), and is generally more robust than baselines to changes in the data generation process.

3.3.2 Preliminaries: Disparate Censorship

We consider a dataset $\{x^{(i)}, \tilde{y}^{(i)}, t^{(i)}, a^{(i)}\}_{i=1}^{N}$, with covariates $x^{(i)} \in \mathbb{R}^d$, labeling/testing decision $t^{(i)} \in \{0, 1\}$, sensitive attribute $a^{(i)}$, and observed label $\tilde{y}^{(i)} \in \{0, 1\}$, a proxy for ground truth $y^{(i)} \in \{0, 1\}$. The proxy label $\tilde{y}^{(i)} = y^{(i)}$ when $t^{(i)} = 1$, and $\tilde{y}^{(i)} = 0$ otherwise (*i.e.*, $\tilde{y}^{(i)} = y^{(i)}t^{(i)}$).

What is disparate censorship? Disparate censorship models "double standards" in label collection decisions (Fig. 5.1, top). It is a variation of selective labeling or outcome measurement error [162, 99]. Disparate censorship uniquely assumes that untested individuals are imputed as *negative.*

We consider disparate censorship in the context of binary classification [42] (Fig. 5.1, top). We justify the model by example. Consider a patient in an emergency room with characteristics x and sensitive attribute a. This patient may have some condition y (currently unobserved) caused by x but *not* a. A clinician may order a diagnostic test (set t to 1) to determine y. The decision is based on x, but could be swayed by biases in a.

To simplify, suppose that tests are perfectly sensitive.[9] Then, we observe ground truth for tested individuals ($t = 1 \implies \tilde{y} = y$). Otherwise, the patient's label is imputed as negative ($t = 0 \implies \tilde{y} = 0$; *i.e.*, untested patients are presumed healthy). However, due to biases in testing decisions t, y may only be available in a biased subset of the data. The causal model of disparate censorship (Fig. 5.1, top) encodes this decision-making pipeline. Beyond healthcare, disparate censorship may arise whenever potentially biased decisions affect data labeling.

Learning under disparate censorship. We aim to learn a mapping $f_\theta : x \to y$ parameterized by θ optimized for discriminative performance (*i.e.*, AUC), but only observe proxy labels $\tilde{y}$. The default approach for learning under disparate censorship is to assume $y = \tilde{y}$ and proceed using supervised learning. However, such an f_θ may encode labeling biases: estimates of $P(\tilde{Y} \mid X)$ may be inflated compared to $P(Y \mid X)$ for those more likely to be labeled. Thus, biased labeling could yield disproportionate impacts on performance across different subgroups of the data.

Note that we can interpret the estimand of interest as the causal effect of testing on the observed label, since, in the language of do-calculus [216],

$$\begin{aligned}\mathbb{E}[Y \mid X] &= \mathbb{E}[Y \mid X, T = 1] = \mathbb{E}[\tilde{Y} \mid X, T = 1] \\ &= \mathbb{E}[\tilde{Y} \mid X, do(T = 1)],\end{aligned} \tag{3.3}$$

which follows from standard causal identifiability derivations given the causal graph of Fig. 5.1 [124]. Intuitively, testing an individual ($do(T = 1)$) reveals their outcome. Thus, a model trained only on tested individuals could consistently estimate $P(Y|X)$, but may not correct for labeling bias. We discuss other approaches in semi-supervised learning in Section 3.3.6.

[9] If not, we can define T to indicate whether a label is *confirmed* correct. This definition captures differences in test sensitivity across groups (*i.e.*, spectrum bias [196]).

3.3.3 Methodology

We propose Disparate Censorship Expectation-Maximization (DCEM) as an approach for learning in the presence of disparate censorship. We first build intuition for how one could mitigate disparate censorship based on the causal model (Section 3.3.3.1). We then derive DCEM (Section 3.3.3.2) and show that it mitigates disparate censorship via a form of regularization (Section 3.3.3.3). We consider alternative designs and their limitations (Section 3.3.3.4). Detailed proofs and definitions are in Appendix A.3.2.

3.3.3.1 Towards mitigating disparate censorship

Recalling the causal model of disparate censorship, suppose that we are naively training a model f_θ to predict $\tilde{y}$. Define groups a and a' and $x \sim \mathcal{X}$. Consider some $\mathcal{X}' \subseteq \mathcal{X}$ so that

$$\mathop{\mathbb{P}}_{x \in \mathcal{X}'}[T \mid X, A = a] << \mathop{\mathbb{P}}_{x \in \mathcal{X}'}[T \mid X, A = a'] \tag{3.4}$$

for all $x \in \mathcal{X}'$. Define $\hat{t} \triangleq P(T \mid X = x, A = a)$ (*e.g.*, probability of receiving a laboratory test) and $\hat{y} \triangleq P(Y \mid X = x)$. By assumption, $\mathbf{x}$ is sufficient for predicting y (*i.e.*, as in Fig. 5.1, top), such that the optimal $\hat{y}$ should be similar across a (within $\mathcal{X}'$). However, Eq. 3.4 states that group a is *undertested* relative to group a': they have a lower $\hat{t}$ within $\mathcal{X}'$. Equivalently, labeling bias *favors* group a'. Thus, our naive model would underestimate $\hat{y}$ in group a (lower $\hat{t}$ than group a' in $\mathcal{X}'$) relative to group a'.

To counterbalance this bias, one could increase $\hat{y}$ (within $\mathcal{X}'$) where group a is more prevalent than group a'; *i.e.*, lower-$\hat{t}$ regions. Since we are interested in discriminative performance, this is analogous to decreasing $\hat{y}$ where $\hat{t}$ is higher, from which the proposed method follows. More broadly, variables associated with labeling bias (A causally affects T) but not the outcome of interest (A does not causally affect Y) may be useful for mitigating labeling bias.

Given our causal model with latent variable Y (Fig. 5.1, top), we base our approach on expectation-maximization (EM) [63]. We can write:

$$\begin{aligned} &P(\tilde{Y}, Y, T, X, A, U) \\ &= P(\tilde{Y} \mid Y, T) P(Y \mid X) P(T \mid X, A) P(X, A, U). \end{aligned} \tag{3.5}$$

Since y is not fully observed, Eq. 3.5 cannot be optimized via standard supervised objectives. Dropping terms that do not involve Y, we can write the maximization of Eq. 3.5 as

$$\max \; P(\tilde{Y} \mid Y, T) P(Y \mid X). \tag{3.6}$$

Optimizing Eq. 3.6 proceeds via EM. We show that the resulting objectives align with reducing $\hat{y}$ in higher-$\hat{t}$ regions and maintain discriminative performance on tested individuals.

3.3.3.2 Disparate Censorship Expectation-Maximization

Informal overview. EM alternates an expectation step (E-step), which imputes guesses for the latent variable(s) (*i.e.*, Y in Eq. 3.6), and a maximization step that optimizes likelihood given the imputed estimates (M-step, *i.e.*, Eq. 3.6). Our E-step imputes preliminary estimates of $P(Y \mid X)$ for untested individuals. Our M-step updates the estimates to counteract labeling biases when $t^{(i)} = 0$, and is equivalent to full supervision when $t^{(i)} = 1$. The E- and M-steps alternate until convergence. Fig. 5.1 (bottom) shows a schematic of DCEM.

E-step. The posterior of $y^{(i)}$ given the observed data is:

$$Q(y^{(i)}) \triangleq \mathbb{E}[y^{(i)} \mid \tilde{y}^{(i)}, t^{(i)}, x^{(i)}, a^{(i)}]. \tag{3.7}$$

We can rewrite Eq. 3.7 as follows:

Theorem 5 (E-step)**.** *The posterior distribution of $y^{(i)}$ given the observed data is equivalent to*

$$Q(y^{(i)}) = \begin{cases} \tilde{y}^{(i)} & t^{(i)} = 1 \\ \mathbb{E}[y^{(i)} = 1 \mid x^{(i)}] & \textit{otherwise} \end{cases} \tag{3.8}$$

Intuitively, the E-step uses $\tilde{y}$ as the label when we have complete label information (recall $t^{(i)} = 1 \implies \tilde{y}^{(i)} = y^{(i)}$); otherwise, we use the posterior estimate $f_\theta(x^{(i)})$ as a smoothed label. Equivalently, the E-step imputes *soft pseudo-labels* for unlabeled data, *i.e.*, probabilistic estimates $\hat{y}^{(i)} \in [0, 1]$. Motivated by approaches that train a pseudo-labeling model on labeled data [7, 238], we pre-train f_θ on tested individuals.

M-step. The M-step maximizes the log-likelihood of Eq. 3.6 given E-step estimates $Q(y^{(i)})$ (Eq. 3.8). There are two terms to model, which is done via an estimator for $y^{(i)}$ trained using $Q(y^{(i)})$ and an estimator for $\tilde{y}^{(i)}$. The latter is obtained by combining an estimate of $t^{(i)}$ with $Q(y^{(i)})$. Concretely, let $\hat{y}^{(i)} \triangleq f_\theta(x^{(i)})$, and let h_ϕ be a model of $P(\tilde{Y} \mid Y, T)$. Maximizing the log-likelihood

of Eq. 3.6 reduces to

$$\max_{\theta} \sum_{i=1}^{N} Q(y^{(i)}) \log \hat{y}^{(i)} + (1 - Q(y^{(i)})) \log(1 - \hat{y}^{(i)}) \\ + Q(y^{(i)}) \left[\tilde{y}^{(i)} \log h_{\phi}(\hat{y}^{(i)}, \hat{t}^{(i)}) \right. \\ \left. + (1 - \tilde{y}^{(i)}) \log(1 - h_{\phi}(\hat{y}^{(i)}, \hat{t}^{(i)})) \right]. \tag{3.9}$$

This leads to the following result:

Theorem 6 (M-step, informal)**.** *Maximizing Eq. 3.9 also maximizes the evidence-based lower bound of Eq. 3.5.*

In practice, we set $h_{\phi}(\hat{y}^{(i)}, \hat{t}^{(i)}) \triangleq \hat{y}^{(i)}\hat{t}^{(i)}$, a smoothed version of the assumption $\tilde{y} = yt$. Defining $\mathcal{L}$ as binary cross-entropy loss, we can rewrite Eq. 3.9:

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)}) + Q(y^{(i)}) \mathcal{L}(\tilde{y}^{(i)}, \hat{y}^{(i)}\hat{t}^{(i)}). \tag{3.10}$$

Eq. 3.10 can be interpreted as a regularized cross-entropy loss with respect to pseudo-label $Q(y^{(i)})$. The first term pushes $\hat{y}^{(i)}$ towards $Q(y^{(i)})$, while the second "encourages" $\hat{y}^{(i)}$ to be consistent with the causal model. To obtain $\hat{t}$, we pre-train and freeze a binary classifier for t, and take the probabilistic estimates as $\hat{t}$.

3.3.3.3 DCEM counterbalances disparate censorship

We show that DCEM imposes a form of "causal regularization" that lowers $\hat{y}$ in untested individuals.

DCEM is a form of causal regularization. By analogy to regularized risk minimization, consider an objective

$$\mathcal{L}(\theta) + \lambda r(\theta), \tag{3.11}$$

for $\lambda \in \mathbb{R}_{+}$ (regularization strength) and a regularizer $r : \Theta \to \mathbb{R}$, where Θ is the parameter space of θ.

Without loss of generality, setting $\lambda = 1$ and matching terms between Eq. 3.11 and Eq. 3.10 yields $r(\theta) = Q(y^{(i)})\mathcal{L}(\tilde{y}^{(i)}, \hat{y}^{(i)}\hat{t}^{(i)})$. While $\hat{t}^{(i)}$ affects the optimization of Eq. 3.10, it is not a multiplier (*e.g.*, λ in Eq. 3.11). To interpret the effect of $\hat{t}^{(i)}$, we propose a definition of causal regularization strength based how the optimal $\hat{y}^{(i)}$ changes.[10]

[10] "Causal regularization" has been defined in the context of causal discovery [11, 128]. Our usage is unrelated: we use a causal model to regularize an estimator.

Definition 7 (Causal regularization strength, informal)**.** *Let $\hat{y}^{(i)}_{OPT}(Q(y^{(i)}), \hat{t}^{(i)})$ be the minimizer of Eq. 3.10. For $\mathcal{L}$ finite & convex on $\hat{y}^{(i)}$ in [0, 1], the causal regularization strength is $R(\cdot) \triangleq |Q(y^{(i)}) - \hat{y}^{(i)}_{OPT}(Q(y^{(i)}), \hat{t}^{(i)})|$.*

Definition 7 quantifies the tradeoff between matching $\hat{y}^{(i)}$ to the E-step estimates and optimizing Eq. 3.10. While $\hat{y}^{(i)}$ is not an optimization parameter, analyzing the optimal $\hat{y}^{(i)}$ can clarify the inductive bias of the M-step. We proceed by considering how causal regularization impacts untested vs. tested individuals. When $t^{(i)} = 0$, the M-step is

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)}) - Q(y^{(i)}) \log(1 - \hat{y}^{(i)}\hat{t}^{(i)}). \tag{3.12}$$

Since $-\log(1-\hat{y}^{(i)}\hat{t}^{(i)})$ increases in $\hat{y}^{(i)}$, the regularization term "encourages" $\hat{y}^{(i)}$ to decrease when $\hat{t}^{(i)} > 0$. The regularization term is *constant* if $\hat{t}^{(i)} = 0$, such that the M-step would not change the E-step estimate. This matches the intuition that one cannot learn about $y^{(i)}$ from individuals that are very different from labeled individuals (*i.e.*, when the *overlap* assumption in causal inference is violated). The regularization strength depends on $\hat{t}^{(i)}$ as follows:

Theorem 8 (informal)**.** *If $t^{(i)} = 0$, causal regularization strength increases in $\hat{t}^{(i)}$.*

The result implies that **causal regularization counterbalances disparate censorship.** Recall that lowering $\hat{y}^{(i)}$ in regions where $\hat{t}^{(i)}$ is higher can mitigate bias. Equivalently, causal regularization must strengthen as $\hat{t}^{(i)}$ increases, which follows from Theorem 8.

Causal regularization aligns with full supervision in tested individuals. When $t^{(i)} = 1$, the M-step is

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(y^{(i)}, \hat{y}^{(i)}) + y^{(i)}\mathcal{L}(y^{(i)}, \hat{y}^{(i)}\hat{t}^{(i)}), \tag{3.13}$$

substituting $y^{(i)}$ for $Q(y^{(i)})$ and $\tilde{y}^{(i)}$. Thus:

Proposition 9. *Eq. 3.13 is minimized when $\hat{y}^{(i)} = y^{(i)}$.*

Proposition 9 states that causal regularization does not change the M-step optimum from matching ground truth when $t^{(i)} = 1$ (*i.e.*, regularization strength = 0). Thus, the M-step objective aligns with fully-supervised loss.

Thus, the M-step (Eq. 3.10) counterbalances disparate censorship by regularizing $\hat{y}^{(i)}$ towards 0 as $\hat{t}^{(i)}$ increases. For $t^{(i)} = 1$, the M-step optimum stays constant, and DCEM should maintain discriminative performance.

3.3.3.4 Alternative designs and their limitations

We consider two alternative designs and their limitations: directly using $t^{(i)}$ in DCEM and propensity score adjustment.

Why not use $t^{(i)}$ directly? We substitute $\hat{t}^{(i)} = t^{(i)}$ into Eq. 3.10 and analyze one summand (without loss of generality):

$$\mathcal{L}(y^{(i)}, \hat{y}^{(i)}) + y^{(i)}\mathcal{L}(y^{(i)}, \hat{y}^{(i)}) \quad t^{(i)} = 1 \tag{3.14}$$

$$\mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)}) \quad t^{(i)} = 0 \tag{3.15}$$

Both losses use the E-step estimate $Q(y^{(i)})$ as supervision. When $t^{(i)} = 1$ (Eq. 3.14), the M-step adds $y^{(i)}\mathcal{L}(y^{(i)}, \hat{y}^{(i)})$, penalizing false negatives 2x as heavily as false positives. This does not affect ranking metrics (*e.g.*, AUC). When $t^{(i)} = 0$ (Eq. 3.15), the M-step drops causal regularization, and thus cannot counterbalance disparate censorship. Directly using $t^{(i)}$ would only help if counterbalancing disparate censorship is unnecessary for good estimation, *i.e.*, when tested individuals are representative of the population.

Why not propensity score adjustment/related causal approaches? Recall that estimating the effect of T on the observed label yields a consistent estimate of $P(Y \mid X)$ (Eq. 3.3, Section 3.3.2). Indeed, $\hat{t}^{(i)}$ is an estimate of $P(T \mid X, A)$, *i.e.*, a *propensity score*, motivating the usage of causal effect estimators that leverage $\hat{t}^{(i)}$. However, propensity score adjustment (*e.g.*, IPW [242] or doubly-robust variations [239, 281, 121]) require an "overlap" assumption $\eta < \hat{t}^{(i)} < 1 - \eta$ for some $\eta = (0, \frac{1}{2})$ and have asymptotic variance of order $O(1/(\eta \cdot (1 - \eta))$, which is sensitive to extreme $\hat{t}^{(i)}$ (*e.g.*, as in AIPW [95]).

However, in finite-sample settings, "sharp" testing decisions lead to weak overlap. Such extreme $\hat{t}^{(i)}$ may arise in threshold-based decisions [67, 220]. For example, a patient either exhibits or does not exhibit the requisite symptoms to warrant testing. This is analogous to inducing covariate shift between tested and untested individuals. In other words, "holes" in the training data emerge when using only labeled examples. Thus, systematic testing bias could exacerbate model performance gaps across population subgroups. While low overlap still impacts DCEM (since DCEM cannot learn when $\hat{t}^{(i)} = 0$), our method instead leverages an evidence-based lower bound to model y under disparate censorship. We further discuss potential improvements in overlap-robustness of the proposed approach in Appendix A.3.2.

3.3.4 Experimental Setup

We validate DCEM with **synthetic data** across different data-generation processes on simulated binary classification tasks (Section 3.3.4.1) and in a pseudo-synthetic sepsis classification task using real clinical data **(MIMIC-III)** [134], across potential laboratory testing policies (Section 3.3.4.2). We then discuss our chosen baselines (Section 3.3.4.3) and evaluation metrics (Section 3.3.4.4).

3.3.4.1 Synthetic Datasets

By definition, y is not fully observed under disparate censorship. Thus, we design a simulation study in order to evaluate various methods with respect to ground truth. The data generation process follows from the assumed causal model of disparate censorship (Fig. 5.1, top):

$$
\begin{aligned}
&a^{(i)} \sim Ber(0.5), x^{(i)} \sim \mathcal{N}(\mu_a \cdot 1_2, 0.03^2 I_{2\times 2}) \\
&t^{(i)} \sim Ber(\sigma(30 \cdot s_T(x^{(i)}, a^{(i)}))) \\
&y^{(i)} \sim Ber(\sigma(10 \cdot s_Y(x^{(i)}) - c_y)),\ \tilde{y}^{(i)} = y^{(i)} t^{(i)}
\end{aligned}
$$

where $I_{2\times 2}$ is the identity matrix, and $s_T : x, a \to \mathbb{R}$, $s_Y : x \to \mathbb{R}$, and $\mu_a \in \mathbb{R}, c_y \in \mathbb{R}$ are simulation parameters. We set $P(A = 0) = 0.5$ and induce confounding between $x^{(i)}$ and $a^{(i)}$ by setting $u^{(i)} = a^{(i)}$. We draw $x^{(i)} \in \mathbb{R}^2$ from group-specific Gaussians, and assume Bernoulli-distributed $t^{(i)}$ and $y^{(i)}$ with parameters defined via $s_T : x, a \to \mathbb{R}$ and $s_Y : x \to \mathbb{R}$, respectively. Intuitively, s_T (s_Y) is a soft "decision boundary" for T (Y). Inspired by observations that clinician testing is can be represented by simpler functions than observed outcomes [197], we choose a non-linear s_Y and a linear s_T.

We simulate $N = 20,000$ individuals for training, validation, and testing each (*i.e.*, 60,000 total). We define settings in terms of testing disparity $q_t = \frac{P(T|A=0)}{P(T|A=1)}$, prevalence disparity $q_y = \frac{P(Y|A=0)}{P(Y|A=1)}$, and testing multiple $k = \frac{P(T=1)}{P(Y=1)}$. Intuitively, q_t controls labeling biases, q_y controls differences between groups, and k controls testing rate. We consider $q_t \in \{1/4, 1/3, 1/2, 1, 2, 3, 4\}$, $q_y \in \{1/4, 1/3, 1/2, 1\}$, and $k \in \{1/4, 1/3, 1/2, 1, 2, 3, 4\}$, and set simulation parameters to yield the desired q_t, q_y, k.[11]

Since s_Y is unknown in practice, we replicate the main experiments across various s_Y as a robustness check. The simulation makes simplifying assumptions (*e.g.*, low dimensionality and $P(A = 0) = 0.5$) but allows full control over y and t. Additional simulation details are in Appendix A.3.3.1.

[11] We skip settings where q_t, q_y, k yield infeasible testing rates.

3.3.4.2 Clinical data: MIMIC-III

Multiple sepsis definitions, such as Sepsis-3 [260], are based on laboratory tests (blood culture) such that patients without a test result are *by definition* negative. Thus, sepsis classification is a potential real-world case of disparate censorship. We curate a sepsis classification task using the MIMIC-III Sepsis-3 cohort [134, 133], an electronic health record dataset.

We aim to distinguish patients who never develop sepsis from those who develop sepsis within 8 hours of an initial 3-hour observation period. If a patient met the Sepsis-3 criteria between 3-11 hours of the first chart measurement, we set $y = 1$, and $y = 0$ if the patient never develops sepsis during their hospital stay. We exclude patients with onset times outside this range and include only White and Black patients to simplify the analysis of bias mitigation. We choose $x \in \mathbb{R}^{13}$ following an existing sepsis prediction model [62], and exclude patients where all features are missing. This yields $N = 5,301$ patients, from which we create a 60-20-20 train-validation-test split. This is a simplified version of a real clinical task, since we exclude patients who develop sepsis later during their hospitalization. Nonetheless, it is helpful for probing the strengths and weaknesses of the proposed approach.

To evaluate model performance, we assume that the observed y reflects ground truth, since $\approx 90\%$ of patients were tested (*i.e.*, received a blood culture) in our cohort. To generate label proxies $\tilde{y}$, we simulate multiple potential labeling biases via a clinically-inspired testing function s_T. We specify a linear s_T based on qSOFA, a score for triaging patients at risk of sepsis [250]. Inspired by observations that clinicians over-weight representative symptoms in diagnostic test decisions [197], we create different versions of s_T via different weightings of qSOFA features. We examine $k \in \{1/4, 1/3, 1/2, 1, 2, 3, 4, 5\}$ and $q_t \in \{1/2, 2/3, 1, 3/2, 2\}$. Details of the sepsis cohort are in Appendix A.3.3.2.

3.3.4.3 Models

As naive baselines, we test a y**-obs model** (training on $\tilde{y}$) and training on **group** a **only**. We select similarly-motivated or applicable baselines from related settings:

- **Noisy-label learning**: Group peer loss [290] (Appendix: Peer loss [178], truncated ℓ_q loss [305] and generalized Jensen-Shannon loss [77]),
- **Semi-supervised learning**: SELF [199] (Appendix: DivideMix [171]),
- **Causal inference**: **tested-only** (training on examples where $t = 1$), and DragonNet [256], using the treatment effect of the sensitive attribute on testing to correct disparate censorship (*i.e.*, learn a correction for $P(Y \mid X) - P(\tilde{Y} \mid X, A)$),
- **Positive-unlabeled learning:** Selected-At-Random EM (SAREM) [21].

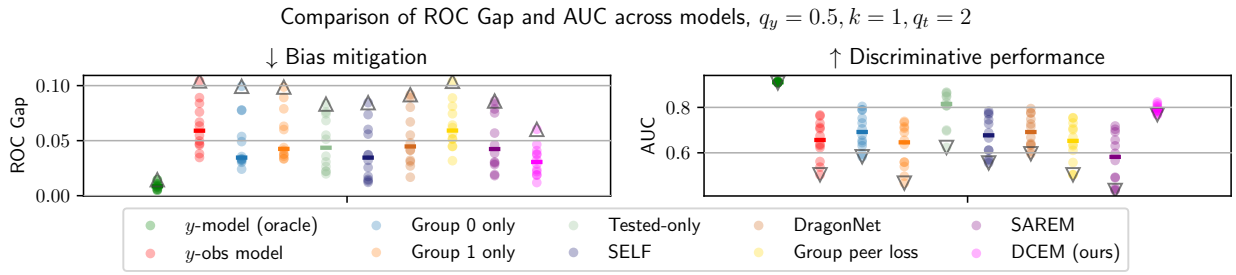


Figure 3.13: Comparison of ROC gap (left) and AUC (right) of selected models at $q_y = 0.5, k = 1, q_t = 2$. Each point represents a different s_Y. Our method (DCEM, magenta) mitigates bias while maintaining competitive AUC compared to baselines, with a tighter range and improved empirical worst-case for both metrics. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

As an oracle, we compare to training on y ("y-model"). We use neural networks for all approaches. Training details, such as hyperparameters, are in Appendix A.3.4.

3.3.4.4 Evaluation metrics

We consider bias mitigation and discriminative performance metrics with respect to y, and measure the robustness of both metrics to changes in the data-generation process.

Discriminative performance. We use the area under the receiver operating characteristic curve (AUC), a standard discriminative performance metric.

Mitigating bias. We use the ROC gap (also called ROC fairness [284] or ABROCA [86]), the absolute area between the ROC curves for each group a. The ROC gap is in [0, 1]. Lower values indicate better bias mitigation. Intuitively, the ROC gap is zero when a classifier with some fixed false positive rate in each group obtains equal true positive rates across groups. Under disparate censorship, a zero ROC gap is achievable if a model *perfectly* predicts y from x.

Robustness. We consider the median AUC and ROC gap over all s_Y (synthetic data setting) or s_T (sepsis classification) and the empirical **worst-case** (AUC: min.; ROC gap: max.) and **range**.

3.3.5 Experiments & Discussion

Our experiments aim to substantiate our main claims:

- In synthetic data, DCEM mitigates bias, maintains competitive discriminative performance and improves robustness, while achieving better tradeoffs between performance and bias mitigation compared to baselines (Section 3.3.5.1).

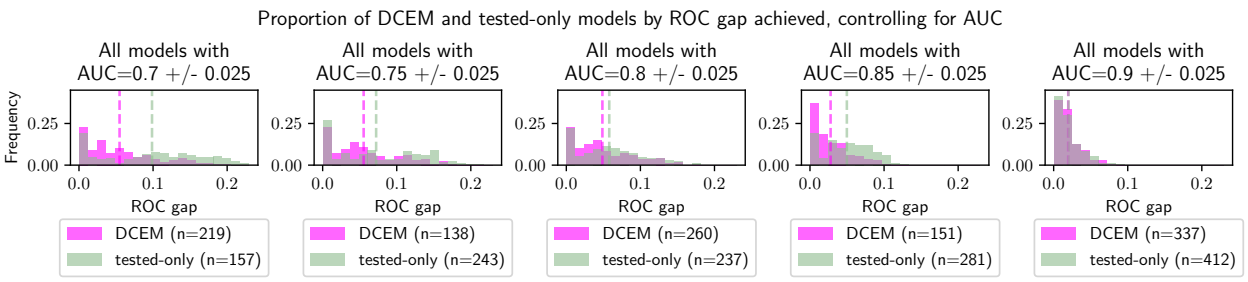


Figure 3.14: Relative frequencies of ROC gaps for DCEM vs. tested-only models at similar AUC (increasing to the right), pooling models across all k, q_y, q_t tested. Dashed lines = mean ROC gap by model. DCEM improves bias mitigation among models with similar AUC.

- On a sepsis classification task, DCEM improves discriminative performance while maintaining good tradeoffs with bias mitigation, and is more robust compared to baselines (Section 3.3.5.2).

We also report full results (Appendix A.3.5.1) and an ablation study of DCEM (Appendix A.3.5.2). We also benchmark causal effect estimators (*i.e.*, as alternatives to the tested-only model) and their overlap robustness compared to DCEM (Appendix A.3.5.3). Further sensitivity analyses can be found in Appendix A.3.5.4 (smoothed $\hat{t}^{(i)}$) and A.3.5.5 (E-step initialization).

3.3.5.1 Results on simulated disparate censorship

Fig. 3.13 shows ROC gaps (left) and AUCs (right) of the baselines most competitive with our approach (DCEM, magenta) at $q_y = 0.5, k = 1, q_t = 2$. In this setting, 25% (*i.e.*, $k \cdot P(Y = 1)$) of individuals are tested, and the base rate of Y in group $a = 0$ is $1/2$ that of group $a = 1$, but group $a = 0$ is twice as likely to be tested. Each point is an ROC gap/AUC value achieved under one decision boundary s_Y. Results for the remaining baselines are in Appendix A.3.5.1. The takeaways align with the main results.

DCEM mitigates bias more effectively than baselines. DCEM achieves a median ROC gap of 0.030 (2nd-best, SELF: 0.034), suggesting that it mitigates bias more effectively than baselines (Fig. 3.13, left). We show similar trends for $q_t \geq 1$, q_y, and $k \in [1/3, 2]$ (Appendix A.3.5.1). At low testing rates, all models mitigate bias poorly. At high testing rates, the tested-only model is sufficient.

For $q_t < 1$ (Appendix A.3.5.1), DCEM mitigates bias compared to the default approach (y-obs model) but no longer dominates the baselines. We hypothesize that DCEM has similar bias mitigation capabilities as baselines, since there is less bias to mitigate. Recalling that $q_y < 1$, since $q_t < 1$, testing probability, $P(Y \mid X)$ *and* $P(\tilde{Y} \mid X)$ are correlated. Learning to predict $\tilde{Y}$ would preserve ordering in $P(Y \mid X)$, reducing impacts on ranking metrics (*e.g.*, ROC gap).[12]

[12]Such settings are related to boundary-consistent noise; see Proposition 1 of [191].

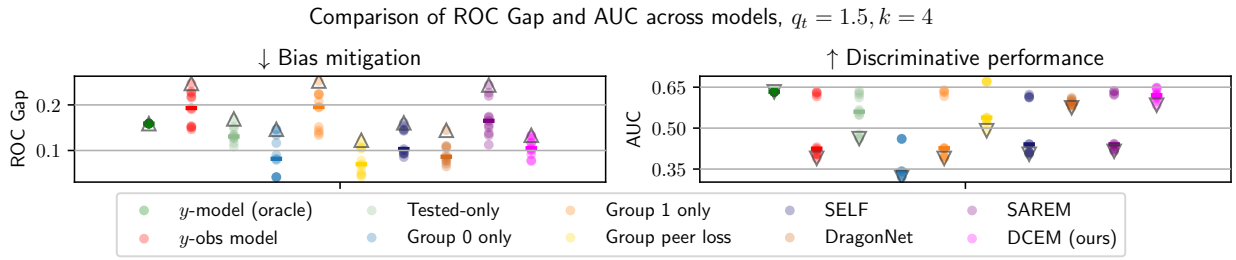


Figure 3.15: ROC gaps (left) and AUC (right) of baselines and DCEM on sepsis classification task at $q_t = 1.5, k = 4$. Each dot represents a different s_T. Our method (DCEM, magenta) maintains competitive or better bias mitigation and discriminative performance compared to baselines. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

DCEM is more robust than baselines to changes in the data-generating process. Fig. 3.13 (left) shows that the maximum ROC gap is lower for DCEM compared to baselines (ours: 0.060 vs. 2nd-best, tested-only: 0.083). We report similar results for the minimum AUC (Fig. 3.13, right; ours: 0.768 vs. 2nd-best, tested-only: 0.623). DCEM also achieves a tighter ROC gap and AUC range. Fig. 3.13 also shows that our method has the tightest ROC gap range (left, DCEM: 0.048 vs. tested-only, 2nd-tightest: 0.063) and AUC range (right, DCEM: 0.055 vs. DragonNet: 0.199).

The results suggest that DCEM maintains robust bias mitigation and discriminative performance across different data-generation processes (s_Y). This is expected, as DCEM optimizes likelihood under the disparate censorship data-generation process by design. In contrast, the baselines may experience selection bias or misspecification, since they discard data or assume certain noise structures/variable dependencies that disparate censorship violates.

DCEM maintains competitive discriminative performance. Fig. 3.13 (right) shows that DCEM outperforms all baselines except for the tested-only model, which our approach lags by 0.028 AUC (DCEM: 0.787 vs. tested-only: 0.815). Other causal estimators achieve similar discriminative performance to the tested-only approach (Appendix A.3.5.3). However, our method improves on the "default" y-obs model, increasing the median AUC by 0.130 (y-obs: 0.657). SELF, which has a similar median ROC gap to DCEM, underperforms DCEM by 0.110 AUC (SELF: 0.677 vs. DCEM: 0.787). Other baselines also underperform. This is expected, since some methods ignore label bias: training on $\tilde{Y}$ alone is misspecified for $E(Y \mid X)$, since it incorrectly assumes that if $T = 0$, then $Y = 0$. The same argument applies to Group 0/1 only approaches.

Some baselines account for label noise/bias, but are mis-specified under disparate censorship since they make different independence assumptions. Group peer loss assumes $T \perp\!\!\!\perp X \mid (Y, A)$, and SELF assumes $T \perp\!\!\!\perp X \mid Y$, ignoring the dependence of biased selective labeling on X. DragonNet accounts for X by adding $P(T \mid X, A = 1) - P(T \mid X, A = 0)$ to the default model's estimates (i.e., $P(\tilde{Y} \mid X)$) as a "correction factor." However, the correction factor may be biased for true negatives:

the oracle is zero, because $P(\tilde{Y} \mid X) = P(Y \mid X)$, but $P(T \mid X, A = 1) - P(T \mid X, A = 0) \neq 0$ in general under systematic labeling bias.

SAR-EM, which is most similar to the proposed approach, models missingness at random (i.e., $T \not\perp\!\!\!\perp X, Y$), but discards reliable negatives. In contrast, the proposed approach incorporates reliable negatives in alignment with our assumptions about labeling bias, allowing it to counterbalance biased selective labeling. Trends are similar for other q_t, q_y and $k \in [1/2, 2]$. Since the tested-only model is a strong baseline, we now compare it directly to DCEM.

DCEM achieves better tradeoffs between discriminative performance and bias mitigation. Among models with similar AUC where AUC < 0.875, DCEM reduces ROC gaps compared to the tested-only model (Fig. 3.14). For example, for AUC $\in (0.825, 0.875)$ (Fig. 3.14, 2nd from right), DCEM improves the average ROC gap by 0.022 (0.028 vs. 0.050), with similar trends at lower AUCs. Among the best-performing models (AUC > 0.875; Fig. 3.14, 1st from right), both methods have similar ROC gaps.

The results suggest that DCEM is not trading discriminative performance for bias mitigation. At a given AUC, DCEM more often yields models with a lower ROC gap than the tested-only model. Since the tested-only approach does not account for label bias, it can achieve relatively high AUC without mitigating bias. In contrast, DCEM explicitly counteracts disparate censorship. A similar comparison to SELF shows that, at low ROC gaps, DCEM likewise finds higher-AUC solutions than SELF (Appendix A.3.5.6).

3.3.5.2 Results on sepsis classification in MIMIC-III

DCEM has better discriminative performance than baselines. Fig. 3.15 compares the ROC gap and AUC of DCEM to selected baselines at testing disparity $q_t = 1.5$, and testing rate multiplier $k = 4$. Each dot corresponds to one variation of s_T (laboratory testing policy). Our method has the highest median AUC among baselines (ours: 0.620 vs. DragonNet: 0.593), nearing the oracle (y-model, 0.633). Note that DCEM has better discriminative performance than the tested-only approach, suggesting that extrapolation from tested to untested individuals is more difficult on the sepsis classification task than the fully synthetic tasks.

DCEM achieves good tradeoffs with bias mitigation. DCEM achieves a smaller median ROC gap compared to five of eight baselines tested. Group peer loss, DragonNet and the Group 0 approach achieve lower median ROC gaps of 0.070, 0.088 and 0.082, respectively (DCEM: 0.105). However, the Group 0 approach catastrophically fails (median AUC: 0.342). Models may perform arbitrarily poorly under disparate censorship if labeling biases sufficiently "conceal" the true decision boundary. Group peer loss (among many other baselines) also exhibits a much wider AUC range than the

proposed approach (Group peer loss: 0.182 vs. DCEM: 0.065), suggesting that its discriminative utility may be limited. DragonNet appears competitive (0.027 AUC lower than DCEM), but would only perform well when the effect of race on testing is close to $|P(Y \mid X) - P(\tilde{Y} \mid X, A)|$, which is violated if labeling biases (large effect of race on testing) are present in negative patients ($P(Y \mid X) \approx P(\tilde{Y} \mid X, A)$).

Many approaches, including DCEM, obtain a lower ROC gap than training on y. Although the oracle obtains the highest median AUC, optimizing discriminative performance on y is not always guaranteed to mitigate bias. DCEM uses labeling probabilities to mitigate bias via causal regularization, while DragonNet directly uses an estimate of the labeling bias as a correction factor. Thus, the results validate that the labeling bias can provide signal for bias mitigation.

DCEM is more robust than most baselines to changes in s_T. DCEM maintains robust bias mitigation capabilities across s_T; *i.e.*, differences in how labelers weigh features in their decisions. Fig. 3.15 shows that DCEM attains a maximum ROC gap of 0.133 (left; DragonNet: 0.144), and a minimum AUC of 0.584 (right; DragonNet: 0.574). Fig. 3.15 also shows that DCEM achieves the tightest ROC gap range (left; DCEM: 0.094 vs. 2nd-best: 0.102) and 2nd-tightest AUC range (right; DCEM: 0.065 vs. DragonNet: 0.018). Many baselines also exhibit a *bimodal* empirical AUC distribution and only perform well under specific labeling behaviors. We examine the sensitivity of baselines to s_T by plotting AUC and ROC gaps against coefficients of s_T (Appendix A.3.5.7).

While DragonNet is competitive on this dataset, its robustness and performance capabilities may not generalize (*e.g.*, simulation results, Fig. 3.13). DCEM is the only approach tested that achieved competitive discriminative performance and bias mitigation on both datasets. Trends in performance and robustness are similar for other k, q_t (Appendix A.3.5.1).

Overall takeaways. In a simulation study of disparate censorship, DCEM mitigates bias while achieving similar or better discriminative performance compared to baselines. The proposed approach is empirically more robust than baselines to changes in the data-generating process. On a sepsis classification task, DCEM mitigates bias while improving or maintaining discriminative performance compared to baselines across different labeling behaviors. Thus, DCEM can potentially mitigate bias with less impact on discriminative performance than existing methods.

3.3.6 Related Work

Selective labeling/disparate censorship. Disparate censorship is a variation of selective labeling [162, 150] and outcome measurement error [99]. Selective labeling problems have been studied in clinical settings [78, 253, 42, 197, 14], social/public policy [244, 151, 163, 179, 145], and fi-

nance [29, 112], among other domains. For an extended literature review of selective labeling problems, see Appendix A.3.1.

Past work has trained ML models under disparate censorship, directly encoding untested individuals as negative [115, 130, 186, 4, 139]. Previous approaches for learning under selective labels leverage heterogeneity in human decisions to recover outcome estimates [162, 150, 44], or use domain-specific adjustments [93, 14]. We propose DCEM, a complementary approach for mitigating bias under disparate censorship without such restrictions.

Semi-supervised learning. Semi-supervised approaches do not assume labels for untested individuals. However, many causally-motivated methods diverge from the causal model of disparate censorship [181, 300, 87, 98, 97, 140, 264] via different independence/causal relationships between variables. Filtering methods [101, 171, 199, 47, 304, 306] assume specific model behavior on noisy examples (*e.g.,*, noise is learned late in training [9]) or labeling bias (randomness/class-dependence), which disparate censorship violates. We also highlight historical expectation-maximization approaches for learning with missing data [92, 91, 5], which place parametric assumptions on the data-generation process. We use neural networks to target the estimands of interest to circumvent parametric assumptions.

Other alternatives include positive-unlabeled learning approaches that assume labeling depends on covariates (*e.g.*, missing not at random) [21, 83, 90, 291]. However, these methods do not leverage correctly-labeled negatives, and naively-incorporating negative examples without causal assumptions may potentially harm model performance or bias mitigation. Other methods for noisy-label learning make assumptions incompatible with our setting, *e.g.* uniform noise within subgroups [290], almost-surely clean & noisy examples [177, 213, 276], different variable independence/directionality relationships [297], that noisy (*i.e.*, out of distribution) examples are rare [288], or other noise constraints [173, 312]. Our approach complements existing work by jointly modeling selective and biased labeling via causal assumptions tailored to a biased decision-making pipeline.

3.3.7 Conclusion

When biased human decisions affect observations of ground truth, applying standard supervised learning techniques to data exhibiting disparate censorship can amplify the harm of ML models to marginalized groups. We propose Disparate Censorship Expectation-Maximization (DCEM), a novel approach to classification, to mitigate such harm. In a simulation study and a sepsis classification task, DCEM mitigates bias and maintains competitive discriminative performance compared to baselines. Limitations of DCEM include potential slow convergence, since EM is iterative. Model evaluation under disparate censorship is also inherently difficult due to the difficulty of obtaining ground truth, motivating future work in dataset curation. Furthermore, DCEM does not

learn a full generative model with all variables. While such a model could target a wider range of estimands, it would also increase the number of terms that need to be modeled. Ultimately, DCEM is a step towards mitigating the disproportionate impacts of disparate censorship. Our work aims to raise awareness of disparate censorship and motivate the study of bias mitigation methods.

3.4 Concluding remarks on disparate censorship: equitable ML models, and beyond

Disparate censorship represents a real-world setting in which a clearly-defined mechanism of bias impacts the intrinsic behavior of an ML model. In particular, disparate censorship can impact the output distribution of ML models, propagating bias in human labeling decisions to decisions informed by ML models. In the push for models for ML-guided clinical decision support, ensuring that such decision support is equitable is central to the responsible delivery of clinical care. By developing a causal model of the relevant bias mechanism under disparate censorship, the design of our method for mitigating disparate censorship in the ML pipeline emerges naturally.

Naturally, an equitable ML model is merely one piece in reinforcing the equitable delivery of clinical care. In other words, the clinical workflow must provide opportunities for the equitable model to affect decision-making. To evaluate the model, some metric for evaluating improvement in equity with respect to external outcomes of interest, such as downstream patient outcomes, or even future testing decisions, is also needed. These choices are not a purely technical: while this thesis explores technical solutions to disparate censorship, real-world mitigation of disparate censorship using our work would need to account for such workflow effects and to design evaluation metrics for equity with respect to clinically-relevant outcomes.

As a simplifying device, this thesis adopts the philosophical convention of treating equity as a normative good: either a model is equitable, which is desirable, or a model is inequitable, which must be avoided. We further assume that the definition and importance of equity never changes throughout repeated interactions with the model: that is; patients with similar true risk for the condition of interest, which may only be partially observed due to incomplete testing, should be treated similarly by the model. Although the perspective that "equity is always desirable" is hopefully uncontroversial, alignment with principles in general may not always reduce to modifying intrinsic properties of an ML model, or increasing or decreasing a single property of interest.

In particular, some principles concern the *functionality* of a model, rather than its intrinsic properties. For example, questions regarding *which* outputs are producible require asking whether a user can steer a model towards arbitrary specific goals, not merely whether such outputs satisfy high-level properties such as fairness. Such values may arise in cases where ML models are required to satisfy heterogeneous constraints or preferences that may vary from user to user, or interaction to interaction. Thus, in the subsequent chapter, we consider steerability, or whether an ML model can "move" its outputs on-demand towards sets of properties.

CHAPTER 4

A Course Correction in Steerability Evaluation: Goal-conditioned Model Behavior at Inference Time

In this chapter, we study the steerability of conditional generative models. Steerability concerns the *functional* layer of the ML pipeline: whether a model can produce behaviors well-aligned with a user's intent. Here, we focus on the ability of LLMs to be steered towards fine-grained user intents. First, we develop a formal definition of steerability. Then, given the limited adoption of LLMs in healthcare applications to date, we shift to the task of text-rewriting, where steerability is required to satisfy multidimensional editing instructions. We evaluate the steerability of multiple recent LLMs using our framework. Our results indicate that many LLMs are not steerable, largely due to "side effects:" instances where the LLM changes textual attributes without any instruction to do so. Baselines such as prompt engineering are ineffective, while interventions such as best-of-N sampling and reinforcement learning show promise, improving steerability by mitigating side-effects despite large computational costs.

4.1 A Course Correction in Steerability Evaluation: Revealing Miscalibration and Side Effects in LLMs

4.1.1 Introduction

Large language models (LLMs) continue to advance on reasoning and instruction-following benchmarks [309, 69]. However, these gains may not yield models that reliably satisfy a wide set of specific user goals, a property called *steerability* [280, 172, 192]. Two fundamental limitations in current benchmark and metric design impede progress in steerability evaluation. First, many benchmarks sample data representative of real-world chat interactions [153, 307] or text scraped from the Internet [229]. Such data skew toward common requests and miss rarer combinations of goals. For example, making text "more formal and longer" is frequent, whereas "more formal but shorter" is rare; a benchmark that uniformly samples the goal-space can evaluate steering towards both equally. Second, many benchmarks implicitly treat performance as scalar. While potentially suitable for tasks such as instruction-following [310] or question-answering [113], single-dimensional metrics cannot measure changes in other dimensions of behavior that may arise in open-ended generation (*e.g.*, [74]). Left unmeasured, such behavioral shifts could conceal harmful behavior (e.g., sycophancy).

We design a steerability evaluation framework to address these gaps, focusing on *steering tasks* in which users aim to *transform* texts in specific ways. We map user goals and LLM outputs into a shared, multi-dimensional *goal-space* with text-to-scalar functions, from which we sample a *steerability probe* comprised of equally-weighted goals. A multi-dimensional goal-space allows us to measure behaviors such as *miscalibration*: too much/too little change along the requested direction, or *side effects*: unintended shifts in dimensions orthogonal to user goals [6]. Figure 4.1 illustrates our framework, showing a user requesting changes in reading level and text length.

Using our proposed framework, we show empirical evidence of challenges in steerability in a text rewriting task. We choose a small number of rule-based goal dimensions to disentangle the accuracy of goal measurement from steerability, and to aid in interpretation of results. First, we find that side effects remain pervasive, even in strong LLMs. As potential mitigation strategies, we find prompt engineering ineffective, while a best-of-N sampling is effective, suggesting that side effects can be mitigated, yet costly due to the need for repeated prompting. A fine-tuning approach (reinforcement learning; RL) improves steerability, but side effects remain. Our results suggests that, while some strategies show promise, they remain insufficient to solve side effects.

In summary, this work makes the following contributions:

- We introduce a framework for measuring LLM steerability with respect to multi-dimensional, multi-level user goals drawn from a goal-space (Section 4.2).

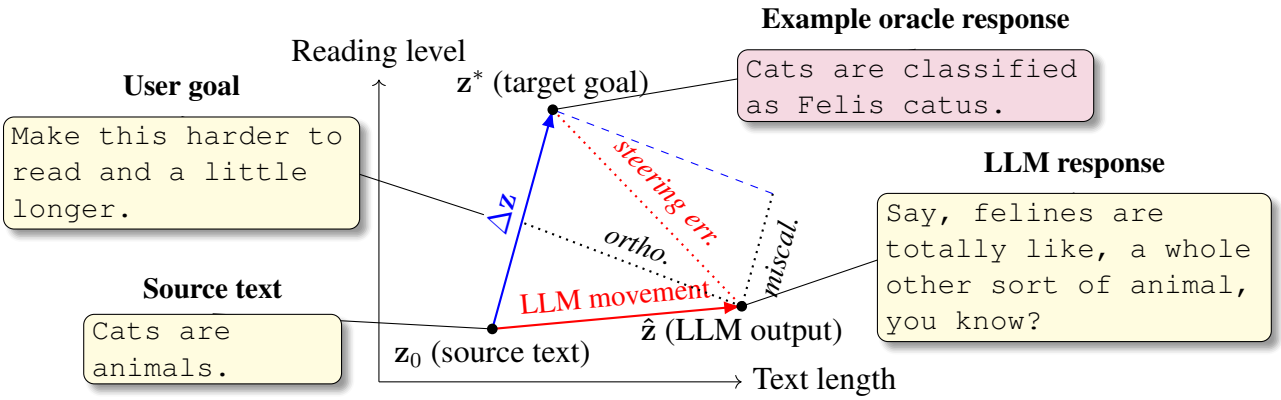


Figure 4.1: Steerability metrics in 2D goal-space (reading level & text length). A user aims to rewrite text ($\mathbf{z}_0$) according to intent $\mathbf{z}^*$, expressed via a prompt (`Make this harder to read...`). The steering error (red dotted line) is the gap between the user's intent $\mathbf{z}^*$ (blue) and the LLM's output $\hat{\mathbf{z}}$ (red). Miscalibration (miscal.) and orthogonality (ortho.) decompose steering error into components parallel and orthogonal to steering error respectively.

- We introduce a methodology for building steerability probes from existing corpora that equally weighs multi-dimensional goals (Section 4.3).
- We benchmark steerability across a variety of LLMs, finding that models frequently exhibit *side effects*, or changes to text in unrequested dimensions (Section 4.4.1).
- We evaluate inference-time steerability interventions, finding prompt engineering ineffective, while best-of-N sampling is effective but costly (Section 4.4.2).
- We evaluate RL fine-tuning and find that it improves steerability, though side effects remain (Section 4.4.3).

The strength of side effects in LLMs across models highlights a potential gap between current LLM capabilities and steerability. We hope our framework can improve evaluation of LLM alignment with diverse sets of human goals, complementing current measures of LLM capability.

4.2 A steerability measurement framework

We aim to measure how well a model follows structured, multi-dimensional user goals in a single-turn setting; e.g., text-rewriting. Here, we formalize steerability in the context of LLM evaluation (Section 4.2.1), and introduce metrics for LLM performance in the space of user goals (Section 4.2.2).

Table 4.1: Comparison of single-turn LLM evaluation strategies by scoring/loss function (ℓ) and a representative evaluation dataset ($P(\mathbf{z}^*, \mathbf{z}_0)$) in terms of LLM output $\hat{\mathbf{z}}$ and user goal $\mathbf{z}^*$.

Metric	**Scoring function (ℓ)**	**Example evaluation dataset ($P(\mathbf{z}^*, \mathbf{z}_0)$)**
Correctness (binary accuracy)	$\mathbf{1}[\hat{\mathbf{z}} = \mathbf{z}^*]$	Instruction-following, reasoning benchmarks (*e.g.*, math)
Ranking accuracy	$\mathbf{1}[R(\hat{\mathbf{z}}_i) > R(\hat{\mathbf{z}}_j)]$ (R: reward)	Pairwise or ordinal preference rankings
Scalar/1D accuracy	$z^* - \hat{z}$; $z^*, \hat{z} \in \mathbb{R}$	Questionnaire-style behavior probes
Steerability (proposed)	$\|\mathbf{z}^* - \hat{\mathbf{z}}\|_2$; $\mathbf{z}^*, \hat{\mathbf{z}} \in \mathbb{R}^n$	Steerability sampled uniformly on $(\mathbf{z}^*, \mathbf{z}_0) \in \mathcal{Z}$

4.2.1 Designing a steerability metric

We aim to evaluate the steerability of a conditional generative model f, which produces outputs $y \in \mathcal{Y}$ via sampling $y \sim f(\cdot \mid x)$ given input $x \in \mathcal{X}$. To evaluate f, one generally measures performance over some user goals $\mathbf{z}^* \sim P(\cdot)$, or *target* goals, where users verbalize intents $\mathbf{z}^*$ via $x \sim P(\cdot \mid \mathbf{z}^*)$. Such metrics consist of (i) an aggregation function over intents $\mathbf{z}^*$ and (ii) a loss function $\ell(\cdot, \mathbf{z}^*)$ that captures concordance between outputs and user goals $\mathbf{z}^*$:

$$\text{metric}(f) \triangleq \mathbb{E}_{\mathbf{z}^* \sim P(\cdot)} \mathbb{E}_{x \sim P(\cdot \mid \mathbf{z}^*)} \mathbb{E}_{y \sim f(\cdot \mid x)} \, \ell(y, \mathbf{z}^*) \tag{4.1}$$

Prior LLM evaluations choose different aggregation and loss functions, summarized in Table 4.1. Instruction following tasks often use a binary ℓ (e.g., correctness [227, 310, 109]) and implicitly assume a small set of canonical goals (e.g., instruction types). Scalar/1D metrics define a continuous ℓ (e.g., P(desired behavior) [209, 277]; concept detection "scores" [**?**]), though such functions remain 1D. Ranking accuracy-based losses [207, 228] emphasize *relative* rather than absolute response quality. Some evaluations rely on chat log data or web scraping [153, 307, 229], or are purpose-built to test specific capabilities [310, 54, 113], which may not be representative of potential users and goals.

These shortcomings may be especially pronounced in *steering tasks* where users aim to transform model outputs along multi-dimensional, multi-level dimensions, such as text-rewriting. In particular, steering tasks may contain a wider range of potential user goals than typically seen in benchmarks. Such tasks may also expose miscalibration, as coarse metrics such as binary accuracy/rankings can lead models to score distinct responses identically, flattening different types of deviations from the user's intent. In addition, since steering tasks may include requests for multi-dimensional changes to text, single dimensional metrics may hide unintended side effects in LLM responses.

We contribute a steerability metric that addresses these limitations by (i) aggregating over a *uniform* distribution of goals, allowing us to better identify poor coverage, and (ii) using a loss function ℓ that measures absolute distance between target goals and model outputs in multiple dimensions. Specifically, let $\mathbf{z}_0$ be a source that to be transformed, and let $\hat{\mathbf{z}}$ be the intent satisfied by the LLM output. Recall that $\mathbf{z}^*$ is the user's intent. Treating $\mathbf{z}^*$, $\hat{\mathbf{z}}$ and $\mathbf{z}_0$ as elements of a shared metric space $\mathcal{Z}$ (*e.g.*, Fig. 4.1), we write:

$$\text{steerability}(f) \triangleq \mathbb{E}_{\mathbf{z}_0,\mathbf{z}^*\sim\mathcal{U}}\, \mathbb{E}_{\hat{\mathbf{z}}\sim f(\cdot|\mathbf{z}_0,\mathbf{z}^*)}[\|\hat{\mathbf{z}} - \mathbf{z}^*\|_2] \tag{4.2}$$

where $\mathcal{U}$ is a uniform distribution over $\mathbf{z}_0$ and $\mathbf{z}^*$.

4.2.2 Measuring steerability in practice

Our steerability metric (Eq. 4.2) puts $\mathbf{z}^*$, $\mathbf{z}_0$, and $\hat{\mathbf{z}}$ in a shared space $\mathcal{Z}$. To define $\mathcal{Z}$, we observe that user goals $\mathbf{z}^*$ for steering tasks often decompose along interpretable dimensions (e.g., "Make this harder to read and a little longer"). Thus, we define $\mathcal{Z}$ to be a set of *dimensions* representing attributes of text (e.g., reading level and length). Formally, define *goal space* $\mathcal{Z} = [0,1]^{|\mathcal{G}|}$, and functions $g_i : \mathcal{Y} \to [0,1]$ for $i \in 1,\ldots,|\mathcal{G}|$ that translate model outputs $y \sim f(\cdot \mid x)$ into goal space, where g_i can be based on existing measures of text features (e.g., Flesch-Kincaid grade level [147], word count). The joint output of all g_i is the *goal-space mapping* of y; *i.e.*, a vector representation of y.

As an example, consider measuring steerability in text-rewriting (Figure 4.1). A user aims to rewrite a *source* (e.g., "Cats are animals") mapping to $\mathbf{z}_0$ in goal-space. Suppose that the user wants a harder-to-read, slightly longer text, which maps to $\mathbf{z}^*$, expressed via a prompt (e.g., "Make this harder to read and a little longer"). We assume $\mathbf{z}^*$ is *feasible*; *i.e.*, that it is possible to make the source harder to read and slightly longer. The LLM produces an output (e.g., "Say, felines are totally like...") satisfying intent $\hat{\mathbf{z}}$, which may not match the user's intent $\mathbf{z}^*$. We quantify the mismatch via *steering error*; *i.e.*, the Euclidean distance between $\mathbf{z}^*$ and $\hat{\mathbf{z}}$. Finally, to ensure *coverage*, we average over a uniformly sample of $\mathbf{z}_0$ and $\mathbf{z}^*$, from which we obtain Eq. 4.2.

However, steering error ($\|\mathbf{z}^* - \hat{\mathbf{z}}\|_2$) does not distinguish between miscalibration, or errors in magnitude, versus side effects, or errors due to unintended changes. Decomposing steering error along direction of the user's intent (*miscalibration*) and the orthogonal error (*orthogonality*), we write:

$$\|\mathbf{z}^* - \hat{\mathbf{z}}\|_2 = \|(\mathbf{z}^* - \mathbf{z}_0) - (\hat{\mathbf{z}} - \mathbf{z}_0)\|_2. \tag{4.3}$$

Projecting Eq. 4.3 onto the desired movement vector $(\mathbf{z}^* - \mathbf{z}_0)$, an orthogonal decomposition yields

the components

$$\text{miscal}(\mathbf{z}^*, \hat{\mathbf{z}} \mid \mathbf{z}_0) = \text{sproj}_{\mathbf{z}^*-\mathbf{z}_0}(\mathbf{z}^* - \hat{\mathbf{z}})/\|\mathbf{z}^* - \mathbf{z}_0\|_2 \tag{4.4}$$

$$\text{ortho}(\mathbf{z}^*, \hat{\mathbf{z}} \mid \mathbf{z}_0) = \text{sproj}^{\perp}_{\mathbf{z}^*-\mathbf{z}_0}(\mathbf{z}^* - \hat{\mathbf{z}})/\|\hat{\mathbf{z}} - \mathbf{z}_0\|_2 \tag{4.5}$$

where sproj and $\text{sproj}^{\perp}$ represent the magnitude of the scalar projection of the steering error onto the desired movement vector and that of the orthogonal complement, respectively (see Figure 4.1). As desired, Eq. 4.4 measures *miscalibration* by capturing over/under-shooting in the direction of the intent, while Eq. 4.5 (orthogonality) captures side effects by quantifying unintended deviation in the orthogonal direction. We normalize both metrics to control for severity in steering error so that small errors in requests for small changes are more serious than such errors in requests for large changes. All three metrics are minimized at zero.

4.3 Experimental setup

Steerability probes are benchmarks designed to measure steerability for a steering task. We describe how we construct an example steerability probe for text-rewriting (Section 4.3.1), candidate steerability interventions evaluated (Section 4.3.2), and our inference setup (Section 4.3.3).

4.3.1 Steerability probe construction

We measure steerability in text-rewriting, a common task likely well-represented in LLM training data. Our probe has two components: (i) goal dimensions defining a goal-space $\mathcal{Z}$ and (ii) a dataset of goals $(\mathbf{z}_0, \mathbf{z}^*) \sim \mathcal{Z}$.

Design principles. For goal dimensions, we choose rules-based evaluators from textual analysis, such that these dimensions are likely valid and fully realizable in goal-space. We can also verify that such dimensions are *independent*, aiding interpretation of LLM behavior such as side effects. For source texts, we combine datasets containing diverse writing styles, from which a uniform subsample can be constructed. This setup can be extended to accommodate more goal dimensions, source texts, or tasks. See Appendix B.1.1 for full probe construction details.

Goal-space. We select reading difficulty (Flesch-Kincaid grade [147]), formality (Heylighen-Dewaele F-score [117]), textual lexical diversity [129], and text length (word count). Though these dimensions may be correlated in training data, each is independently manipulable in theory (e.g., syllables per word & sentence length affect Flesch-Kincaid; whereas Heylighen-Dewaele measures

the part-of-speech distribution). Requests mentioning these dimensions appear in real-world chats (e.g. WildChat/LMSys [307], Appendix B.1.4.5). Metric descriptions are in the Appendix. For RL fine-tuning, we focus on 2D goal-space (reading difficulty, formality) to isolate challenges in steerability in a simple setting where goal dimensions are conceptually distinct but likely correlated in real-world text. As a secondary validity check, we verify that LLM-as-judge can detect changes in all chosen goal dimensions (see Appendix B.1.1.2).

Although these metrics are not the only way to operationalize the goal dimensions under investigation, these metrics are established in textual analysis. As a secondary validity check, we verify that LLM-as-judge can detect changes in all goal dimensions (Appendix B.1.2.3).

Source texts and goals. We sample seed texts from news articles (CNN/DailyMail [249]), social media (RedditTIFU [146]), English novels (BookSum [157]), and movie synopses (SummScreenFD [251]), to cover a wide stylistic range (total $N = 8,303$). We compute goal-space mappings for seed texts and min-max scale the empirical middle 95% of each goal dimension to $[0, 1]$, clipping values outside that range, such that goal dimensions are on comparable scales. We then resample $\mathbf{z}_0$ to be uniform over over goal space $\mathcal{Z}$ via reweighting. For each $\mathbf{z}_0$, we choose three *active* goal dimensions at random, and sample $\mathbf{z}^*$ within ± 0.1 to 0.7 of the original value, copying components of $\mathbf{z}_0$ to $\mathbf{z}^*$ for inactive dimensions. Our main probe consists of 64 source texts with 32 goals each ($N = 2,048$). See Appendix B.1.1.1 for pre-processing details. All reported results are statistically significant at level $\alpha = 0.05$ based on a paired, two-sided Wilcoxon rank-signed test, with other tests used as specified.

For RL fine-tuning, our training probe consists of 384 source texts with 16 goals each ($N = 3,072$). We select *one* active goal dimension and report metrics post-RL on 64 held-out source texts with 16 goals each ($N = 1,024$) in 2D goal-space with one active goal dimension unless specified.

Default prompt. To turn $(\mathbf{z}_0, \mathbf{z}^*)$ into prompts, we use a template-based prompt that names active goal dimensions with modifiers "slightly" for changes < 0.2, and "much" when changes are > 0.5, and no modifier otherwise (e.g., "make this [slightly/much] [more/less] formal;" see also Appendix B.1.2.2). To avoid penalizing prompt ambiguity rather than steerability failures, we discretize $\mathbf{z}^*$ and $\hat{\mathbf{z}}$ using the same bins implied by the prompt modifiers (cut points at 0, ± 0.2 and ± 0.5) when reporting steerability metrics.

4.3.2 Candidate steerability interventions

We evaluate common single-turn techniques for influencing model behavior. We choose a set of methods applicable to multi-dimensional, multi-level intents, namely, prompt engineering, best-of-N sampling, and RL fine-tuning.

Prompt engineering. Prompt engineering is the design of a strategy for verbalizing intent $\mathbf{z}^*$, which may span direct instructions (e.g., Figure 4.1), chain-of-thought [293], or negative prompting (e.g., "don't change anything else") [243]. We extend the default prompt by testing the inclusion of negative prompts and specific instructions (e.g., "increase formality by changing X"), a chain-of-thought style directive (e.g., "explain proposed edits"), and an underspecified prompt as a naive upper bound on steering error. While non-exhaustive, this set reflects common strategies proposed in prior work applicable to text rewriting. See Appendix B.1.2.2 for examples.

Best-of-N sampling. Best-of-N selects the response with the lowest steering error out of N attempts, assessing whether models are even capable of producing responses with low steering error. To encourage diverse but fluent samples, we use min-p sampling ($p = 0.2$) with temperature 1 and a 0.1 frequency penalty [200].

RL fine-tuning. RL fine-tuning optimizes model parameters via online RL, using steering error as the negative reward. Since sampling directly from uniform $\mathcal{U}$ may be infeasible, we reweight training examples from a dataset $\mathcal{D}$ by estimating the density ratio $\mathcal{U}/\mathcal{D}$ via classifier-based methods [26]:

$$\min_f \mathbb{E}_{(\mathbf{z}_0,\mathbf{z}^*)\sim\mathcal{D}} \mathbb{E}_{\hat{\mathbf{z}}\sim f(\cdot|\mathbf{z}_0,\mathbf{z}^*)}[\hat{w}(\mathbf{z}_0,\mathbf{z}^*)\cdot\|\mathbf{z}^*-\hat{\mathbf{z}}\|_2^2]. \tag{4.6}$$

To optimize Eq. 4.6, we use a policy gradient method based on leave-one-out proximal policy optimization (LOOP) [45]. We fine-tune a Llama3.1-8B model via rank-stabilized LoRA [136]. We generate rollouts using the same decoding parameters as best-of-N sampling. We discuss modifications to LOOP in Appendix B.1.3.2, and hyperparameters in Appendix B.1.3.3.

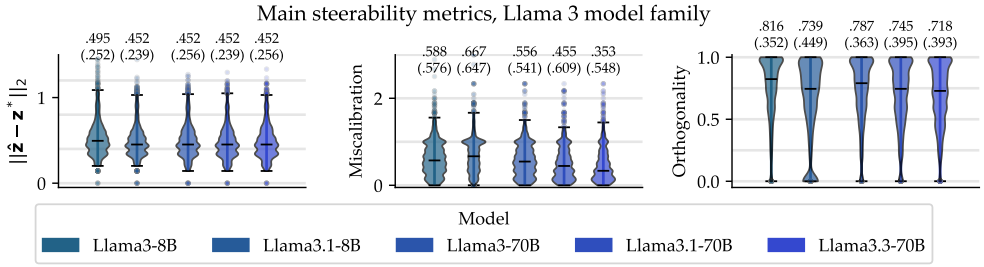


Figure 4.2: Median (IQR) of steering error (left), miscalibration (middle), and orthogonality, Llama3 family. Steering error does not improve with model size (left), but miscalibration does (middle). Orthogonality drops slightly (right), but remains skewed.

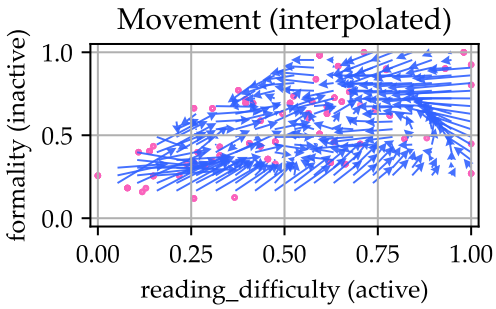


Figure 4.3: Vector flow of goal-space movement (blue), Llama3.3-70B, in requests to change reading difficulty but not formality. Horizontal movement is desired, but not vertical movement. Source texts in red.

4.3.3 LLM inference setup

Models. We evaluate GPT (3.5 turbo, 4 turbo, 4o, 4.1 [3, 205, 123]), Llama3 (Llama3 to 3.3, 8B/70B [72]), Deepseek-R1 variants (8B/70B, distilled [270]), and Qwen3 (4B/32B/30B-A3B) [271], and o1-/o3-mini [127, 206]. LLM inference is performed using the OpenAI API (GPT) or vLLM (all others) [159], with greedy sampling and a context length of 32,000 tokens unless specified.

Output post-processing. To ensure metrics are computed over valid rewrites, we post-process responses to remove boilerplate text (e.g., "Sure, here's...") and reasoning tokens (e.g., `<think>` blocks). We also filter refusals, degenerate behavior (e.g., repetitive looping), or rewrites unrelated to the source using LLM-as-judge and manual review of responses flagged by the LLM (see Appendix B.1.2.3; B.1.4.4).[1]

4.4 Empirical results

We evaluate steerability in text-rewriting using the proposed metrics. Our results suggest that current LLMs are not steerable, which we largely attribute to side effects. Further analysis suggests goal dimensions may be spuriously entangled (Section 4.4.1). As candidate interventions, we try prompt engineering, which is ineffective, and best-of-N sampling, which requires extensive sampling (Section 4.4.2). We then try RL fine-tuning in 2D goal-space, which rivals best-of-128 and disentangles goals, but side effects remain (Section 4.4.3).

[1]Due to filtering, metrics are reported on slightly different response distributions. The effect is negligible: in our main results, rejected responses comprise $\leq$6 ($\approx$0.29%) of outputs in any probe.

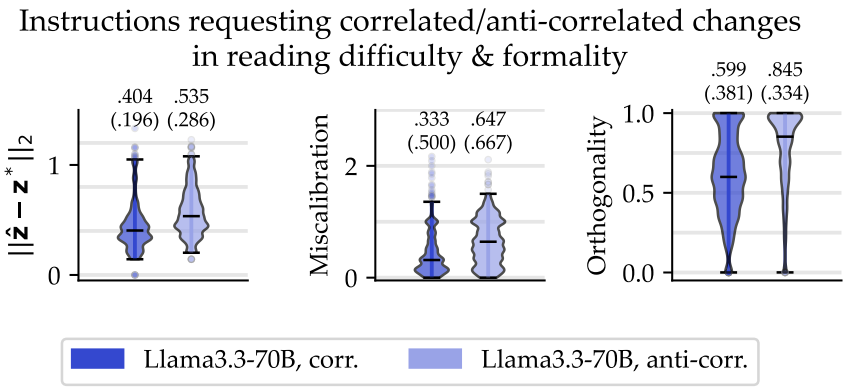


Figure 4.4: Median and IQR steerability, Llama3.3-70B, in correlated (darker) vs. anti-correlated (lighter) requests for change in reading difficulty and formality. Llama3.3-70B struggles more with anti-correlated changes.

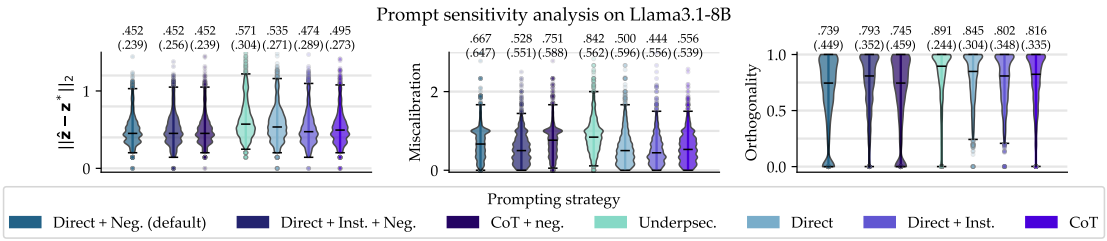


Figure 4.5: Median and IQR of steering error (left), miscalibration (middle), and orthogonality (right) of Llama3.1-8B across prompting strategies. More detailed prompts and removal of the negative prompt marginally improve miscalibration over the default. However, side-effects remain severe.

4.4.1 Large language models are not steerable

Even strong LLMs induce side effects. Neither larger nor newer models meaningfully improve steering error.[2] Median steering error remains high, 0.452 for the largest model (Llama-3.3; Figure 4.2, left), far from ideal despite outperforming a random baseline (0.770; sampling random goal levels in each dimension). Miscalibration improves (Figure 4.2, center) with model size (e.g., Llama3.1-8B vs. 70B: 0.667 $\rightarrow$ 0.455). Some residual miscalibration is expected, since the model may not be calibrated to the magnitude of "slightly/much" in our prompts.

Median orthogonality remains high and skewed towards 1 even as model size increases (Figure 4.2, right) with Llama3.3-70B performing best with an orthogonality of 0.718. While several pairwise differences are statistically significant, models remain in a high-orthogonality regime on average. Similar trends hold in GPT, Deepseek, Qwen3, and o1/o3 models, where larger/newer

[2]Some pairwise tests yield statistical significance, but effect sizes are small.

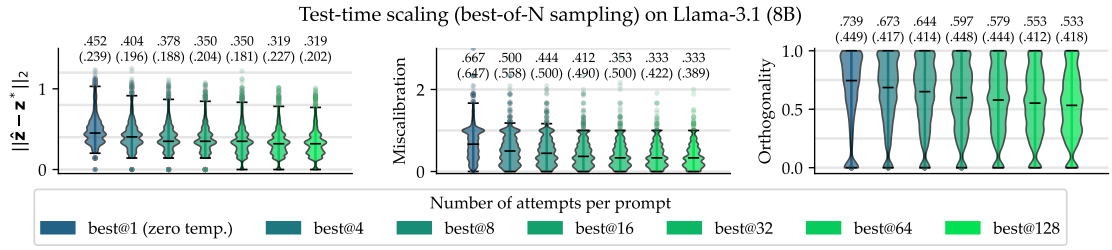


Figure 4.6: Median and IQR for best-of-$\{4, 8, \ldots, 128\}$ approaches on Llama3.1-8B, with a direct + negative prompt. Increasing N consistently improves steerability, but improvements scale slowly.

models reduce miscalibration but have little effect on orthogonality (Appendix B.1.4.1). Note that, even as miscalibration and orthogonality decrease, median steering error may not due to normalization (Eq. 4.4-4.5; errors are penalized in proportion to the requested/observed movement). To further study side effects, we analyze a 2D goal subspace.

Goal dimensions may be entangled. We investigate side effects in a 2D (reading difficulty, formality) subspace using a vector flow diagram of goal-space movement (Figure 4.3, Llama3.3-70B, blue vectors). We include instructions requesting changes to reading difficulty (x-axis) but not formality (y-axis), such that vertical movement is a side effect. Figure 4.3 shows a "current" from the lower left (informal & easy to read) to the top right, suggesting that, when asked to increase reading difficulty without direction on formality, LLMs still increase formality.

Appendix B.1.4.2 shows additional movement vectors and flows. We also conduct a preliminary study of coupling between goal dimensions, which suggests that the entanglement is LLM-induced.

While harder-to-read texts are often more formal, they need not be under our chosen measurement functions (Flesch-Kincaid grade, reading difficulty; Heylighen-Dewaele score, formality). LLM behavior appears to reflect this correlation: when stratifying steerability probe results based on whether the prompt requested correlated (e.g., make it harder to read and more formal) vs. anti-correlated changes to reading difficulty and formality (e.g., make it harder to read and *less* formal), Llama3.3-70B is less steerable on anti-correlated requests compared to correlated requests (steering error, 0.535 vs. 0.404; Figure 4.4, diff.: 0.131, Mann-Whitney $U = 77944.5$), with similar results in other model families (GPT, Deepseek, Qwen3; see Appendix). Thus, side effects may harm steerability in requests running contrary to similar correlations.

On coverage. While our probe is designed to target a uniform distribution of goals, results are similar whether or not we sample source texts uniformly in goal-space (Appendix B.1.4.1). Thus, steerability failures are unlikely to be concentrated in overrepresented goals in our evaluation.

	Steering error	Miscalibration	Orthogonality
Base model (pre-RL)	0.300 (0.150)	0.986 (0.464)	0.147 (0.328)
Best@128 (pre-RL)	0.210 (0.168)	0.683 (0.539)	0.121 (0.283)
Miscalibration-only reward	0.210 (0.138)	0.542 (0.429)	0.366 (0.395)
Orthogonality-only reward	0.386 (0.248)	1.463 (1.004)	**0.025** (0.134)
Full steering error	**0.119** (0.135)	**0.294** (0.391)	0.160 (0.292)

Table 4.2: Main results for RL, with an ablation study of the reward model. Mean (std. dev.) of steerability metrics across evaluation probe ($N = 1,024$: 64 held-out source texts; 16 goals each).

Request type	Pre-RL orth.	Pre-RL, best@128 orth.	Post-RL orth.
Corr. requests	0.216 (0.279)	0.238 (0.253)	0.322 (0.283)
Anti-corr. requests	0.330 (0.399)	0.341 (0.320)	0.317 (0.264)
Mean Gap (absolute)	0.114	0.103	0.005

Table 4.3: Mean (std. dev.) of (from left to right) orthogonality for pre- vs. post-RL model on correlated (top; e.g., increase both dimensions) vs. anti-correlated requests (middle; e.g., change dimensions in opposite directions). RL shrinks the gap in side-effects (bottom) between correlated and anti-correlated requests, despite only supervised via 1D instructions.

Takeaway #1: side effects impede steerability. Despite progress in LLM reasoning and model capacity, LLMs continue to exhibit side effects. Entanglement between goal dimensions contributes to side effects, limiting steerability for intents that contradict correlations between goal dimensions.

4.4.2 Inference-time steering is costly

We now study whether inference-time strategies can improve steerability. First, we evaluate whether prompt engineering can elicit responses that satisfy user goals. Second, we leverage best-of-N sampling to test whether such responses are in the support of the model's output distribution.

Model	Sentence BLEU
Pre-RL	0.864 (0.239)
Pre-RL, best@128	0.761 (0.245)
Post-RL	0.529 (0.239)

Table 4.4: Mean (std. dev.) sentence-level BLEU (original vs. rewrite) by dataset, pre- & post-RL.

Prompt engineering does not solve side effects. More detailed prompting strategies compared to the default (e.g., chain-of-thought style or adding instructions) tend to improve miscalibration, as does removing the negative prompt (median: 0.667 $\rightarrow$ 0.444, no negative prompt + added instructions; Figure 4.5, middle). Yet orthogonality remains skewed towards 1 despite improvements under some strategies (e.g., direct + negative prompts; Figure 4.5, right). Thus, mitigating side effects with prompt engineering alone may be challenging. Results for all strategies are in Appendix B.1.4.1.

Best-of-N sampling is a costly solution. Since side effects remain severe across prompting strategies, we investigate whether responses that reduce side effects exist in the model's sampling distribution via best-of-N sampling. Best-of-4 with Llama3.1-8B lowers steering error (Figure 4.6, left), outperforming best-of-1 across all prompting strategies and models evaluated (GPT-4.1 vs. Llama3.1-8B: 0.429 $\rightarrow$ to 0.404, see Appendix B.1.4.1). Median orthogonality at best-of-4 also outperforms the top best-of-1 model (GPT-4.1 vs. Llama3.3-70B: 0.718 $\rightarrow$ 0.673, see Appendix B.1.4.1). This improvement with N suggests that responses better-aligned with goals lie within the model's support but are rare in the model's sampling distribution. Best-of-N also scales poorly, lowering median steering error by 0.031 at most when doubling N (Figure 4.6, left).

Takeaway #2: Inference-time steerability is possible but inefficient. We find that prompt engineering alone may not be powerful enough to surface responses with low steering error. While best-of-N sampling demonstrates the existence of such responses, they remain rare under the base model's output distribution. Our results motivate fine-tuning to increase the likelihood of low steering-error responses.

4.4.3 RL yields progress towards steerable models

Gains under best-of-N sampling suggest that low steering error responses exist but are rare under an LLM's sampling distribution. We hypothesize RL can shift the output distribution towards such generations. Indeed, RL improves steerability, adopting different rewriting strategies compared to the base model, but does not eliminate side effects.

The post-RL model rivals best-of-128 sampling. In a 2D goal-space (reading difficulty & formality), RL improves steerability in Llama3.1-8B. We report mean and standard deviation to capture improvements in the tails (Table 4.2). Post-RL steerability rivals best-of-128 sampling in steering error (pre-RL best@128 vs. post-RL mean: 0.210 $\rightarrow$ 0.119), though orthogonality lags the base model (pre-RL vs. post-RL mean: 0.147 $\rightarrow$ 0.121). Furthermore, optimizing only miscalibration or orthogonality worsens the other (e.g., RL w/ steering error: 0.294; orthogonality-only: 1.463), which we conjecture may be due to underspecification: flat-reward regions could

worsen overfitting (e.g., all formality levels are equal-reward when optimizing miscalibration in reading difficulty only).

RL shifts the model's rewriting strategies. To analyze whether post-RL steerability improvements are meaningful, we examine changes in generation patterns. First, RL mitigates copy-pasting behavior. Before fine-tuning, the base model copy-pastes the source text in 135 of 1,024 (13.2%) prompts evaluated, a trivial method to minimize orthogonality. Post-RL, the copy-paste behavior vanishes. BLEU [210] between rewrites and source texts also drops (Table 4.4; pre-RL vs. post-RL: $0.864 \rightarrow 0.529$), suggesting that the post-RL model adopts a less conservative editing strategy to satisfy user goals. Pre- vs. post-RL flow diagrams (Appendix B.1.4.2) support our analysis. Second, RL generalizes to unseen instructions. Despite training with 1D instructions, the post-RL model better handles 2D anti-correlated instructions. We report mean and standard deviation to capture improvements in the tails. The difference in mean orthogonality between correlated vs. anti-correlated requests largely vanishes, dropping from 0.114 pre-RL to 0.005 post-RL (Table 4.3, right), suggesting improved independence in controlling each goal dimension. We show violin plots summarizing other metrics in Figure **??**. Analysis of an anti-correlated rewrite (Appendix B.1.4.3) further illustrates this behavior.

Takeaway #3: RL yields partial progress towards steerability. In a 2D goal-space, we improve the steerability of Llama3.1-8B. These improvements reflect meaningful changes in the model's rewriting strategies, such as reducing copy-paste behavior (lower BLEU score post-RL) and improved disentanglement of goal dimensions (lower orthogonality post-RL in anti-correlated requests). Nonetheless, orthogonality can still be improved, highlighting the need for further work to eliminate side effects.

4.5 Discussion & Conclusion

We propose a framework for measuring steerability: whether a model can reliably follow diverse, multi-dimensional goals. Existing LLM evaluations directly leverage data from real-world interactions or Internet text, which may not be representative, or use single-dimensional metrics, which do not capture side effects in open-ended generation. Our steerability probe design mitigates these gaps by uniformly sampling goals and measuring multiple dimensions of text. Empirically, LLMs struggle with steerability due to side effects. Inference-time interventions such as prompt engineering and best-of-N sampling offer minor or costly gains. However, RL fine-tuning shows promise as a partial solution. Our work suggests that steerability may be a fundamental challenge

for LLM alignment, requiring shifts in model behavior beyond inference-time tweaks. We hope that our framework provides a foundation for measuring LLM alignment with diverse sets of user goals.

Limitations. We focus on steering along verifiable text attributes, leaving goals such as style, to future work. Even as we anchor to such goals with objective measurements, there remains inherent subjectivity not only in one defines "goals," but also in how one measures them. We also only evaluate LLMs in single-turn settings. However, our framework is easily extended to multi-turn settings or generative models beyond text (e.g., multimodal LMs). Our study of interventions is non-exhaustive: we do not vary prompt formatting and apply RL to only an 8B model in 2D goal-space. Larger models may have higher post-RL upside, but optimizing steerability in higher dimensional goal-space could introduce new challenges. Ultimately, our framework is a principled foundation for evaluating LLM steerability, that we hope complements current evaluations of LLM capabilities and improves alignment with diverse human goals.

CHAPTER 5

Gaming and Strategic Adaptation: Facilitating Oversight over Deployment-time Interactions

In this chapter, we study the strategic manipulation of ML models. Strategic manipulation concerns the systemic layer of the ML pipeline. In contrast to steerability, which relates to whether the functionality is sufficiently *broad* to satisfy user goals, systemic alignment concerns model-user interactions that tend to emerge under the particular incentives, regulations, and idiosyncrascies of a system, and whether such interactions remain within certain acceptable bounds. We study these questions in the context of Medicare Advantage, a U.S. public health insurance program, as a case study. Concretely, in Medicare Advantage, the government operates a risk-adjusted reimbursement model based on reported diagnoses to compute capitation payments to insurance companies, creating an incentive to overreport diagnoses. We frame this setting as a multi-agent strategic adaptation problem, in which multiple insurers (*i.e.*, "agents") report diagnoses according to some utility maximization problem balancing the benefits of overreporting (increased payment) with the costs of overreporting (fines and prosecution), which may vary across agents based on their risk tolerance for overreporting (the *gaming deterrence* parameter.) While we show that exactly estimating gaming deterrence is not possible, a ranking of agents by gaming deterrence can be constructed through pairwise comparisons between agents by solving a causal effect estimation problem. We apply our approach to create a national ranking of Medicare Advantage organizations (MAOs) using a weighted Bradley-Terry model based on head-to-head comparisons between MA contracts. We find that our ranking correlates with gaps in the rate of relevant treatments: across multiple conditions, the probability of a beneficiary receiving a treatment relevant to their diagnosis decreases in higher-ranked MA contracts (more aggressive).

5.1 Who's Gaming the System? A Causally-Motivated Approach for Detecting Strategic Adaptation

5.1.1 Introduction

Machine learning (ML) models often guide decisions that impact individuals or entities. Attributes describing an individual or entity are often inputs to such models. In response, such entities may modify their attributes to obtain a more desirable outcome. But changing one's attributes may be costly due to the difficulty of generating supporting evidence, or penalties for fraud. This behavior is called *gaming* or *strategic adaptation* [103]. Strategic adaptation frames gaming as "utility maximization:" *agents* change their attributes to maximize a payout, but incur a cost for modifying attributes.

As an illustrative example, we turn to the health insurance industry. In the United States (U.S.), contracted health insurance companies report their enrollees' diagnoses to the government, which calculates a payout based on reported diagnoses via a publicly available model [224]. The *payout* is intended to support care of the enrollee in relation to the diagnosis. Companies may attempt to maximize payouts by reporting extraneous diagnoses, an illegal practice known as "upcoding" [89]. Despite increasing awareness of upcoding [89, 258, 85, 71], upcoding costs U.S. taxpayers over $12B U.S. dollars annually [50], even with substantial investment in audits ($100.7M U.S. dollars, 2023 [81]) and payout changes to adjust for gaming [155, 82]. Since audits may not scale and often overlook fraud [241, 257], tools for flagging gaming-prone agents could help target audits. Beyond health insurance, gaming emerges in responses to credit-scoring algorithms [52, 15] and driver responses to rider allocation algorithms in ride-sharing apps [168].

In this work, we study how one can identify agents with the highest propensity to strategically manipulate their inputs given a dataset of agents and their observed model inputs. A supervised approach is infeasible since fraud/gaming labels are unavailable in our setting. A common paradigm for fraud/gaming detection is unsupervised anomaly detection. However, gamed attributes may not be outlier-like. The perspective of gaming as utility-maximizing behavior in strategic classification provides an alternative to existing fraud/gaming detection methods. Many works in strategic classification (*e.g.*, [103]) assume known utility functions and identical feature manipulation costs across agents, which assists in identifying an agent's "optimal" gaming behavior. However, such assumptions may not always apply. For example, in U.S. Medicare, due to the rarity of penalties for upcoding, it is unclear how to quantify the *cost* of fraud. Furthermore, due to the large number of companies contracted with U.S. Medicare, with distinct incentives (*e.g.*, for-profit vs. non-profit groups) and disjoint populations of patients, there may be heterogeneity in feature manipulation costs.

To bridge this gap, we propose a novel framework for modeling agent utilities by introducing a **gaming deterrence parameter**, which scales the perceived cost (to the agent) of gaming. First, we show that directly estimating the gaming parameter is not possible: the best we can do is a lower bound on the gaming deterrence parameter. However, by re-casting gaming detection as a causal effect estimation problem, where each agent represents a "treatment," we prove that a *ranking* of agents based on their gaming deterrence parameter is recoverable. Thus, we propose a causally-motivated ranking algorithm that produces a ranking of agents. Practically, agents most likely to game under our ranking could be flagged for further monitoring/auditing. While our framework is inspired by health insurance fraud, it applies more broadly to instances of gaming where multiple agents are gaming an ML-guided decision.

We evaluate the performance of causal effect estimators for gaming detection in a synthetic dataset. Empirically, causal approaches rank the worst offenders higher than existing non-causal approaches that screen based on payouts/randomly, as well as anomaly detection methods. We then verify in a real-world U.S. Medicare claims dataset that causal effect estimation yields rankings correlated with the prevalence of for-profit healthcare providers, a suspected driver of gaming [258, 259].

In summary: we **1)** extend strategic classification to model differences in gaming behavior across agents (Section 5.1.3), **2)** prove that point-identifying an agents' gaming parameter is impossible without strong assumptions (Section 5.1.4), **3)** show that, by recasting gaming detection as causal effect estimation, one can recover a ranking of agents based on their gaming deterrence parameters (Section 5.1.4), **4)** demonstrate empirically that our framework identifies the worst offenders with fewer audits than baselines (Section 5.1.5), and **5)** show in a real-data case study that our approach yields rankings correlated with suspected drivers of gaming (Section 5.1.5). Code to replicate our experiments will be made publicly available at https://github.com/MLD3/gaming_detection.

5.1.2 Related works

Machine learning-based anomaly detection. Fraud detection is often framed as an unsupervised anomaly/outlier detection problem [232, 60, 176, 105, 96, 174, 298]. Such approaches assume gamed model inputs are outliers in some distribution. However, this assumption may be incorrect if gaming is common in the data, or if gaming results in only small changes to observed agent attributes. In borrowing from strategic adaptation, we frame gaming as utility maximization, rather than making distributional assumptions about gamed agent attributes.

Strategic classification & gaming in machine learning. A large body of work in strategic classification aims to design incentives to mitigate gaming/strategic behavior by agents. However,

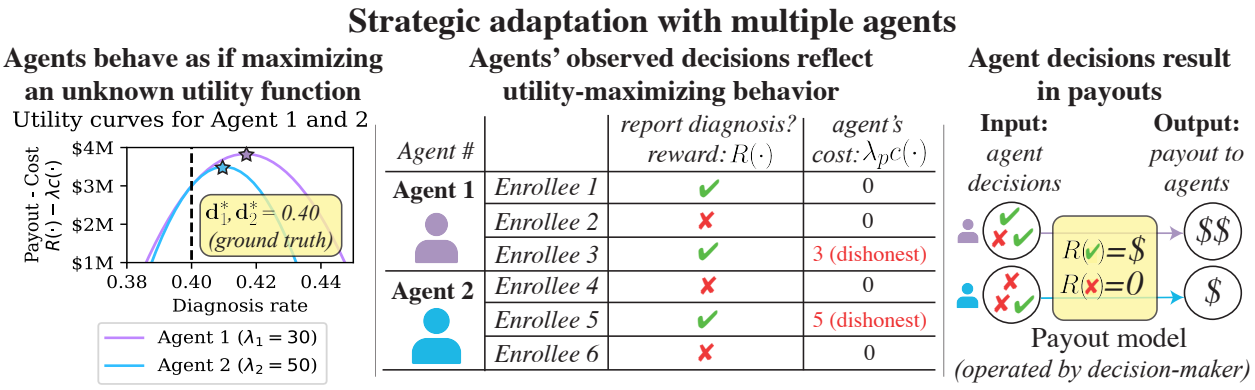


Figure 5.1: **Left:** Two agents with gaming deterrence parameters $\lambda_1 = 30$ (purple) and $\lambda_2 = 50$ (blue) maximize utility (reward R - cost c) with respect to diagnosis rate. Gaming costs *increase* in $\lambda_{(\cdot)}$, and lower an agent's optimal diagnosis rate (stars). **Center:** Agents' observed decisions reflect utility-maximizing behavior. **Right:** A decision-maker computes a payout based on agent decisions.

such works assume that feature manipulations costs are known/can be estimated across agents, or are identical [103, 48, 170, 193, 303, 165, 166, 28, 266]. In contrast, in our setting, feature manipulation costs are unknown and may differ across agents, but exhibit some shared structure that facilitates comparisons across agents. Shao et al. [254] assumes that agent gaming capacities may differ by placing bounds on manipulation, which may be unrealistic. Closest to our work is that of Dong et al. [70], which also assumes unknown and differing agent costs, but does not leverage similarities in gaming across agents. Our work supplements the strategic classification literature by studying a related yet fundamentally distinct problem: rather than designing incentives to mitigate strategic behavior, which is infeasible under unknown manipulation costs, we leverage differences in costs across agents to identify agents more likely to game.

5.1.3 Background & Problem Setup

We review strategic adaptation and extend it to model differences in gaming across agents.

What is strategic adaptation? Our work builds upon strategic classification [103]. Consider a pre-existing model $f : \mathcal{D} \mapsto \mathbb{R}$ that maps agent attributes $d \in \mathcal{D}$ to a payout. An *agent* may leverage its knowledge of f to change their attribute(s) d according to some function Δ:

$$\Delta(d) \triangleq \arg\max_{\tilde{d} \in \mathcal{D}} R(\tilde{d}; f) - g(\tilde{d}, d) \tag{5.1}$$

where $R : \mathcal{D} \to \mathbb{R}$ is the *payout* of changing d to $\tilde{d}$, and $g : \mathcal{D} \times \mathcal{D} \to \mathbb{R}_+$ is the *cost* of manipulating d. When $\mathcal{D} \subseteq \mathbb{R}$, g is often assumed to be "separable;" *i.e.*, for some function c, $g(\tilde{d}, d) = c(\tilde{d} - d)$. $\Delta(\cdot)$ describes how an agent manipulates d to obtain a higher payout from f. This behavior is

called *strategic adaptation* or *gaming*. For simplicity, we assume $R = f$; *i.e.*, the model f directly determines the payout.

Modeling agent variation in gaming. To extend strategic adaptation to multiple agents, we add a non-negative *gaming deterrence parameter* $\lambda_p \in \mathbb{R}^+$ to Eq. 5.1. Consider an observational dataset $\mathcal{D}_p \triangleq \{(\mathbf{x}_i, d_i)\}_{i=1}^{M_p}$ of an agent p's decisions $d_i \in \{0, 1\}$ given some information $\mathbf{x}_i \in \mathcal{X}$. For simplicity, we assume d_i is binary, though the proposed framework generalizes to non-binary decisions and arbitrary numbers of independent decisions. For example, a health insurance plan p chooses whether to report that an enrollee has a diagnosis d_i given enrollee characteristics $\mathbf{x}_i$. Agent assignment is *mutually exclusive* (*e.g.*, individuals are enrolled in one health insurance plan).

If the agent knows that d_i will be used as input to a payout model, they may have an incentive to *increase* d_i (without loss of generality) to obtain a higher payout. Let $\bar{d} = \frac{1}{M_p} \sum_{i=1}^{M_p} d_i$, and suppose each agent p chooses $\bar{d}$ according to the following utility-maximization problem:

$$P(d_i = 1 \mid p) \equiv \Delta_p(d_p^*) \triangleq \underset{\bar{d} \in [0,1]}{\arg\max}\ R(\bar{d}) - \lambda_p c(\bar{d} - d_p^*) \tag{5.2}$$

where $R : [0, 1] \to \mathbb{R}$, $c : \mathbb{R} \to \mathbb{R}_+$, and d_p^* is the ground truth value of $\bar{d}$ given the $\mathbf{x}_i$ seen by agent p. Thus, $\Delta_p(d_p^*)$ is the gamed/observed decision rate of agent p in a population where the ground truth decision rate is d_p^*. Although $\bar{d} \in [0, 1]$ since it is a proportion, our framework applies to $\bar{d}$ on arbitrary intervals. This formulation assumes that each $\mathbf{x}_i$ is equally likely to be gamed, that $\mathbf{x}_i$ are truthfully observed, and that any difference between $\Delta_p(d_p^*)$ and d_p^* is due to gaming.

We focus on the gaming deterrence parameter λ_p, which scales the cost of manipulation $c(\cdot)$. λ_p is non-negative and represents an agent's "aversion" to gaming. Lower values of λ_p mean that agent p is more willing to game. Thus, identifying agents most willing to game means finding agents with the lowest λ_p. We summarize multi-agent strategic adaptation in Figure 5.1. Next, we introduce assumptions on the reward function R, cost function c, and ground truth d_p^*.

Assumption 1 (Shared rewards & costs)**.** *Reward (R) and cost (c) functions are shared across agents.*

Sharing R encodes the assumption that all agents are reacting to the same payout model, while sharing c across agents encodes the belief agents must take similarly-costly actions to manipulate features (*e.g.*, if insurance plans must follow/fraudulently report specific procedures to justify a diagnosis). Many works in strategic classification implicitly make a similar assumption (*e.g.*, [103, 18]).

Assumption 2 (Increasing rewards)**.** *The reward function R is strictly increasing in $\bar{d}$.*

Increasing rewards formalizes the agent's incentive to perturb its decisions (*i.e.*, inputs to the payout model) from $d_i = 0$ to 1 in Eq. 5.2.

Assumption 3 (Cost convexity)**.** *The cost function c is strictly convex and minimized at 0 such that $c(0) = 0$ and $c'(0) = 0$, and increases for all agents p for any $\bar{d} \geq d_p^*$.*

One possible c is $c(x) = x^2$. Strict convexity ensures a unique cost-minimizing action, and $c(0) = 0$ ensures that d_p^* is ground truth, such that increasing $\bar{d}$ incurs greater cost (*e.g.*, Fig. 5.1, left).

Assumption 4 (Diminishing or linear returns)**.** *The reward function R is concave in d_i.*

For example, R may be a $\log$ or affine function. Assumption 3 (strictly convex c) and 4 ensure that R cannot grow fast enough to offset manipulation costs. Furthermore, due to Assumptions 2- 4:

Remark 10 (Gaming is utility-maximizing)**.** *Given any agent p and d_p^*, we have that $\Delta_p(d_p^*) \geq d_p^*$.*

Equivalently, optimal gaming entails increasing d_i from the ground truth. Note that Assumptions 2- 4 are more general versions of assumptions placed on rewards/costs in the strategic classification literature (*e.g.*, [103, 70, 170]).

Assumption 5 (Non-strategic behavior is feasible)**.** $d_p^* \in [0, 1]$ *is a constant depending solely on* $\mathbf{x}_i$.

Due to Assumption 5, ground truth d_p^* may vary by agent due to differences in $\mathbf{x}_i$ (*e.g.*, health insurance plans serve populations with varying levels of health).

5.1.4 Theoretical analysis: finding agents most likely to game

We aim to identify agents most likely to game a decision-making model, *i.e.*, agents with the lowest gaming parameters λ_p. Here, we prove that λ_p cannot be point-identified without further assumptions (Section 5.1.4.1), but ranking λ_p is possible via causal effect estimation (Section 5.1.4.2). Detailed proofs are in Appendix C.1.2.

5.1.4.1 Partial identification of the gaming parameter

Here, we show that given our assumptions, λ_p is only partially identifiable (cannot be uniquely determined):

Proposition 11. *Define $R'(\cdot)$ as $\frac{dR}{dd_p^*}$ and $c'(\cdot)$ as $\frac{dc}{dd_p^*}$. For any agent p, given Assumptions 1- 4 and an observed $\Delta_p(d_p^*)$,*

$$\lambda_p \in \left[\frac{R'(\Delta_p(d_p^*))}{c'(\Delta_p(d_p^*))}, \infty \right), \tag{5.3}$$

and the bound is sharp.

Intuitively, different values of the unknown d_p^* yield different estimates of λ_p consistent with the observed $\Delta_p(d_p^*)$. Thus, uncertainty in d_p^* results in uncertainty in λ_p. Equivalently, point-identifying λ_p requires *perfect* knowledge of d_p^*. Thus, without further assumptions, λ_p is only partially identifiable. The lower bound is attained for $d_p^* = 0$ (all $d_i = 1$ are manipulated), while $\lambda_p \to \infty$ as $\Delta_p(d_p^*) \to d_p^*$ (no manipulation). Intuitively, increases in λ_p further disincentivize increases to $\Delta_p(d_p^*)$, such that $\Delta_p(d_p^*)$ gets closer to d_p^*.

A naïve approach to gaming detection would be to rank individuals using the above bound. To see why this is problematic, consider an Agent 1 ($\lambda_1 = 10$) and Agent 2 ($\lambda_2 = 30$), and let $R(x) = x$ and $c(x) = x^2$. Suppose Agent 1 is a health insurance plan serving a relatively healthy population ($d_1^* = 0.05$), while Agent 2 serves a population with a higher burden of illness ($d_2^* = 0.12$). Via Eq. 5.2, we have $\Delta_1(d_1^*) = 0.10$, while $\Delta_2(d_2^*) \approx 0.14$. Substitution into Eq. 5.3 yields $\lambda_1 \geq 5$ and $\lambda_2 \geq 3.66$, flipping the true ranking of λ_p. Thus, acting on this bound may incorrectly penalize agents when a high $\Delta_p(d_p^*)$ is appropriate; *e.g.*, insurance plans serving sicker populations.

5.1.4.2 Identifying a ranking of the gaming parameter

Since we showed that point-identifying λ_p is impossible without further assumptions, we relax gaming detection to a *ranking* problem. Intuitively, differences in agent behavior under similar conditions may indicate different gaming capacities, from which the proposed approach follows.

Ranking λ_p by estimating counterfactuals. Recall that we aim to find agents with the lowest λ_p. To build a ranking strategy, first consider ranking two agents p and p'. Define $\mathcal{D}_{p,p'} \triangleq \{(\mathbf{x}_i, d_i, p_i) \mid i = 1, \ldots, n \text{ and } p_i \in \{p, p'\}\}$, where p_i is an indicator for the agent that observed example i. We follow the Neyman-Rubin potential outcomes framework [119] and assume that our data follow the structure of the causal graph shown in Figure 5.2 (left). Let $d_i(p)$ be the value of d_i if p_i was set to p (*i.e.* had agent p been *compelled* to make a decision). Such variables are called *counterfactuals*. For demonstration, Figure 5.2 (right) shows a toy dataset with counterfactuals $d_i(p), d_i(p')$, where "?" are unobserved decisions d_i. Dropping unobserved data, the average $d_i(p)$ is thus $\Delta_p(d_p^*)$ by definition. Our strategy hinges on the estimation of $d_i(p)$, a counterfactual. In general, our ability to do so rests on the following *identifiability* assumptions [216]:

Assumption 6 (Conditional exchangeability)**.** *For all i, $d_i(p_i) \perp\!\!\!\perp p_i \mid \mathbf{x}_i$, where $d_i(p_i)$ is the potential outcome of d_i under agent p_i.*

Assumption 7 (Consistency)**.** *For all i, $d_i(p_i) = d_i$.*

Assumption 8 (Positivity/overlap)**.** *For any agent p and $\mathbf{x} \in \mathcal{X}$, $0 < \mathbb{P}[p \mid \mathbf{x}] < 1$.*

Via Assumptions 6- 8, we have that $\mathbb{E}[d_i(p) \mid \mathbf{x}_i] = \mathbb{E}[d_i \mid \mathbf{x}_i, p]$ and $\mathbb{E}[d_i(p') \mid \mathbf{x}_i] = \mathbb{E}[d_i \mid \mathbf{x}_i, p']$ [124]. Hence, the causal effect is *identifiable*. Thus, when $d_i(p)$ and $d_i(p')$ are fully observed, we can estimate $\Delta_p(d_p^*)$ and $\Delta_{p'}(d_p^*)$ as follows:

$$\hat{\Delta}_p(d_p^*) = \frac{1}{n_p} \sum_{\substack{(\mathbf{x}_i, d_i, p_i) \in \mathcal{D}_{p,p'} \\ p_i = p}} d_i(p) \qquad \hat{\Delta}_{p'}(d_p^*) = \frac{1}{n_p} \sum_{\substack{(\mathbf{x}_i, d_i, p_i) \in \mathcal{D}_{p,p'} \\ p_i = p}} d_i(p') \tag{5.4}$$

where n_p is the number of observations by agent p. Here, despite the identifiability of $d_i(p)$, λ_p remains non-identifiable in practice. Given an invertible mapping from d^* to x, identifying λ_p is straightforward given the utility function, as λ_p encodes information about the gap between $d(p)$ and d^*. However, we only assume that d^* causally influences x with no further constraints. Without observations of d^*, it is unclear how such a mapping can be identified, or whether it exists. [1] However, since our goal is to identify worst offenders, it suffices to *rank* agents by λ_p, for which we demonstrate a monotonic relationship between λ_p and $\Delta_p(d_p^*)$:

Theorem 12. *Under Assumptions 1- 5, and $\Delta_{p'}(d_p^*)$ defined as*

$$\Delta_{p'}(d_p^*) \triangleq \arg\max_{\bar{d} \in [0,1]} R(\bar{d}) - \lambda_{p'} c(\bar{d} - d_p^*), \tag{5.5}$$

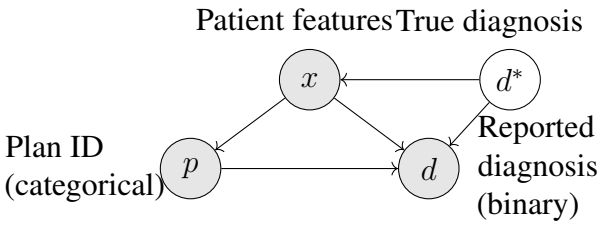
we have that $\Delta_p(d_p^*) < \Delta_{p'}(d_p^*)$ if and only if $\lambda_p > \lambda_{p'}$.

Theorem 12 tells us that estimation of $\Delta_p(d_p^*)$ and $\Delta_{p'}(d_p^*)$ can be used to rank λ_p vs. $\lambda_{p'}$. Eq. 5.5 differs from Eq. 5.2: while R, c, and d_p^* are the same, λ_p changes to $\lambda_{p'}$. While subtle, the distinction is key to gaming detection: $\Delta_{p'}(d_p^*)$ denotes what agent p' *would have* done in a population with ground truth d_p^*, as opposed to what agent p' *actually did* in their population (ground truth $d_{p'}^*$). Then, we proceed by estimating the effect of "swapping" agents:

Corollary 13. *Define $\tau(p, p') \triangleq \mathbb{E}_{\mathbf{x}_i}[\mathbb{E}[d_i(p) \mid \mathbf{x}_i]] - \mathbb{E}_{\mathbf{x}_i}[\mathbb{E}[d_i(p') \mid \mathbf{x}_i]]$. Then, given Assumptions 1-8, $\tau(p, p') > 0$ if and only if $\lambda_p < \lambda_{p'}$.*

Since $\mathbb{E}[d_i(p) \mid \mathbf{x}_i] = \mathbb{E}[d_i \mid \mathbf{x}_i, p]$, it is an unbiased estimator of $\Delta_p(d_p^*)$ by definition, and substitution into Theorem 12 yields the result. Thus, one can rank λ_p by estimating the effect of switching from agent p to p' on the observed decision rate for *all* pairs of agents p, p'. Each effect estimate is a pairwise comparison of agents, from which a ranking of $\lambda_{(\cdot)}$ follows. We summarize causally-motivated gaming detection in Fig. 5.3, with pseudocode in Fig. 5.4.

[1] One may also revisit Proposition 11 with knowledge of x. Doing so may improve the bounds on λ_p from Eq. 5.3 (*e.g.*, through sensitivity analyses [185, 137]). As our goal is to identify worst offenders, we need not exactly estimate or even bound treatment effects. See Appendix C.1.2 for further comment.



x_i	d_i	p_i	$d_i(p)$	$d_i(p')$
0.2	1	p	1	?
-0.3	0	p	0	?
-0.1	1	p'	?	1
0.4	0	p'	?	0

Figure 5.2: **Left:** Causal graph for gaming detection with confounders $\mathbf{x}$, agent indicator p, ground truth diagnosis d^*, and agent decision d. **Right:** Toy dataset with observed factual outcomes $d_i(p)$ and $d_i(p')$. "?" denotes missing counterfactual outcomes.

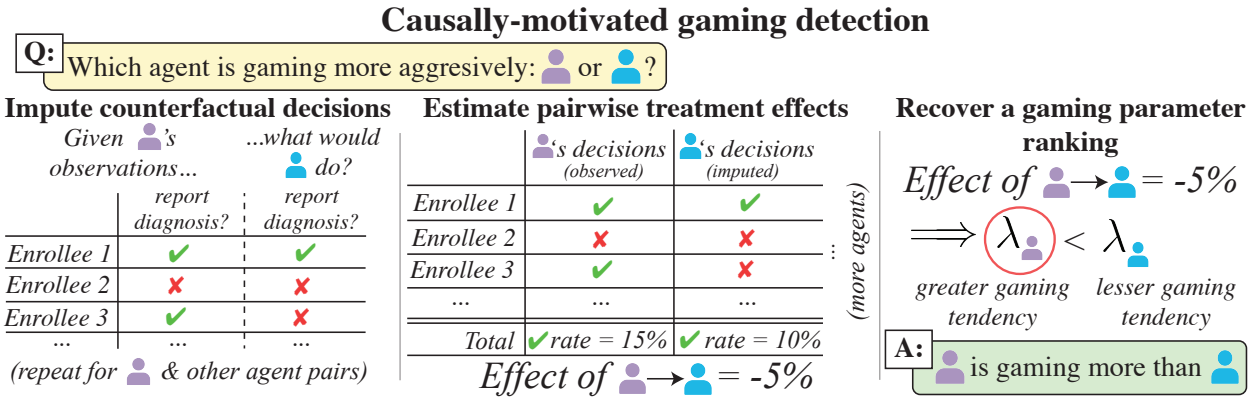


Figure 5.3: Causally-motivated gaming detection. **Left:** First, we impute counterfactual decisions for each agent. **Middle:** The imputed counterfactuals yield average treatment effects (ATEs) across *pairs* of agents. **Right:** Using ATE estimates to rank agents yields a ranking of the gaming parameter λ_p. We show one direction of comparison across agents without loss of generality. In practice, we impute decisions for both directions of comparison and average (given blue agent's observations, impute purple agent's decisions).

Obtaining a well-ordered ranking. Since we aim to rank λ_p, our chosen estimator should yield a well-ordered ranking. Via Corollary 13, the oracle treatment effect τ suffices, but τ must generally be estimated. We show that a "sufficiently" accurate estimate $\hat{\tau}$ also yields the desired result:

Proposition 14. *Let $\tau(\cdot)$ be the oracle treatment effect function as defined in Corollary 13, and $\hat{\tau}$ be its sample estimate. Given Assumptions 1- 8, for any $\varepsilon > 0$, if $\sup |\hat{\tau}(p, p') - \tau(p, p')| \leq \varepsilon$, then, for all p, p' such that $\min_{p,p'} |\tau(p, p')| > \varepsilon$, $\hat{\tau}(p, p') > 0$ if and only if $\lambda_p < \lambda_{p'}$.*

The result is immediate: sufficiently low estimation error in $\hat{\tau}$ (*i.e.*, $\leq \varepsilon$) cannot "flip" any pairwise rankings where $\tau > \varepsilon$. Thus, any consistent estimate of τ yields (asymptotically) a well-ordering of λ_p. We defer to past asymptotic analyses of causal effect estimation for further discussion [55, 201].

a) Causally-motivated gaming detection.

Input: $\mathcal{D} = \{(\mathbf{x}_i, d_i, p_i)\}_{i=1}^N$ and a list of P agents $\mathbf{p}$
Output: sorted list of P agents

$\hat{\tau} \leftarrow$ fitted causal effect estimator using $\mathcal{D}$
$\mathbf{T} \leftarrow$ empty P-by-P matrix
for all p_i, p_j where $i < j$ **do**
    $\mathbf{T}[i, j] \leftarrow \hat{\tau}(p_i, p_j)$
end for
$\mathbf{\Lambda} \leftarrow$ sort $\mathbf{p}$, using $\mathbf{T}$ to make comparisons
return $\mathbf{\Lambda}$

b) Python-style pseudocode.

```
train_data, test_data = split(dataset)
model = fit_causal_estimator(train_data)
comparisons = {}
for p_i, p_j in agents where p_i != p_j:
  # counterfactual est., plugging in agent IDs
  agent_i_cf = model.predict(test_data, p_i)
  agent_j_cf = model.predict(test_data, p_j)
  comparisons[(p_i, p_j)] = agent_i_cf -
      agent_j_cf

ranking = sort(agents, comparisons) # use the
    comparisons dictionary as a lookup table
return ranking
```

Figure 5.4: Pseudocode for causally-motivated gaming detection. Causal effect estimators take pairs of agents (p, p') and their observations $\mathbf{x}^{(\cdot)}$ as inputs, and output the mean difference in predicted decision rate.

5.1.5 Empirical results & discussion

We aim to demonstrate that causal inference can be used for gaming detection. First, we discuss our setup (Section 5.1.5.1). Then, we show in synthetic data (Section 5.1.5.2) that causal methods require fewer audits than existing non-causal methods to catch the worst offenders.

5.1.5.1 Setup

We describe the datasets, evaluation method, and gaming detection methods under consideration.

Datasets. In real datasets, ground truth gaming rankings are often unavailable. Thus, we validate our framework in a **synthetic dataset**. We hand-select 20 λ_p values (one per agent; see Appendix C.1.3.1 for raw λ_p values) and simulate confounding by generating covariates from agent-specific Gaussians with means $\boldsymbol{\mu}_p \in \mathbb{R}^2$. To control confounding strength, we set $\boldsymbol{\mu}_p = g(\log(\lambda_p))$, where g is an affine transformation such that the range of $\boldsymbol{\mu}_p$ across agents equals a chosen range parameter $R_\mu \in \{0, 0.1, \ldots, 1.0\}$. . Smaller R_μ implies less confounding, since $\boldsymbol{\mu}_{(\cdot)}$ varies less across agents. We generate 500 observations $\mathbf{x}^{(i)} \in \mathbb{R}^2$ and ground-truth $d^{*(i)}$ per agent via

$$\mathbf{x}^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}_{p^{(i)}}, \sigma^2 \mathbf{I}_{2\times 2}) \qquad d^{*(i)} \sim Ber(\alpha^{*(i)}); \quad \alpha^{*(i)} = \sigma(\mathbf{w}^\top \mathbf{x}^{(i)} + b),$$

where $\mathbf{w} \sim \mathcal{U}(0,1)^2$, such that increasing $\mathbf{x}$ increases $\alpha^{*(i)}$ (and thus $P(d^{(i)}=1)$), b is chosen such that $\alpha^{*(i)}$ for the mean $\mathbf{x}$ is $\approx 5\%$, and $\sigma^2 = 1$.[2] We simulate gamed agent decisions $d^{(i)}$ as follows:

$$d^{(i)} \sim Ber(\alpha_p^{(i)}); \quad \alpha_p^{(i)} = \underset{\tilde{d}_p \in [0,1]}{\arg\max}\ \log(\tilde{d}_p) - \lambda_p(\tilde{d}_p - d^{*(i)})^2.$$

Recall that the decisions $d^{(i)}$ are also agent inputs to a payout model. We generate 10 datasets (each $N = 10,000$; 20 agents $\times$ 500 observations) for all 11 levels of confounding (as measured by the range of means R_μ). Causal inference assumptions hold in the synthetic data: all confounders $\mathbf{x}^{(i)}$ are observed (Assump. 6), consistency holds by construction (Assump. 7), and overlap holds since all $\mathbf{x}^{(i)} \mid p$ are supported on $\mathbb{R}^2$ (Assump. 8). Full synthetic data generation details are in Appendix C.1.3.1.

To benchmark causal effect estimation in a more realistic setting, we apply causal methods to gaming detection in **U.S. Medicare** claims. Medicare is the public health insurance system in the U.S. for residents aged 65 and over. Since private insurance claims data is not widely available, we conduct a gaming case study in healthcare providers. In Medicare, the U.S. government pays healthcare providers on a per-service basis [1]. Thus, providers may be incentivized to label enrollees with as many diagnoses as possible to secure extra payment from the government. We select a 0.2% sample of all Medicare enrollees with a claim in 2018; *i.e.*, those who utilized a service covered directly by Medicare ($N = 37,893$). We use demographic information and diagnoses in 2018 as covariates and select the rate of uncomplicated diabetes diagnosis in 2019 as the outcome. Given differences in healthcare policy and access across U.S. states, we pool data at the U.S. state level and treat each state as an "agent." Additional cohort details are in Appendix C.1.3.2.

Evaluating rankings. Given an observational dataset of the form $\{(\mathbf{x}_i, d_i, p_i)\}_{i=1}^N$, with covariates $\mathbf{x}_i$, observed decisions d_i, and agent indicators p_i, gaming detection algorithms output an ordinal agent ranking in terms of the gaming parameter λ_p. We aim to measure the efficiency of a predicted ranking given some level of resources committed by a decision-maker (*i.e.*, # of agents audited).

Ground truth rankings are available in synthetic data. Thus, we measure the top-5 **sensitivity** at k (S_k), the % of top-5 worst offenders in the predicted top-k, and the **discounted cumulative gain (DCG)** at k, a weighted sum of ground-truth "relevance scores" for the top-k predicted agents, across audit intensities $k \in \{1, \ldots, 20\}$. For a K-agent dataset, we define the relevance score as $K+1$ minus the ground-truth rank (*e.g.*, true rank 1 = relevance K, true rank 2 = relevance $K-1$, etc.). Concretely, for r_i defined as the ith ranked agent in a predicted ranking, and rank$(\cdot)$ as the

[2]For ease of implementation, $\mathbf{x}^{(i)}$ is generated conditioned on $p^{(i)}$. This remains consistent with the causal DAG (Fig. 5.2, right), since the DAG factorizes as $P(D \mid P, X)P(P \mid X)P(X)$, or equivalently, $P(D \mid P, X)P(X \mid P)P(P)$, as in our data-generation process.

function returning the *ground-truth* ordinal rank with respect to λ_p, our ranking evaluation metrics are computed as follows:

$$S_k \triangleq \frac{1}{5}\sum_{i=1}^{k} \mathbb{1}[\text{rank}(r_i) \leq 5] \qquad \text{DCG}_k \triangleq \sum_{i=1}^{k} \frac{K - \text{rank}(r_i)}{\log_2(i+1)}. \tag{5.6}$$

Note that sensitivity is an all-or-nothing measure of audit quality given a fixed audit intensity k. However, DCG rewards higher predicted rankings for top-k worst offenders, regardless of the absolute ranking position. Furthermore, DCG weights decrease with predicted rank (Eq. 5.6), which prioritizes correctly ranking the worst offenders over correctly ranking agents unlikely to game. To summarize audit efficiency across k, we also report **area under the top-5 sensitivity curve (AUSC)** across k.[3]

In contrast, ground truth gaming rankings are unavailable in the Medicare cohort. As an exploratory analysis, we compare an estimated gaming ranking to 104 state-level healthcare statistics from the 2003-2017 National Neighborhood Data Archive [144] and 2018 Medicare Provider of Service files [2] relating to healthcare access and hospital information (*e.g.*, ownership and size). We report the top five statistics most positively and negatively correlated with our predicted rankings in terms of **Spearman rank-correlation**. State-level summary statistics used are enumerated in Appendix C.1.3.2.

Models. To demonstrate the utility of causal effect estimation for gaming detection, we compare non-causal baselines to causal effect estimators. Non-causal approaches include a payout-only ranking (based on $P(d_i = 1)$ per agent) and random ranking. We also compare to existing approaches in anomaly detection, which do not make causal assumptions, but assume that gamed decisions are outlier-like. We test k-nearest neighbor outlier detection (KNN) [232], empirical-cumulative-distribution-based outlier detection (ECOD) [174], and deep isolation forests (DIF) [298]. These methods use $(\mathbf{x}_i, d_i)$ as inputs to an "anomaly score" model. We use average within-agent anomaly scores as a ranking. We discuss other works in algorithmic anomaly/fraud detection in Appendix C.1.1.

We implement the following causal effect estimators. **PSM** fits a propensity score model to match points in one agent's population to its nearest neighbor in the other agent's population with respect to propensity score estimates [242]. The **S-learner** trains one outcome prediction model for all agents, while the **T-learner** trains one model per agent [158]. **DragonNet** jointly models the outcome (one prediction "head" per agent) and propensity score to control for confounding [256].

[3]AUSC is similar to the area under the ROC curve (AUROC) with the top-k worst offenders defined as the "positive class," except that the x-axis is the # of audits rather than false positive rate. Unlike AUROC, random performance is less than 0.5, but tends to 0.5 as the number of agents $K \to \infty$ (Appendix C.1.2.6).

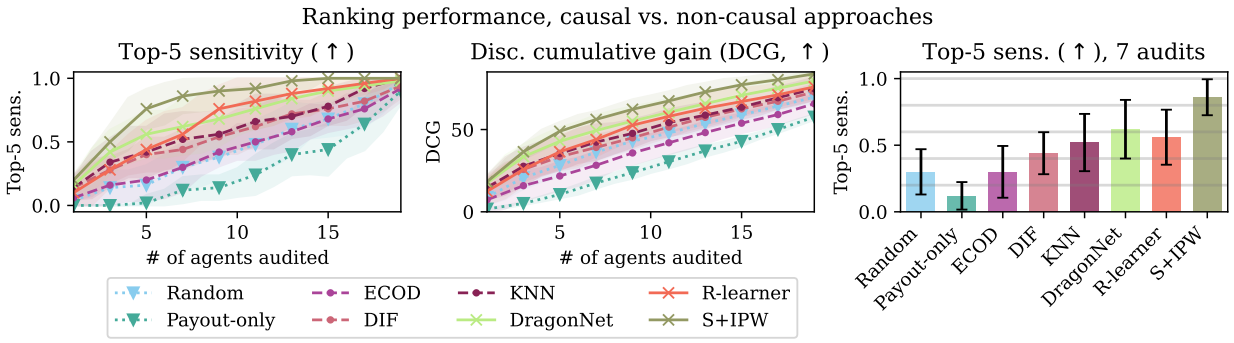


Figure 5.5: Mean top-5 sensitivity (left) and DCG (center) across # of agents audited, and top-5 sensitivity with 7 audits (right) at mean range 0.9, with $\pm\sigma$ error. Causal methods improve over non-causal baselines. ▽: naïve baselines. ○: anomaly detectors. ×: causal effect estimators.

S+IPW fits the S-learner with estimated sample weights that reweight the observed distribution to resemble an unconfounded distribution (randomized treatment assignment) [242]. The **R-learner** fits an outcome and propensity model, then regresses the residuals on one another to obtain a final unconfounded estimator [201]. Hyperparameters for all baselines and causal effect estimators are in Appendix C.1.5.

Implementation details. We use neural networks for all modeling (causal effect estimators + DIF). We use one-hot encoding for treatments (agent indicators). Matching across pairs of agents occurred without replacement (one-to-one), dropping unmatched individuals. We use generalizations of IPW and R-learners to multiple treatments, namely permutation weighting [8] and the structured intervention network [135], respectively. We perform a 7:3 dataset train-test split, training all models on the larger split. All rankings are computed on the test split. Early stopping is performed on a 20% validation split randomly sampled from the training set. Full modeling and training details, including the architecture and hyperparameters used, are in Appendix C.1.5.

5.1.5.2 Gaming detection in synthetic data

Causal effect estimators identify gaming more efficiently than non-causal baselines. Figure 5.5 shows the top-5 sensitivity and DCG of rankings produced by causal vs. non-causal gaming detection approaches at high confounding (mean range: 0.9). Since the S-, T-learner, and DragonNet perform similarly (Appendix C.1.4.2), we show only DragonNet here. Since PSM underperforms due to challenges with multi-treatment confounding control, we also defer results for PSM to Appendix C.1.4.2. Across audit intensities, causal approaches outperform non-causal methods in terms of sensitivity (*e.g.*, at 7 audits, Fig. 5.5, right; S+IPW: 0.860±0.135 vs. KNN: 0.520±0.215) and DCG (S+IPW: 56.1±5.11 vs. KNN: 42.3±10.8).[4] Trends are similar at other levels of confounding

[4]We report $\pm\sigma$ error bars as in the figures.

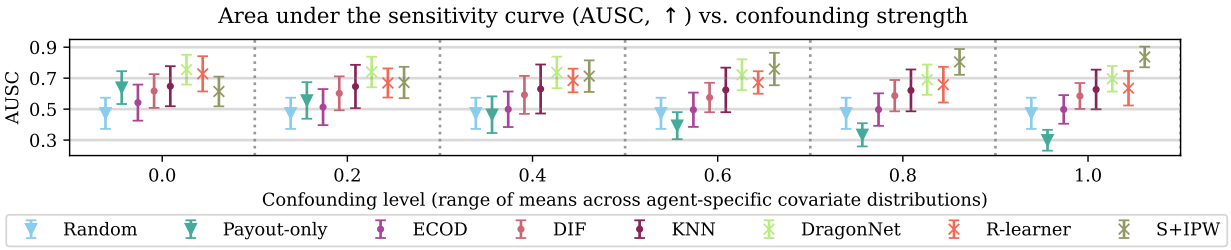


Figure 5.6: Area under the sensitivity curve (AUSC) for causal vs. non-causal methods across levels of confounding, with $\pm\sigma$ error. As confounding increases, a payout-only ranking degrades. Anomaly detection performance does not vary across confounding strength, maintaining slightly better than random rankings. Causal methods generally maintain higher mean AUSC than baselines across confounding levels. $\triangledown$: naïve baselines. $\circ$: anomaly detectors. $\times$: causal effect estimators.

(Appendix C.1.4.1), though the gain in ranking performance of causal approaches over non-causal methods diminish at lower levels of confounding (Figure 5.7; S+IPW AUSC: 0.614±0.096 vs. KNN: 0.648±0.129, mean range 0.0).

A payout-only approach yields worse than random ranking with sufficient confounding between covariates and agent decisions (payout-only AUSC: 0.299±0.067 vs. random: 0.473±0.101, Figure 5.6, mean range 1.0; *e.g.*, if healthier patients are enrolled in more gaming-prone plans). Anomaly detection methods (KNN, ECOD, DIF) are ill-suited for detecting gaming in dense regions of covariate space by design, while causal approaches would excel due to improved overlap. If outliers are more likely to be manipulated (*e.g.*, if populations with lower ground-truth d_p^* are more likely to be gamed), an anomaly detection method would identify gaming in such points. Indeed, anomaly detectors empirically outperform random ranking but lag causal methods. We further discuss why anomaly detection methods underperform causal approaches in Appendix C.1.4.1.

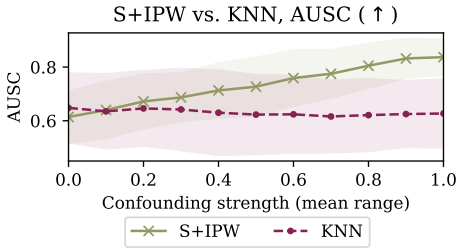


Figure 5.7: AUSC of S+IPW (causal) vs. KNN (non-causal) across confounding strength with $\pm\sigma$ error. The advantage of S+IPW over KNN decreases as confounding diminishes.

Trends in ranking performance vary across causal effect estimators. DragonNet and the R-learner degrade slightly as confounding increases (DragonNet AUSC: 0.755±0.096 → 0.696±0.083; R-learner: 0.728±0.114 → 0.635±0.111; mean range 0.0 → 1.0), likely due to slightly worse overlap. However, S+IPW *improves* as confounding increases (AUSC: 0.614±0.096 → 0.837±0.067; mean range 0.0 → 1.0). This is likely since the oracle IPW weights deviate from a uniform weighting as confounding increases. In such settings, estimation error in IPW weights may be less likely to incorrectly up-weight points that an oracle propensity score would down-weight, and vice versa. While such error would bias the pointwise treatment effect estimate, for the purposes of ranking, causal effect estimators only need to correctly estimate the *sign* of the treatment effect. Thus, we

hypothesize that our ranking may be less affected by such errors in the propensity score estimate. We leave formal analyses of the properties of IPW with respect to the sign of causal effect estimates to future work. A sensitivity analysis of all causal approaches is in Appendix C.1.4.2.

Takeaways. In a synthetic dataset, causal effect estimation approaches identified gaming more efficiently than non-causal baselines across levels of confounding. The empirical results provide proof-of-concept for causal effect estimation as a gaming detection method.

5.1.6 Conclusion

We propose a causally-motivated framework for ranking agents by gaming propensity in the context of strategic adaptation. We show that the gaming parameter is only partially identifiable, but a ranking of a set of agents based on the gaming deterrence parameter is identifiable via causal inference. We demonstrate the utility of causal effect estimation for gaming detection on synthetic data and a case study of upcoding in Medicare.

Limitations & broader impact. We assume agents always increase d_i with respect to ground truth, and that gaming explains all differences in agent behaviors, ignoring factors such as agent "quality" (*e.g.*, quality of care). Utility-maximizing behavior and conditional exchangeability are strong assumptions, but are statistically unverifiable. Many works in game theory and causal inference share these limitations. We caution that policies informed by our framework could reinforce imbalanced power dynamics (*i.e.*, are individual citizens [193] or more powerful entities gaming a model) via extraneous or weaponized accusations of gaming, since not all entities have equal capacity to respond to such claims. In particular, gaming by individuals may potentially reflect structural inequities rather than inherently pathological behavior. To mitigate such risks, we suggest "shadowing" studies (decisions visible, but not acted upon) alongside existing audit mechanisms before adoption.

5.2 Streamlining Oversight in Medicare Advantage: A New Approach to Measuring Coding Intensity via Contract Rankings

5.2.1 Introduction

Over half of all Medicare beneficiaries, or 34.4 million individuals, are enrolled in Medicare Advantage (MA). The Centers for Medicare and Medicaid Services (CMS) reimburses MA contracts a per-beneficiary, risk-adjustment payment ("capitation") intended to cover costs for their members' care, where sicker patients result in higher reimbursements, as determined based on Hierarchical Condition Categories (HCCs), or pre-defined groupings of diagnosis codes [53, 224]. While this system is intended to support care and mitigate adverse selection against beneficiaries with greater medical needs, it also creates a powerful financial incentive for insurers to "game" the model by overreporting HCCs to maximize subsequent payments, leading to high "coding intensity [189]." Excessive coding intensity is estimated to waste up to $40 billion USD in overpayments from CMS, straining the financial solvency of Medicare [189, 190].

The primary oversight tool employed by CMS is the Risk Adjustment Data Validation (RADV) audit. These audits target contracts, a bundle of Medicare coverage options offered by insurers, or Medicare Advantage organizations (MAOs). RADV audits estimate improper payments by taking a random sample of 35 enrollees per contract, requesting information from insurers to substantiate reported diagnoses, and estimating payment error based on unsubstantiated diagnoses [40]. However, the growing scale of the MA program presents a challenge to effective oversight. Recent announcements signal an intent to expand the scope of RADV audits to all eligible contracts. Thus, prioritizing contracts exhibiting excessive coding intensity for auditing could mitigate backlogs and streamline overpayment recovery [39].

Prior work measuring coding intensity generally relies on comparing diagnosis patterns in Medicare Advantage to those expected based on proxies of patient health derived from Traditional Medicare (TM) [189, 156, 154, 56]. Specifically, methods proposed by the Medicare Payment Advisory Committee (MedPAC) compute a score based on spending in TM conditional on age, biological sex, and dual-enrollment status, and other methods compute similar TM-derived baselines and assume any spending or excess risk score beyond the TM benchmark is inappropriate [189, 156]. However, applying such approaches to individual MA contracts risks conflating differences in coding practice with geographical variation in practice or differences in underlying beneficiary health. Other approaches propose hypothetical mechanisms of coding intensity; for example, examining whether diagnoses from chart reviews and in-home health risk assessments are associated with differences in the rate of specific treatments ancillary to reported diagnoses [292]. However, such approaches

are may become unreliable if insurers shift their coding strategies. Without a resource-intensive audit, it is difficult to determine whether gaps in diagnosis rates correspond to reporting diligence or extraneous diagnoses.

This study introduces a national approach for comparing coding intensity across MA contracts while addressing key limitations of prior work. Differences in underlying health are adjusted by stratifying beneficiaries on prior-year risk scores and restricting comparisons to contracts operating in the same geographic markets to control for regional practice patterns [262, 294]. To further isolate variation in diagnostic reporting behavior across contracts, we follow prior work and study beneficiaries who switched from TM to MA, to control for differences in prior diagnostic behavior across MAOs [35]. We then combine these adjusted contract-to-contract comparisons into a national ranking of coding intensity.

While excessive coding intensity has long been recognized as a problematic source of overpayments in MA, the recent push to audit all contracts by CMS makes methods for prioritizing contracts for auditing increasingly salient [39]. A high-quality ranking of contracts could improve overpayment recovery efficacy by reducing backlogs in the audit process. The proposed approach, which makes fewer assumptions than previous methods about mechanisms of coding intensity, directly compares MA contracts' reported diagnosis rates to one another, allowing the method to remain applicable under shifts in insurer strategy in response to new payment or auditing policies enacted by CMS. Overall, the results suggest that the underlying method provides a scalable, generalizable basis for audit prioritization.

5.2.2 Methods

Study cohort. The dataset consists of a 20% sample of MA beneficiaries in 2019. Only claims eligible for risk adjustment based on the Version 22 HCC definition are considered, which defines mappings from ICD-10 codes to reimbursement-eligible categories, and the 2019 listing of eligible CPT codes. The resultant codes are a subset of inpatient, outpatient, and carrier claims. In contrast to current MedPAC practice, which examines growth in risk scores within concurrent TM and MA cohorts, this study examines switchers, or beneficiaries in 2018 TM who switched to MA in 2019. The switchers' design ensures that past MA coding patterns do not pollute estimates for coding intensity constructed using adjustment on 2018 risk.

This study is restricted to non-dual, aged, non-ESRD beneficiaries, since these beneficiaries are subject to the same risk adjustment model, and thus similar diagnostic misreporting incentives. Further exclusions include beneficiaries who moved states, since MA offerings are geographically scoped, and those with any 2019 hospice claims, since hospice is not covered by MA. To ensure all MA switchers are subjected to similar exposure to MA coding strategies, beneficiaries were only

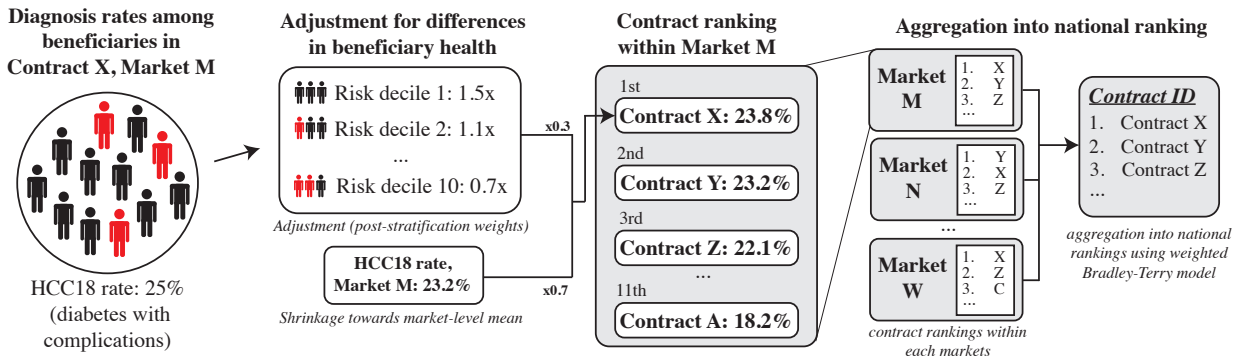


Figure 5.8: Overview of the proposed contract ranking pipeline. From left to right: for each diagnosis, we compute the per-contract, per-market prevalence. We adjust for differences in beneficiary health using post-stratification weights and apply shrinkage towards the market-level mean, yielding a contract ranking for the market. These rankings are aggregated across diagnoses and markets in a weighted Bradley-Terry model to produce a final, national ranking.

included if they were alive at the end of the study period (December 2019) consistent with MedPAC practice and stay with the same MA contract for the entirety of 2019 [189]. Finally, beneficiaries in counties where only a single contract is observed are excluded: since the proposed approach leverages local comparisons between contracts, such comparisons are ill-defined in one-contract counties.

Data Extraction. We extract HCCs from diagnosis codes from the 2018 TM and 2019 MA encounter files using the same 20% random sample as the study cohort. Inpatient and skilled nursing facility claims for TM only are sourced from the Medicare Provider Analysis and Review (MedPAR) file. For downstream validation, we also leverage prescription information sourced from Part D events files. All data extraction and analysis were performed using open-source Python packages.

5.2.2.1 Statistical analysis

The method outputs a ranking of MA contracts based on coding intensity. Motivated by causally-motivated gaming detection, a technique to rank entities such as MA contracts that modify their behavior in response to an incentive, this ranking is constructed using pairwise comparisons of diagnosis rates in geographically co-located contracts, adjusted for prior-year risk observed in TM [43]. Diagnosis rates are computed with respect to HCCs defined under the V22 CMS-HCC model. To ensure stable estimates and avoid overinterpreting extremely sparse diagnoses, we restrict the HCC set to those with a prevalence of at least 0.5% within the switchers' cohort. A list of the 45 eligible Version 22 HCCs (of 79 total) is shown in Appendix Table C.3.

Estimating rankings. The proposed approach generally proceeds in three stages: within-market adjustment and ranking, then aggregation across markets. Markets are defined as counties where beneficiaries have the same choice of MA contracts. In each market, adjusted HCC rates are computed for each contract using post-stratification as follows. First, national risk deciles are computed via the HCCPy package using reported ICD-10 codes, age, and biological sex in 2018 TM, using CMS-HCC weights for the community non-dual aged cohort. Then, per-decile diagnosis rates are computed, with shrinkage towards the market-level mean to allow for well-defined estimates in small contracts. Finally, a weighted average of diagnoses by risk decile is calculated to produce an adjusted diagnosis rate for each contract and market. These weights are derived from the prevalence of each risk decile in the national cohort, ensuring the adjustment targets a nationally representative population of beneficiaries. Note that decile weights are not exactly equal to 10% due to ties in CMS-HCC model risk scores. After adjustment, per-contract HCC rates are sorted into a within-market ranking of contracts.

To combine within-market rankings, we observe that each ranking implies a set of "matchups" between contracts for each HCC. Each matchup is assigned an importance value based on the difference in HCC rates between contracts, market size (number of beneficiaries), and HCC model weight, corresponding to the risk-adjusted payment associated with an HCC. The results from these matchups are used to fit a Bradley-Terry model, a model used to estimate relative team strength in competitions given incomplete matchups between teams [32, 38]. Such aggregation ensures that contracts that tend to code at higher rates than their geographically co-located competitors are ranked high, while those that often code below their peers are ranked low.

This adjustment, ranking, and aggregation process is re-run across 1000 within-market, within-contract bootstrap samples, and the median and interquartile range (IQR) of the final rank of each contract is reported. As a summary of the rankings, the share of contracts by majority Medicare plan type (Health Maintenance Organization/HMO or Preferred Provider Option/PPO) within each decile of our ranking is also reported. A formal description of the ranking pipeline is in the Appendix.

5.2.2.2 Evaluating Rankings

Validating the rankings against ground truth is challenging, as audits are sparse. We mitigate this with internal and external validity checks that verify our rankings align with properties expected of a ranking consistent with gaming detection, as well as a simulation study to help contextualize what levels of accuracy would be meaningful. We caution that these checks are best interpreted as supplementary evidence rather than a definitive claim of correctness.

First, to understand the importance of adjustment, we report the Kendall tau correlation between the proposed ranking and naive rankings constructed based on unadjusted and adjusted average per-contract risk scores based on 2019 MA [141]. Uncertainty in the ranks is compared to that of

a random ranking to ensure the proposed approach is meaningfully differentiating contracts from one another. Then, as a proxy for external validity, differences in the rate of relevant treatments (prescriptions or procedures) in the same year conditional on particular HCCs are analyzed, based on the assumption that legitimate diagnoses are associated with higher rates of relevant follow-up treatments [292, 302]. Exact HCC-to-treatment mappings selected can be seen in the Appendix. To mitigate missingness in prescription drug data, per-contract treatment rates are only computed in beneficiaries with 12 months of Part D coverage.

To supplement these tests, a simulation study was also performed to assess the practical utility of rankings as a function of their quality. The simulation study measures how quickly contracts that code intensely would be identified for various distributions of diagnosis rates. Noise is added to the ranking to simulate rankings with varying levels of accuracy as measured using the c-index [106]. Further details on this simulation study can be found in the Appendix.

5.2.3 Limitations

This study leverages a switchers' cohort to adjust for prior-year diagnoses unaffected by gaming incentives. However, this decision restricts the dataset size, as switchers account for a minority of Medicare beneficiaries, which may bias the results due to small sample size or potential non-representativeness. As a result, some contract ranks are highly uncertain, with wide rank IQRs. Furthermore, since this work uses observational data, proper adjustment is critical to the interpretation of the results. Unobserved confounders that drive both beneficiary choice of contract and reported diagnosis rates (e.g. through favorable selection) may not be captured in our data, and can bias comparisons between contracts, introducing noise to the final ranking. To mitigate these limitations, the validation strategy for the proposed approach aims to quantify similarity between the proposed ranking and those generated based on previously studied signals of gaming. In addition, a simulation study is performed to evaluate robustness to errors in the ranking.

This ranking is also agnostic to contract size/enrollment except through the effect of shrinkage, since the proposed ranking method is based on local comparisons of contract-level diagnosis rates. This means our ranking should not be interpreted as a measure of total misreporting volume. This design choice allows for flagging gaming in small contracts that may otherwise continue to misreport diagnoses, while retaining the option to make enrollment-aware auditing decisions. In addition, only claims submitted in 2019 are analyzed. This may not be indicative of current rankings. However, the approach is general and can be applied to other years or claims combined from multiple years.

Lastly, the proposed approach assumes that coding intensity above the market average by an MA contract is inherently suspicious. Without gold-standard validation of diagnostic misreporting, this assumption is statistically unverifiable. However, we support our method with internal and external

Table 5.1: Summary statistics for switchers' cohort ($n = 128{,}455$ beneficiaries).

Characteristic	**Value**
Median age in 2018 (IQR), years	72 (68–77)
Female, %	55.2
Race/ethnicity, %	
White	84.4
Black	9.0
Hispanic	1.1
Asian	1.5
North American Native	0.3
Unknown / Other	3.9
Average CMS-HCC risk score in 2018 (IQR)	0.64 (0.35–1.08)
Median number of HCCs per beneficiary (IQR)	1 (0–2)

validity checks that examine whether the proposed ranking demonstrates properties of a ranking consistent with gaming detection, in addition to a simulation study to help contextualize what levels of accuracy would be meaningful.

5.2.4 Study Results

After applying exclusion criteria (Figure 5.9), the resultant cohort contains 128,455 Medicare beneficiaries who switched from TM into MA, enrolled in 427 MA contracts representing 2,388 counties in 1,894 markets. Table 5.1 reports further summary statistics of the cohort. Furthermore, the median contract size in the cohort was 49 (IQR: 11.5-138.0) beneficiaries, each serving a median of 10 (IQR: 4.0-23.0) counties. The median age of beneficiaries was 72 years in 2018, with an interquartile range of 68 to 77 years (9 years). The switchers' cohort skewed slightly female (55.2%), and was predominantly White (84.4%), with a median HCC count of 1 (IQR: 0 to 2) and a pre-switch V22 CMS-HCC risk score of 0.64 (IQR: 0.35-1.08) pre-switch.

Median market-level adjusted diagnosis rates with 95% empirical CIs are shown in Figure 5.10 for diabetes with (left) and without complications (right) in Los Angeles County, California (N=3,199), the largest county in the switchers' cohort by enrollment, color-coded by their position in the national ranking (red = high, blue = low). County-level estimates are combined across thousands of U.S. counties and multiple HCCs to yield the national-level ranking shown in Figure 5.10. We present the top 20 contract-level rankings by median rank (Figure 5.11); the full ranking can be seen in Figures C.26 to C.30. As a summary of the full rankings, we plot the share of contracts by dominant plan type (HMO or PPO) in Figure 5.12 for each decile of our ranking, defined based on

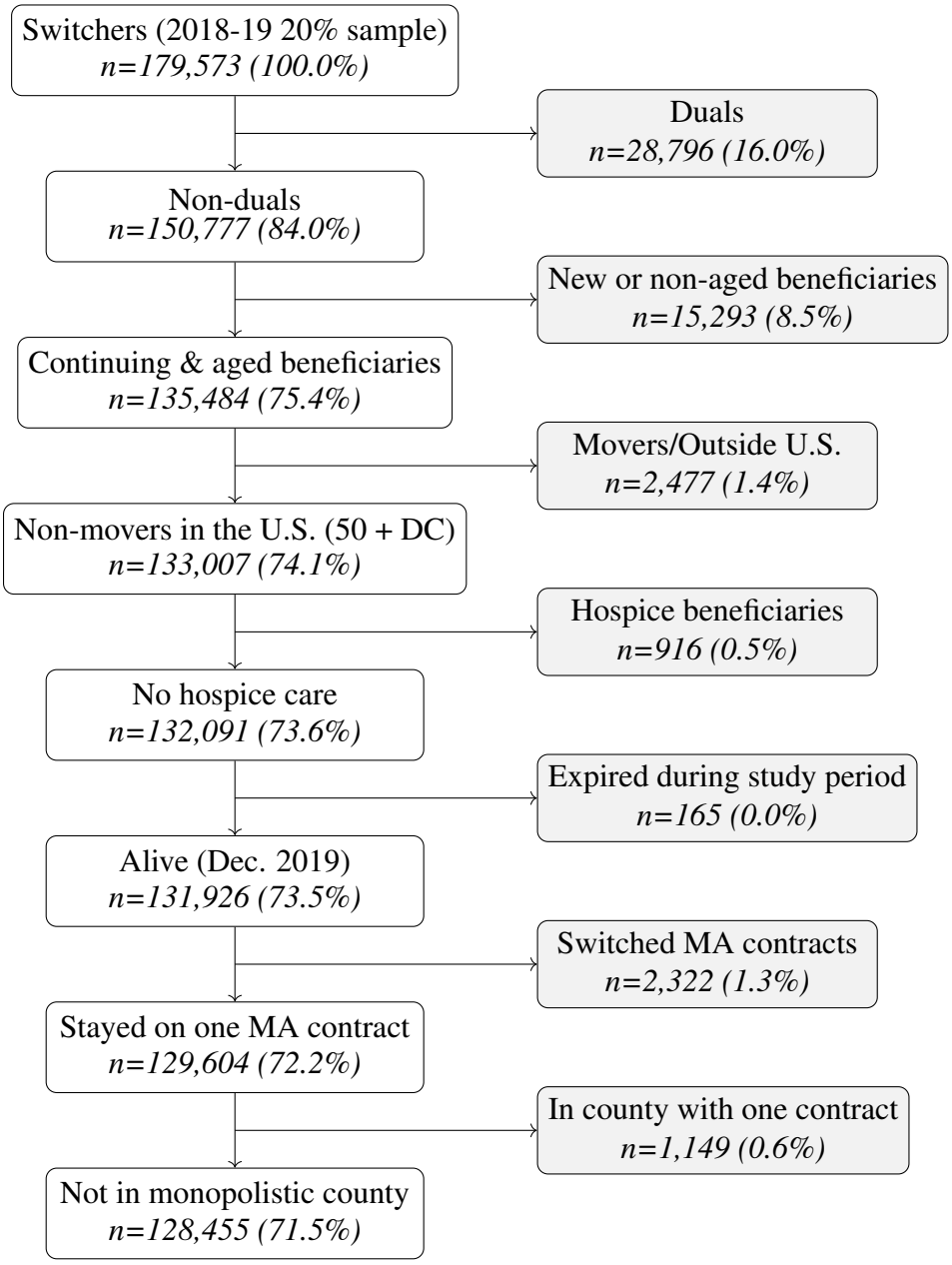


Figure 5.9: Cohort diagram with exclusion criteria for TM-to-MA beneficiaries analyzed. Starting with a switchers cohort, we focus on non-dual beneficiaries to isolate... that did not move states, were not on hospice care and survived to the end of the study period (Dec. 2019), stayed on a single MA contract, and did not reside in a monopolistic county (only one MA contract available). .

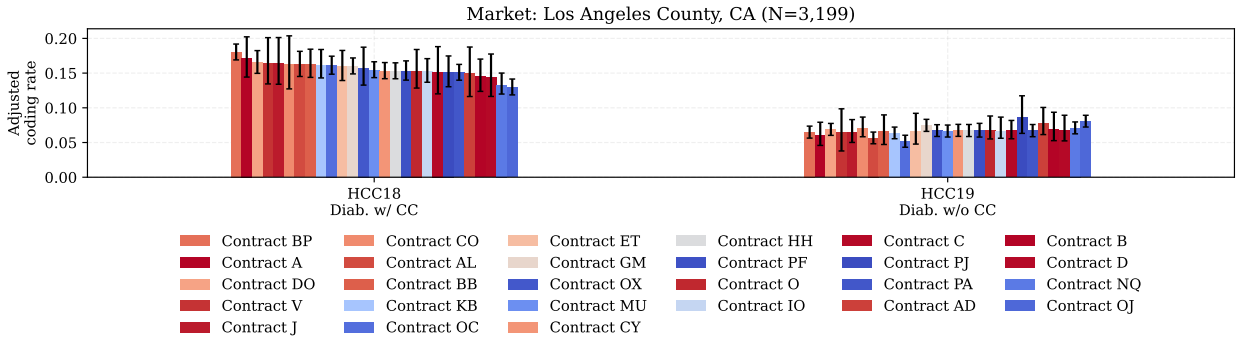


Figure 5.10: Local, shrinkage-based and adjusted diagnosis rates for contracts (names anonymized) in Los Angeles County, CA for diabetes with (left) and without complications (right). Red indicates high national rank (top 50%); blue indicates low national rank (bottom 50%), with increased saturation indicating more extreme (high/low) ranks. While county-level signals are noisy, we aggregate such patterns across HCCs and across counties to produce the national ranking. Contract identifiers are anonymized using alphabetical labels (A, B, . . . , Z, AA, AB, . . .).

whether a contract predominantly (>50%) offers HMOs or PPOs. The share of HMOs is largest in high-coding intensity deciles and decreases in favor of PPOs and cost-based plans as coding intensity decreases.

Internal Validation. The Kendall tau correlation between our ranking and a national unadjusted ranking (observed per-contract mean CMS-HCC score) is 0.191 (empirical 95% CI: 0.158 - 0.223). The correlation increases to 0.249 (95% CI: 0.215 - 0.281) when comparing with a naive adjusted national ranking that ignores potential differences between markets. Full correlations between all pairs of rankings tested can be seen in Appendix C.5. In the permutation test, error bars in the permuted cohort also widened compared to the original ranking, as expected: in the original ranking, the middle 50% of rank IQRs span 41 to 112 spots (9.6 - 26.2% of all contracts), while in the permuted ranking, the middle 50% of rank IQRs span 157 to 289 spots (36.8% - 67.7% of all contracts).

Partial external validation. For all five conditions analyzed, namely, diabetes with complications, psychiatric conditions, arrhythmia/vascular COPD, and Parkinson's, we observe positive gaps in relevant treatment rates across deciles of contracts (Figure 5.13). These gaps range from 0.13% per decile (diabetes with complications) to 1.08% (COPD). Raw treatment rates and enrollment counts for each decile are reported in the Appendix in Tables C.6 through C.10.

Simulation study. Our simulation indicates that, all else equal, higher variance in misreporting rates and lower mean misreporting rates make it easier to produce rankings that recover overpayments more quickly. When auditing 10% of contracts, a perfect ranking (c-index = 1) can catch

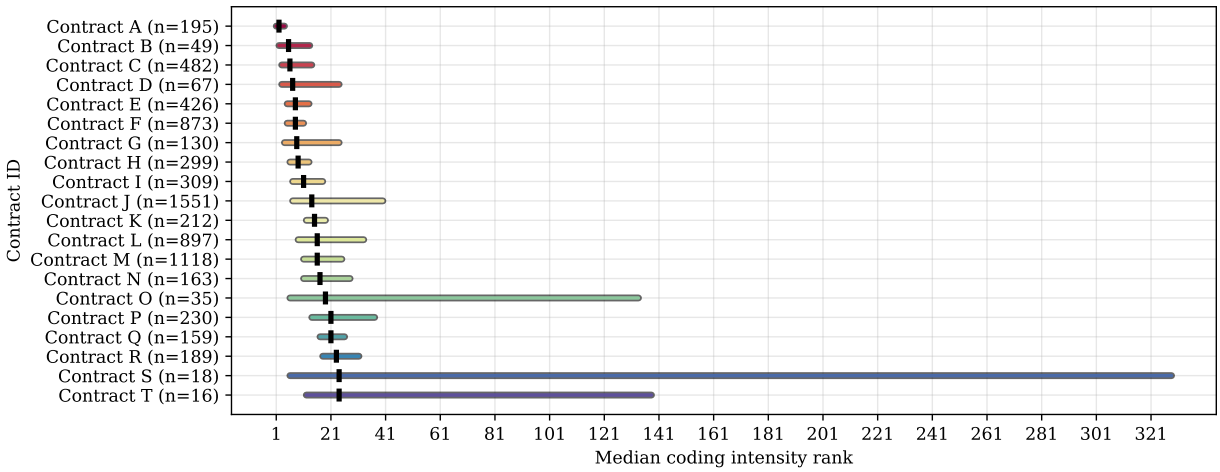


Figure 5.11: Median rank and IQR (width of the bars) across bootstraps (n=1000) of coding intensity rank for top 20-ranked contracts (sorted by median rank in ascending order), with number of switchers in each contract. A higher rank (further left) indicates more intense coding.

over 80% of all misreporting (Figure 5.14, left), at the smallest misreporting rate (10%) and highest variance (6%) tested. Under such conditions, even a moderately inaccurate rating (c-index: 0.65 - 0.70) can recover 83.6% of misreporting at a 10% audit rate (Figure 5.14, left), while a random audit would only be expected to recover 10%. However, for very high misreporting rates (30%) and low variance (2%), an audit with a similarly inaccurate ranking would recover fewer than 20% of misreporting, though this recovery still represents a marginal improvement over a random audit (Figure 5.14, right).

5.2.5 Discussion

CMS' recent announcement of an aggressive auditing strategy for MA makes prioritizing contracts for auditing increasingly timely, given limited oversight resources. In response, we propose a general approach for ranking MA contracts by gaming propensity that leverages direct comparisons between reported diagnosis rates within markets, with adjustment for prior-year risk and geographical variation in clinical practice norms. Although the proposed rankings are positively correlated with naïve rankings constructed using only national diagnosis rates, the modest magnitude of these correlations suggests that our approach adds information beyond national measures, a conclusion reinforced by the degradation of rank stability in the permutation test. Beyond statistical checks, we find that HMO-dominated contracts tend to be ranked higher than PPO-dominated contracts, echoing prior analyses [56, 89]. Furthermore, higher-ranked contracts tend to have lower rates of relevant treatments as well, consistent with previous work [292]. A simulation study further

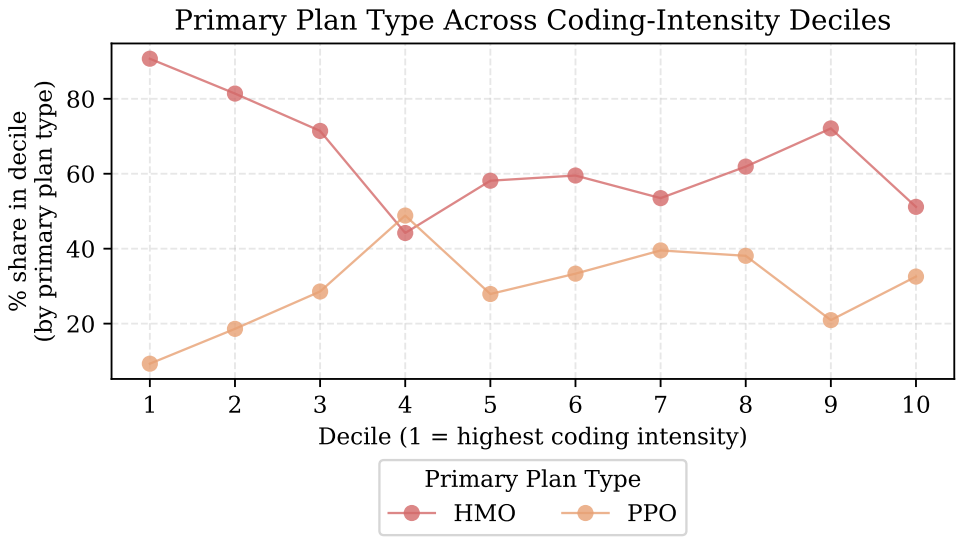


Figure 5.12: Share of contracts by primary plan type (HMO vs. PPO) by decile (1 = highest coding intensity). The share of HMOs relative to PPOs decreases along the coding intensity ranking.

establishes that, when diagnosis rates are low, but spread in diagnosis rates is high, even a noisy ranking can result in orderings that recover overpayments over 8 times faster than a random ranking.

Taken with prior work, the results reinforce concerns about systematic excessive coding intensity in MA as a major source of overpayments in MA. Our results complement prior analyses of excessive coding intensity focused on heterogeneity in coding intensity across MA contracts, echoing prior findings of differential coding across MA insurers [154, 56]. While differences between MA contracts identified in this study are not certain to be overpayments, that payments in MA exceed those in TM even after adjustment for combinations of enrollee case-mix or geographical factors is well-documented using various methods [189, 156, 89].

The proposed approach represents an alternative paradigm for measuring coding intensity, eschewing the assumption that coding above the TM baseline alone is abnormal. While there is merit to comparisons against TM, doing so may be fraught for the purposes of comparison between MA contracts for audit prioritization. In particular, MAOs have historically responded to TM-anchored estimates with claims that potential errors in TM coding rates bias estimates of overspending [279, 278] Thus, our approach represents a data-driven middle ground: we compare MA contracts to one another without making claims about appropriate coding rates.

While changes to Medicare payment policy, such as modeling changes that redefine reimbursement-eligible ICD codes, blanket payment reductions (e.g., coding inflation factors), or incorporating additional data sources such as patient health survey data might temporarily mitigate MA coding intensity, they may also create new avenues of gaming not present in prior

risk-adjustment models [187]. Approaches to MA contract ranking for audit prioritization that do not require direct knowledge of MA coding strategies may remain valid even as MAOs potentially shift coding strategies in response to policy changes. Although this study does not propose any novel mechanisms of excessive coding intensity, the fact that the ranking produced by the proposed approach independently aligns with previously studied signals of coding intensity supports the usage of MA contract comparisons as a means for audit prioritization.

If adopted by CMS, the efficacy of such an auditing approach depends on the accuracy of the ranking. While ranking accuracy is not possible to validate exactly, the simulation study provides insight here: if misreporting is concentrated in a small number of contracts, then it is possible in principle to recover a large proportion of overpayment with few audits. Conversely, if most contracts code at similar rates, even a perfect ranking will be slow to recover overpayment. The study finds that overall diagnosis rates for most HCCs are generally low ($<10\%$), suggesting that our cohort is closer to the regime represented by the left side of Figure 5 (low variance, low diagnostic reporting rate). Thus, even a moderately inaccurate ranking could still yield much faster overpayment recovery for CMS than a random audit.

5.2.6 Conclusion

As CMS continues to expand its auditing efforts in MA, methods to prioritize contracts at high risk for gaming become increasingly critical. This work presents a data-driven approach to ranking MA contracts by gaming propensity by way of pairwise comparisons between MA contracts. Without anchoring to TM or knowledge of potential gaming mechanisms, our method independently recovers potential signals of gaming, namely, gaps in treatment rates for conditions of interest post hoc. In settings with low reported diagnosis rates and high variation in misreporting rates, our ranking can flag the worst offenders far more efficiently than a random ranking. Ultimately, our ranking could inform audit prioritization strategies for CMS that remain robust to shifts in MAO coding behavior.

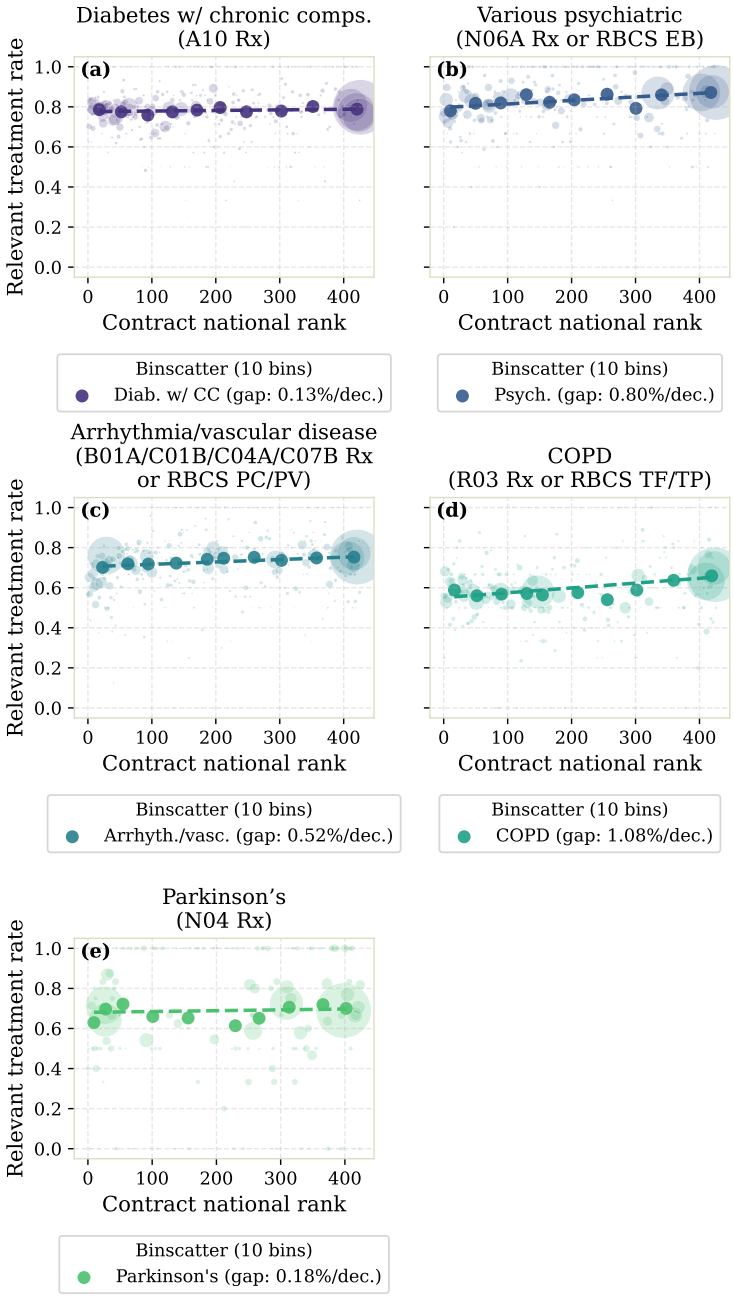


Figure 5.13: Binscatter plot of relevant treatments with respect to rankings computed using our approach over **a)** diabetes with complications (HCC18), **b)** various psychiatric conditions (HCC58), **c)** arrhythmias/vascular disease (HCC96, 107, and 108), **d)** COPD (HCC111), and **e)** Parkinson's disease (HCC78). Transparent dots = treatment rate for a contract, where the size of the dot corresponds to the number of beneficiaries with the HCC in the contract. Solid dots = per-decile treatment rates, with an enrollment-weighted trendline (dashed) summarizing per-decile treatment rates.

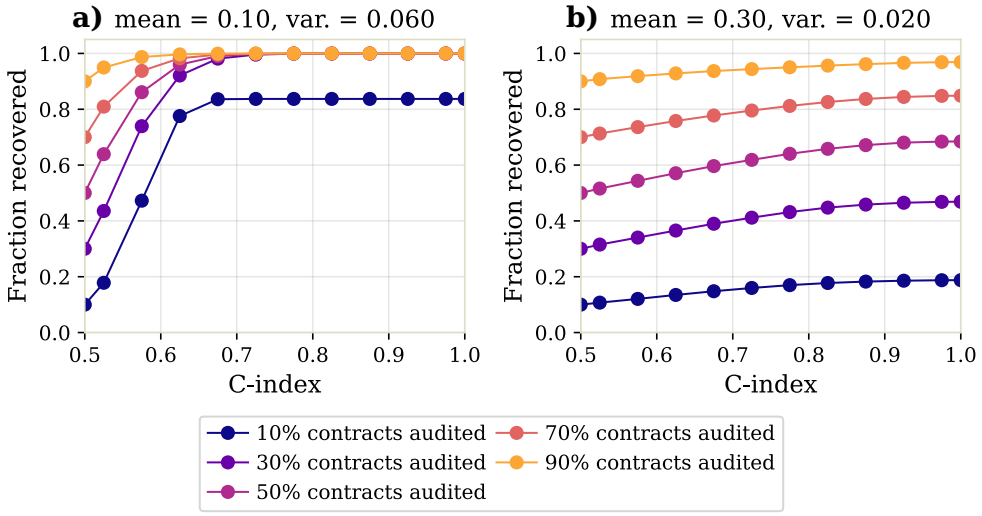


Figure 5.14: Graphs representing the fraction of misreported diagnoses audited as a function of the c-index (accuracy of a ranking) for different auditing intensities (10%, 30%, 50%, 70%, and 90%) and contract-level diagnosis rates. **a)** displays the fraction of recovered diagnoses when the mean diagnosis rate is 0.1 and the spread of the diagnosis rate is 0.06; **b)** displays the fraction of recovered diagnoses when the mean diagnosis rate is 0.3 and the spread of the diagnosis rate is 0.02.

5.3 Concluding remarks on gaming

The CMS-HCC model is an instance of manipulation in ML, in which users of an ML model in production exploit its functionality to obtain desirable outcomes. When the outputs of such models are used to allocate resources, such as public funds as in the CMS-HCC model, contention for resources can amplify the negative societal impacts of model manipulation.

Our study on gaming represents a setting in which the interests of two sets of ML model users are opposed: those of the model developer (*e.g.*, the U.S. government) versus those affected by the model's decisions (*e.g.*, insurers). Clearly, no model can be aligned with both sets of values: the insurer's assumed profit motive trades off with the public interest in minimizing government spending. Thus, our work, and by extension, systemic alignment, cannot be neutral. We take the side of the model developer, and propose an approach for prioritizing audits in the absence of ground-truth labels indicating manipulation.

Yet the systemic layer of ML alignment does not necessitate taking the side of the model developer. It is not necessarily true that manipulation is normatively bad: depending on the context in which an ML model is used, manipulation may be a symptom of structural inequities rather than inherently pathological behavior. Mathematical solutions thereof may further reinforce these inequities: improving the model robustness against manipulation often requires imposing more conservative decision thresholds that allocate fewer total resources, benefiting the institution at the expense of individuals in need facing the model decision, a tension explored in prior work [193].

Accordingly, systemic alignment requires considering not only which agents may engage in manipulation, but also the broader set of norms that the model ultimately establishes. The tools developed in this chapter are designed for deployment at the level of large health insurers and are intended to function within an institutional audit apparatus. However, when analogous mechanisms are applied to individuals, who possess substantially less capacity to contest determinations or absorb adverse consequences, we risk entrenching existing inequities under the guise procedural neutrality. Thus, operationalizing systemic alignment requires not only technical refinements, but also account for whose interests are prioritized, and who is rendered vulnerable.

CHAPTER 6

Conclusion

In this thesis, we explored the alignment of ML models with high-level principles and values specific to real world domains. In particular, there remains a gap between objectives used to optimize ML models and the values we would like such models to encode in their output distribution. Thus, we considered alignment at multiple layers of the ML pipeline: 1) intrinsic factors, which relate to the model's underlying behavior, 2) functional factors, which relate to whether a model can be steered towards user goals, and 3) interactional factors, which relate to how users interact with or manipulate the model. We argued that alignment at all layers of the ML pipeline is critical to responsible usage of ML.

6.1 Summary of contributions

Our contributions relate to alignment at multiple layers of the machine learning pipeline, from intrinsic data properties to functional model behavior, and ultimately, systemic dynamics that lead to certain user-model interactions.

First, we examine intrinsic factors in alignment through the lens of disparate censorship, or systematic biases in laboratory testing, in Chapter 3. Prior work has leveraged laboratory tests directly as labels for risk-stratification models [4, 115, 80, 235]. We document disparate censorship in practice in two large U.S. academic hospitals, develop theory that inequitable behavior under disparate censorship can only be avoided in restrictive cases, and propose an expectation-maximization based solution that treats disparate censorship as a noisy-label problem, in which recover of a true, "debiased" label is the objective.

Second, we examine functional factors in alignment through the lens of steerability in conditional generative models in Chapter 4, where steerability pertains to whether a model can produce outputs aligned with user goals. Prior evaluations of LLM performance generally span binary correctness-based instruction-following benchmarks [227, 310, 109] or continuous one-dimensional metrics such as estimated probabilities of desired behavior or concept-detection scores [209, 277]. Other

datasets for training are sourced from chat logs, web-scraped interactions, or capability-specific test suites [153, 307, 229, 54, 113], which may reflect only a narrow subset of potential users and goals. We propose a novel evaluation framework for measuring steerability, and benchmark a variety of LLMs using our framework. Our study surfaces side effects in LLM behavior: cases where LLMs induce unreqeusted changes in text, and find RL to be a partial but promising solution for improving steerability.

Lastly, we examine systemic factors through the lens of strategic manipulation in Chapter 5. We focus on the CMS-HCC model, a U.S. government risk-adjustment model for reimbursing private insurers for paying for healthcare for their beneficiaires in the Medicare Advantage program. This setting is particularly challenging: in contrast to previous fraud detection approaches, gaming may not be outlier-like [232, 60, 176, 105, 96, 174, 298], and the lack of ground-truth labels impedes supervised approaches [17, 102, 282]. Furthermore, prior analyses of gaming generally anchor to spending or underlying beneficiary health as measured in Traditional Medicare [188, 155], which may be confounded by differences in patient populations, coding practice, or care environments. We develop theory about causally-motivated gaming detection, relating construction of a ranking of gaming propensity to a causal inference problem. We then adapt our approach to MA claims data, finding that higher rankings correlate with lower treatment rates, matching previously-studied signals of gaming.

6.2 Future extensions

Several technical extensions to our work are applicable for improving alignment and evaluation thereof at each layer of the ML pipeline considered.

With respect to intrinsic factors of alignment, our work in disparate censorship highlights a fundamental challenge: in real clinical data, we rarely know whether groups differ only in marginal covariates, only in conditional risk, or in both. Because each setting produces different downstream effects under undertesting, recovering the correct representation of risk from perturbed labels may be non-identifiable without additional assumptions. Future work could explore practical tests or heuristics that help determine which setting best approximates a given clinical context. More generally, beyond disparate censorship, how one can infer true outcomes when both the features and the label-generation process are entangled with structural disparities remains an open but salient problem in aligning the intrinsic behavior of ML models.

For functional factors, applying our steerability probe design to practical settings may require moving beyond rules-based dimensions, and developing principled ways to learn or elicit goal-spaces. This can emerge in chat-like interactions as studied in our text-rewriting task, where aspects of human preferences such as helpfulness, harmfulness, or honesty may not be directly

quantifiable [12]. Autonomously eliciting such goal-spaces is also potentially of interest (cf. [13]), with the goal of learning to steer towards self-imposed goals. Learning and steering in goal-space is particularly essential to applying conditional generative models to high-stakes domains, such as in protein design, where objectives are multi-faceted and may trade off with one another. While we study RL as a partial remedy for steerability in a 2D setting, steering in higher dimensional goal space among goals that may be at odds poses interesting questions for how best to navigate such tradeoffs. Understanding how to steer in such complex tradeoff landscapes is essential for functional alignment at scale.

For systemic factors, beyond oversight, a natural extension is to develop explicitly manipulation-proof models. While the connection between strategic classification and Stackelberg equilibria is well-studied, learning under unknown costs that may vary across agents remains underexplored []. Codifying sets of assumptions that make learning possible could provide a starting point for developing manipulaton-resistant models in practice.

6.3 Closing remarks

Alignment in the real world is, in some sense, a wicked problem: correct or incorrect solutions are unclear, each problem instance is unique, and there is no clear stopping rule [237]. And yet, within the complexity of alignment lies unbounded opportunity to refine how we construct ML models that align with relevant high-level values and principles. Our contributions can provide a not only technical stepping stones for improving alignment in healthcare or health policy oriented settings, but also serve as a template for alignment from the bottom-up: starting with concrete institutional or individual values, then constructing tools that illuminate or reshape the interactions between ML models and the real world.

APPENDIX A

Appendix for Disparate Censorship: A Case Study of Biases at Training Time

A.1 Supplementary material for "Racial differences in laboratory testing as a potential mechanism for bias in AI"

A.1.1 Methods

A.1.1.1 Laboratory test extraction

In MIMIC-IV, each laboratory value was stored in a separate record. For laboratory tests with a single associated value (e.g., BNP, Troponin T), the admission identifier associated with all lab events containing the relevant laboratory value were collected to determine admissions in which the relevant laboratory test was ordered. For hospital visits without an admission identifier (i.e., ED visits that did not result in admission), the chart time of the relevant lab value was cross-referenced with the start and end times of the visit, from which an ED visit identifier was associated with each relevant laboratory test. At BIDMC, laboratory tests with multiple associated values (e.g., CBC, metabolic panels), all relevant laboratory values for each test were collected and grouped by chart time, then cross-referenced with an admission identifier or ED visit identifier as described above. Laboratory test records were constructed similarly in the U-M data.

A.1.1.2 Additional data processing and matching details

For BIDMC, age information needed to be reconstructed from their shifted admission dates, since MIMIC-IV shifts timestamps for each patient for de-identification. To perform exact matching, visits with unknown or miscellaneous complaints were excluded from the matched analysis, as were visits with unknown or missing ED triage score.

Table A.1: Case-normalized free-text chief complaint for 25 most common chief complaints in the unmatched cohort at BIDMC among White and Black patients.

Complaint	White (n=244,387)	Black (n=92,437)
abd pain	9,566 (3.9)	3,565 (3.9)
chest pain	7,923 (3.2)	3,812 (4.1)
s/p fall	5,597 (2.3)	1,077 (1.2)
dyspnea	4,698 (1.9)	1,899 (2.1)
wound eval	3,217 (1.3)	1,135 (1.2)
back pain	2,726 (1.1)	1,282 (1.4)
headache	2,566 (1.1)	1,494 (1.6)
fever	2,546 (1.0)	467 (0.5)
weakness	2,045 (0.8)	620 (0.7)
syncope	1,934 (0.8)	509 (0.6)
altered mental status	1,793 (0.7)	557 (0.6)
dizziness	1,564 (0.6)	676 (0.7)
n/v	1,514 (0.6)	587 (0.6)
etoh	1,508 (0.6)	541 (0.6)
abnormal labs	1,505 (0.6)	540 (0.6)
brbpr	1,502 (0.6)	521 (0.6)
seizure	1,470 (0.6)	521 (0.6)
lower back pain	1,347 (0.6)	994 (1.1)
mvc	1,305 (0.5)	972 (1.1)
palpitations	1,241 (0.5)	440 (0.5)
abdominal pain	1,164 (0.5)	427 (0.5)
abd pain n/v	1,161 (0.5)	608 (0.7)
chest pain dyspnea	1,077 (0.4)	563 (0.6)
sore throat	996 (0.4)	681 (0.7)
rash	961 (0.4)	654 (0.7)

A.1.2 Supplementary results

A.1.2.1 Sensitivity Analysis (Alternative Matching Strategy)

Since ED triage score is potentially a collider with respect to race and other confounders, we run experiments with an alternate matching strategy, performing 1:1 exact matching on race, biological sex, and chief complaint (i.e., excluding ED triage score with respect to the main paper). This yields a population of 109,656 and 154,384 ED visits at BIDMC and U-M, respectively. This represents 59.3% of Black patients and 22.4% of White patients at BIDMC, and 75.5% of Black patients and 17.6% of White patients at U-M. Overall trends were similar to those identified in the primary analysis. However, Black patients were significantly less likely than White patients to receive a blood culture (BIDMC: 10.7% vs. 9.4%, difference: 1.2%, 95% CI: 0.9% to 1.6%, $p<.001$) at both

Table A.2: Case-normalized free-text chief complaint for 25 most common chief complaints in the unmatched cohort at U-M among White and Black patients.

Complaint	White (n=439,005)	Black (n=102,269)
chest pain	22,140 (5.0)	5,825 (5.7)
shortness of breath	18,186 (4.1)	4,025 (3.9)
abdominal pain - generalized	15,173 (3.5)	3,532 (3.5)
back pain	8,643 (2.0)	2,102 (2.1)
fall	8,590 (2.0)	936 (0.9)
alcohol intoxication	6,913 (1.6)	2,482 (2.4)
altered mental status	7,212 (1.6)	1,194 (1.2)
wound check	6,860 (1.6)	1,457 (1.4)
abnormal lab	6,673 (1.5)	1,409 (1.4)
flank pain	6,975 (1.6)	995 (1.0)
vomiting	6,066 (1.4)	1,661 (1.6)
headache	5,974 (1.4)	1,736 (1.7)
leg pain	4,817 (1.1)	1,433 (1.4)
fever	5,382 (1.2)	774 (0.8)
dizziness	4,884 (1.1)	1,119 (1.1)
fatigue	4,525 (1.0)	989 (1.0)
eye problem	4,266 (1.0)	767 (0.7)
motor vehicle crash	3,577 (0.8)	994 (1.0)
syncope	3,789 (0.9)	686 (0.7)
seizures	3,317 (0.8)	891 (0.9)
abdominal pain - lower	3,277 (0.7)	903 (0.9)
abdominal pain - right lower quadrant	3,526 (0.8)	471 (0.5)
cough	2,556 (0.6)	949 (0.9)
postop problem	2,999 (0.7)	462 (0.5)
palpitations	3,015 (0.7)	435 (0.4)

institutions (instead of U-M only). Trends are similar at U-M (Figure A.1). We note that, under this matching strategy, White patients remained significantly more likely to be admitted following the ED visit at both institutions (BIDMC: 38.4% vs. 34.9%, difference: 3.5%, 95% CI: 2.9% to 4.1%, p<.001). Trends in hospital admission rates are similar at U-M (Figure A.2). Full results are reported in Figure A.1 and Tables A.3 & A.4.

A.1.2.2 Intersectional Analysis

Black male patients were significantly less likely than White male patients to receive a CBC (BIDMC: 60.2% vs. 57.9%, difference 2.3%, 95% CI: 1.2% to 3.3%, p<.001) or metabolic panel (BIDMC: 60.8% vs. 58.9%, difference 1.9%, 95% CI: 0.9% to 2.9%, p<.001). Similarly, White

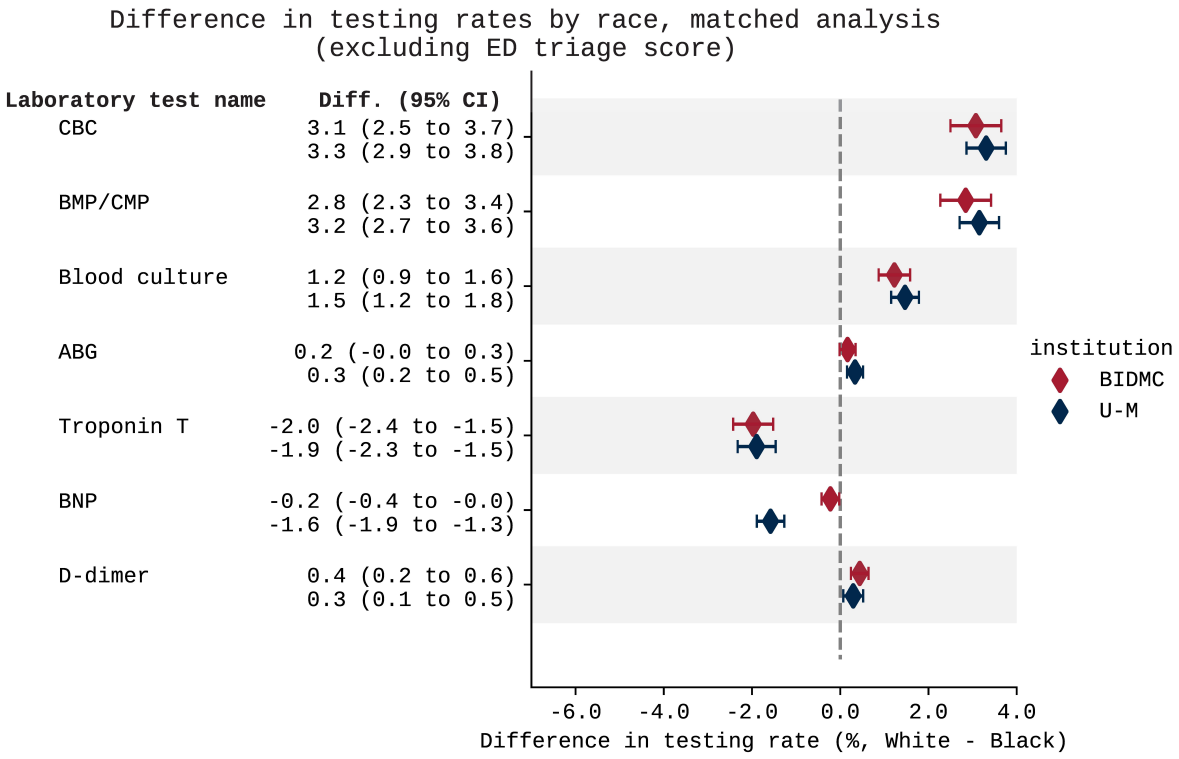


Figure A.1: Sensitivity analysis of matching strategy showing differences in testing rates between White and Black patients matched on age, sex, and chief complaint only.

patients were significantly less likely than Black patients to receive a troponin (BIDMC: 18.6% vs. 20.3%, difference -1.7%, 95% CI: -2.5% to -0.8%, p<.001) test. Trends are similar in female patients and at U-M (Figure A.3). In contrast to the primary analysis, after disaggregating by biological sex, there was no significant difference in blood culture or BNP rates across race in both male and female patients. Note that we use a Bonferroni correction factor of 28 for this disaggregated analysis. Full results are reported in Figure A.3 and Tables A.9 & A.10.

A.1.2.3 Subgroup analysis with respect to admission status

In the subgroup analysis of visits resulting in discharge, matching on age, sex, ED triage score, and chief complaint yielded a population of 57,062 and 95,880 ED visits at BIDMC and U-M, respectively. At BIDMC, this matched subgroup included 48.2% and 23.8% of all visits resulting in discharge by Black and White patients, respectively. At U-M, this matched subgroup included 66.8% and 17.6% of visits resulting in discharge by Black and White patients, respectively. Within

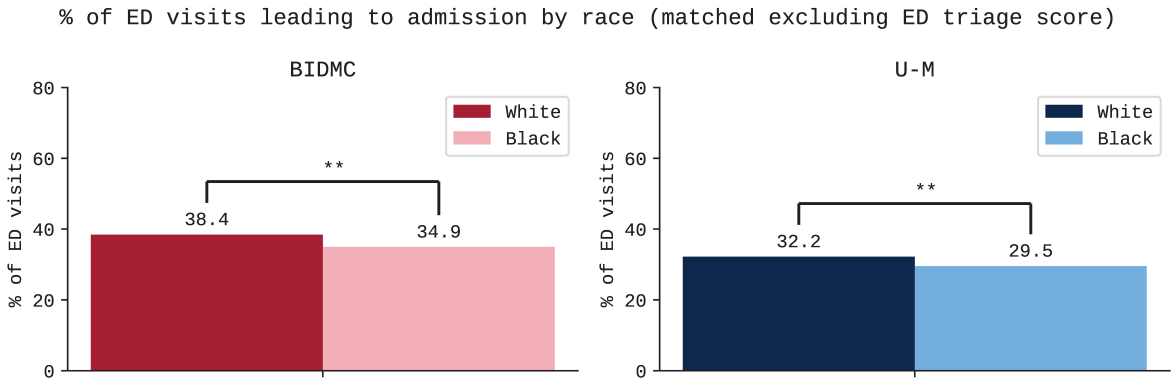


Figure A.2: Differences in rate of admission to hospital between White and Black patients matched on age, sex, and chief complaint.

such visits, White patients were significantly more likely than Black patients to receive a blood culture (BIDMC: 6.8% vs. 5.7%, difference 1.1%, 95% CI: 0.8% to 1.5%, p<.001) or d-dimer (BIDMC: 3.8% vs. 3.1%, difference 0.7%, 95% CI: 0.4% to 1.0%, p<.001). Trends were similar at U-M (Figure 4, top). See Tables A.13 and A.14 (Supplement) for the testing rates shown in Figure A.4 (top).

Among patients admitted to the hospital, matching on age, sex, ED triage score, and chief complaint yielded a population of 28,710 and 37,368 ED visits at BIDMC and U-M, respectively. At both institutions (Figure 4, bottom), White patients were significantly less likely than Black patients to receive a troponin test (BIDMC: 35.1% vs. 38.9%, difference -3.8%, 95% CI: -5.8% to -1.9%, p<.001). The remaining testing differences in this subgroup were not statistically significant at both institutions. See Tables A.15 and A.16 (Supplement) for the testing rates shown in Figure A.4 (bottom).

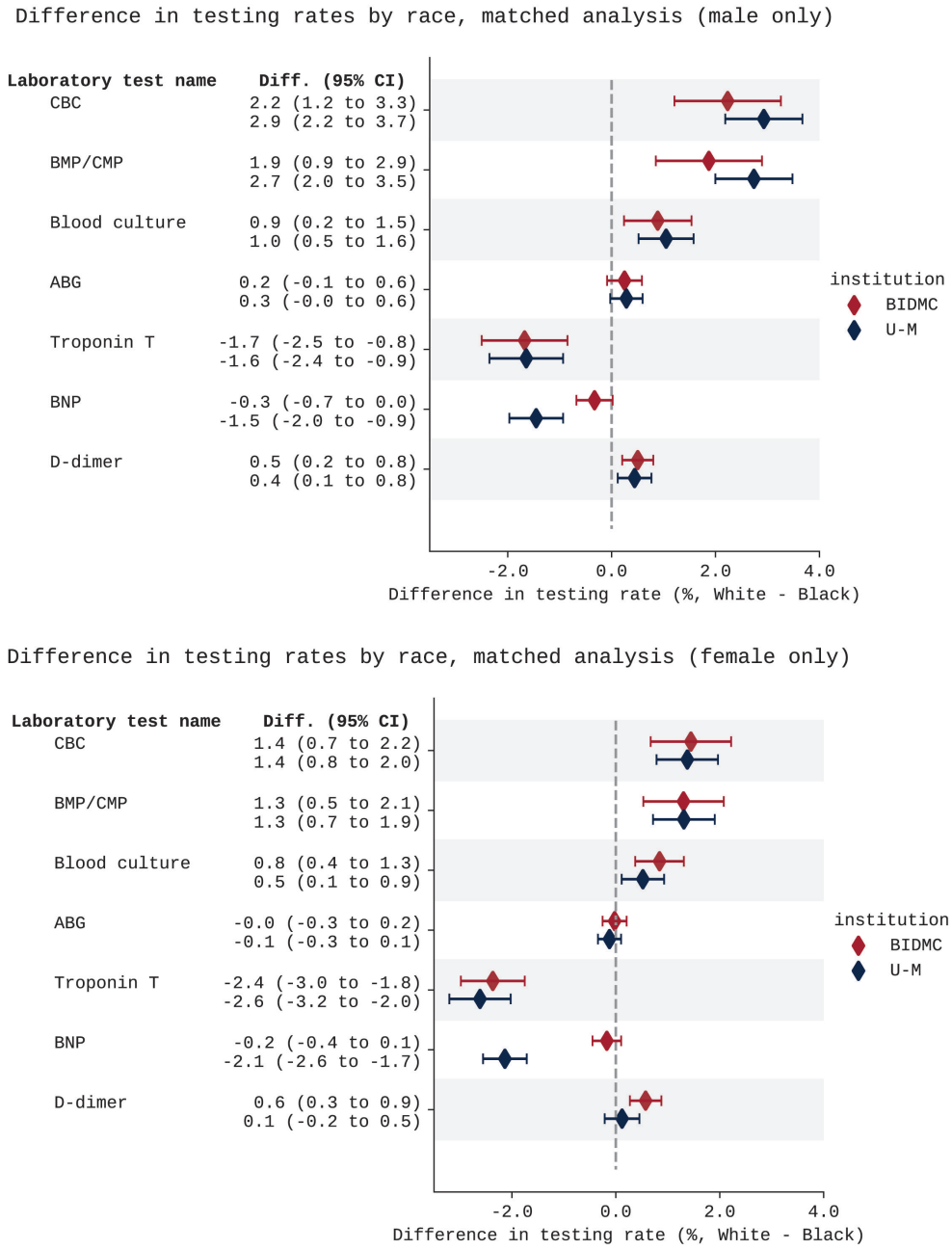


Figure A.3: Differences in testing rates between White and Black patients matched on age, biological sex, chief complaint, and ED triage score, stratified by male (top) and female (bottom) patients.

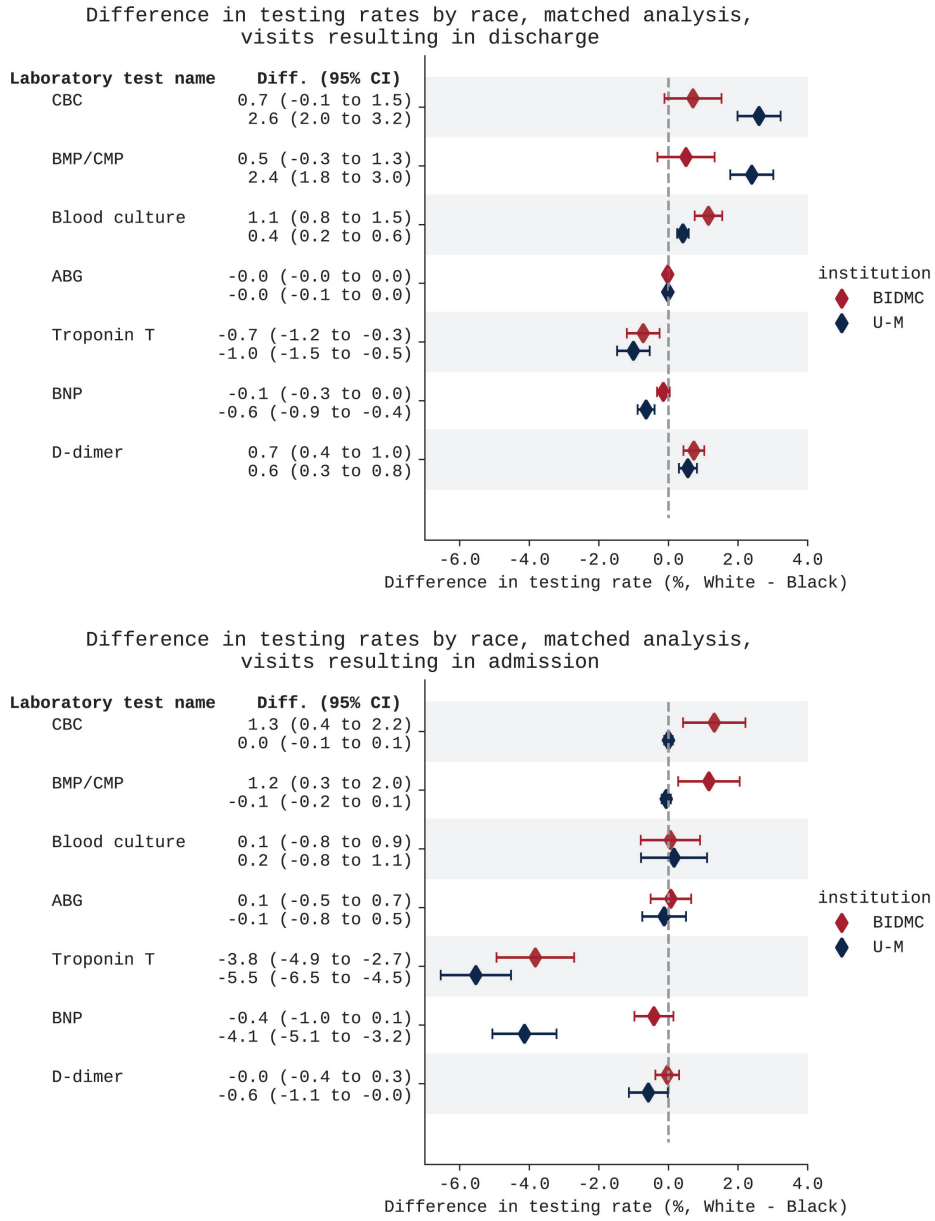


Figure A.4: Subgroup analysis of testing differences by admission status (discharge vs. admission) between White and Black patients matched on age, sex, chief complaint, and ED triage score.

Table A.3: Full results for differences in testing rates at BIDMC between White and Black patients matched on age, sex, chief complaint, and ED triage score.

Test	White (n=47,160)	Black (n=47,160)	P value
Complete blood count	29,724 (63.0)	28,903 (61.3)	<.001
Metabolic panel	29,823 (63.2)	29,108 (61.7)	<.001
Blood culture	4,870 (10.3)	4,465 (9.5)	<.001
Arterial blood gas	1,103 (2.3)	1,066 (2.3)	.42
Troponin T	8,119 (17.2)	9,113 (19.3)	<.001
Brain natriuretic peptide	1,342 (2.8)	1,451 (3.1)	.04
D-dimer	1,573 (3.3)	1,315 (2.8)	<.001

Table A.4: Full results for differences in testing rates at U-M between White and Black patients matched on age, sex, chief complaint, and ED triage score.

Test	White (n=70,755)	Black (n=70,755)	P value
Complete blood count	52,072 (73.6)	50,635 (71.6)	<.001
Metabolic panel	51,914 (73.4)	50,561 (71.5)	<.001
Blood culture	8,019 (11.3)	7,493 (10.6)	<.001
Arterial blood gas	2,298 (3.3)	2,263 (3.2)	.61
Troponin T	17,186 (24.3)	18,746 (26.5)	<.001
Brain natriuretic peptide	7,174 (10.1)	8,481 (12.0)	<.001
D-dimer	3,965 (5.6)	3,784 (5.4)	.04

Table A.5: Testing rates at BIDMC, White and Black pateitns.

Test	White (n=244,387)	Black (n=92,437)	P value
Complete blood count	170,081 (69.6)	55,890 (60.5)	<.001
Metabolic panel	170,482 (69.8)	56,198 (60.8)	<.001
Blood culture	34,024 (13.9)	9,273 (10.0)	<.001
Arterial blood gas	10,248 (4.2)	2,297 (2.5)	<.001
Troponin T	46,018 (18.8)	16,843 (18.2)	<.001
Brain natriuretic peptide	9,105 (3.7)	2,900 (3.1)	<.001
D-dimer	5,148 (2.1)	2,411 (2.6)	<.001

Table A.6: Testing rates at U-M, Black and White patients.

Test	White (n=439,005)	Black (n=102,269)	P value
Complete blood count	335,263 (76.4)	72,078 (70.5)	$<$.001
Metabolic panel	334,946 (76.3)	71,990 (70.4)	$<$.001
Blood culture	65,038 (14.8)	11,083 (10.8)	$<$.001
Arterial blood gas	19,939 (4.5)	3,318 (3.2)	$<$.001
Troponin T	117,086 (26.7)	25,820 (25.3)	$<$.001
Brain natriuretic peptide	54,645 (12.5)	11,694 (11.4)	$<$.001
D-dimer	21,332 (4.9)	5,090 (5.0)	.12

Table A.7: Cohort demographics and subsequent hospital admission rate for BIDMC among all ED visits after matching on age, biological sex, chief complaint, and ED triage score.

	White (n=47,160)	Black (n=47,160)	P value
Median age, years (IQR)	46 (30 - 61)	46 (30 - 61)	N/A
Sex (% Female)	29,481 (62.5)	29,481 (62.5)	N/A
ED triage score (%)[a]			
1	1,342 (2.9)	1,342 (2.9)	
2	11,800 (25.0)	11,800 (25.0)	
3	30,397 (64.5)	30,397 (64.5)	N/A
4	3,579 (7.6)	3,579 (7.6)	
5	42 (0.1)	42 (0.1)	
Mean ED triage score (SD)	2.77 (0.63)	2.77 (0.63)	N/A
Admitted to hospital (%)	17,775 (37.7)	16,523 (35.0)	$<$.001

[a] Percentages may not sum to 100.0% due to rounding. Null or missing ED triage scores are removed prior to matching.

Table A.8: Cohort demographics and subsequent hospital admission rate for U-M among all ED visits after matching on age, biological sex, chief complaint, and ED triage score.

	White (n=70,755)	Black (n=70,755)	P value
Median age, years (IQR)	44 (30 - 59)	44 (30 - 59)	N/A
Sex (% Female)	40,929 (57.8)	40,929 (57.8)	N/A
ED triage score (%)[a]			
1	435 (0.6)	435 (0.6)	
2	24,785 (35.0)	24,785 (35.0)	
3	36,352 (51.4)	36,352 (51.4)	N/A
4	8,765 (12.4)	8,765 (12.4)	
5	418 (0.6)	418 (0.6)	
Mean ED triage score (SD)	2.77 (0.69)	2.77 (0.69)	N/A
Admitted to hospital (%)	22,172 (31.3)	21,072 (29.8)	<.001

[a] Percentages may not sum to 100.0% due to rounding. Null or missing ED triage scores are removed prior to matching.

Table A.9: Differences in testing rates at BIDMC between White and Black patients matched on age, sex, and chief complaint, all patients.

Test	White (n=54,744)	Black (n=54,744)	P value
Complete blood count	34,598 (63.2)	32,915 (60.1)	<.001
Metabolic panel	34,699 (63.4)	33,142 (60.5)	<.001
Blood culture	5,845 (10.7)	5,173 (9.4)	<.001
Arterial blood gas	1,368 (2.5)	1,277 (2.3)	.07
Troponin T	9,243 (16.9)	10,323 (18.9)	<.001
Brain natriuretic peptide	1,554 (2.8)	1,676 (3.1)	.03
D-dimer	1,720 (3.1)	1,478 (2.7)	<.001

Table A.10: Differences in testing rates at U-M between White and Black patients matched on age, sex, and chief complaint, all patients.

Test	White (n=77,417)	Black (n=77,417)	P value
Complete blood count	57,204 (73.9)	54,643 (70.6)	<.001
Metabolic panel	57,023 (73.7)	54,582 (70.5)	<.001
Blood culture	9,330 (12.1)	8,193 (10.6)	<.001
Arterial blood gas	2,813 (3.6)	2,553 (3.3)	<.001
Troponin T	18,585 (24.0)	20,051 (25.9)	<.001
Brain natriuretic peptide	7,866 (10.2)	9,089 (11.7)	<.001
D-dimer	4,281 (5.5)	4,053 (5.2)	.01

Table A.11: Differences in testing rates at BIDMC between White and Black patients matched on age, sex, chief complaint, and ED triage score among all patients, by biological sex.

	Female			Male		
Test	W[a](n=29,481)	B[b](n=29,481)	P	W (n=17,679)	B (n=17,679)	P
CBC[c]	19,086 (64.7)	18,660 (63.3)	<.001	10,638 (60.2)	10,243 (57.9)	<.001
MP[d]	19,076 (64.7)	18,692 (63.4)	<.001	10,747 (60.8)	10,416 (58.9)	<.001
BCx[e]	2,860 (9.7)	2,612 (8.9)	<.001	2,010 (11.4)	1,853 (10.5)	.007
ABG[f]	612 (2.1)	619 (2.1)	.84	491 (2.8)	447 (2.5)	.15
Troponin T	4,826 (16.4)	5,524 (18.7)	<.001	3,293 (18.6)	3,589 (20.3)	<.001
BNP[g]	860 (2.9)	911 (3.1)	.22	482 (2.7)	540 (3.1)	.07
D-dimer	1,158 (3.9)	989 (3.4)	<.001	415 (2.3)	326 (1.8)	<.001

[a] W: White
[b] B: Black
[c] CBC: Complete blood count
[d] MP: Metabolic panel
[e] BCx: Blood culture
[f] ABG: Arterial blood gas
[g] BNP: Brain natriuretic peptide

Table A.12: Differences in testing rates at U-M between White and Black patients matched on age, sex, chief complaint, and ED triage score among all patients, by biological sex.

	Female			Male		
Test	W[a](n=40,929)	B[b](n=40,929)	P	W (n=29,826)	B (n=29,826)	P
CBC[c]	31,092 (76.0)	30,529 (74.6)	<.001	20,980 (70.3)	20,106 (67.4)	<.001
MP[d]	30,909 (75.5)	30,373 (74.2)	<.001	21,005 (70.4)	20,188 (67.7)	<.001
BCx[e]	4,144 (10.1)	3,931 (9.6)	.01	3,875 (13.0)	3,562 (11.9)	<.001
ABG[f]	1,091 (2.7)	1,141 (2.8)	.29	1,207 (4.1)	1,122 (3.8)	.08
Troponin T	9,521 (23.3)	10,591 (25.9)	<.001	7,665 (25.7)	8,155 (27.3)	<.001
BNP[g]	3,880 (9.5)	4,754 (11.6)	<.001	3,294 (11.0)	3,727 (12.5)	<.001
D-dimer	2,627 (6.4)	2,578 (6.3)	.49	1,338 (4.5)	1,206 (4.0)	.008

[a] W: White
[b] B: Black
[c] CBC: Complete blood count
[d] MP: Metabolic panel
[e] BCx: Blood culture
[f] ABG: Arterial blood gas
[g] BNP: Brain natriuretic peptide

Table A.13: Difference in testing rates at BIDMC between White and Black patients in ED visits resulting in discharge, matched on age, sex, chief complaint, and ED triage score.

Test	White (n=28,531)	Black (n=28,531)	P value
Complete blood count	14,515 (50.9)	14,314 (50.2)	.09
Metabolic panel	14,543 (51.0)	14,399 (50.5)	.23
Blood culture	1,945 (6.8)	1,617 (5.7)	$<$.001
Arterial blood gas	5 (0.0)	11 (0.0)	.13
Troponin T	2,493 (8.7)	2,700 (9.5)	.003
Brain natriuretic peptide	337 (1.2)	379 (1.3)	.11
D-dimer	1,090 (3.8)	882 (3.1)	$<$.001

Table A.14: Difference in testing rates at U-M between White and Black patients in ED visits resulting in discharge, matched on age, sex, chief complaint, and ED triage score.

Test	White (n=47,940)	Black (n=47,940)	P value
Complete blood count	29,623 (61.8)	28,375 (59.2)	$<$.001
Metabolic panel	29,456 (61.4)	28,308 (59.1)	$<$.001
Blood culture	945 (2.0)	746 (1.6)	$<$.001
Arterial blood gas	62 (0.1)	71 (0.2)	.49
Troponin T	7,612 (15.9)	8,095 (16.9)	$<$.001
Brain natriuretic peptide	1,692 (3.5)	2,000 (4.2)	$<$.001
D-dimer	2,224 (4.6)	1,956 (4.1)	$<$.001

Table A.15: Difference in testing rates at BIDMC between White and Black patients in ED visits resulting in admission, matched on age, sex, chief complaint, and ED triage score.

Test	White (n=14,355)	Black (n=14,355)	P value
Complete blood count	11,802 (82.2)	11,613 (80.9)	.004
Metabolic panel	11,894 (82.9)	11,727 (81.7)	.009
Blood culture	2,334 (16.3)	2,326 (16.2)	.90
Arterial blood gas	979 (6.8)	969 (6.8)	.81
Troponin T	5,041 (35.1)	5,590 (38.9)	$<$.001
Brain natriuretic peptide	871 (6.1)	931 (6.5)	.14
D-dimer	314 (2.2)	319 (2.2)	.84

Table A.16: Difference in testing rates at U-M between White and Black patients in ED visits resulting in admission, matched on age, sex, chief complaint, and ED triage score.

Test	White (n=18,684)	Black (n=18,684)	P value
Complete blood count	18,617 (99.6)	18,617 (99.6)	1
Metabolic panel	18,599 (99.6)	18,611 (99.6)	.38
Blood culture	6,063 (32.5)	6,033 (32.3)	.75
Arterial blood gas	1,996 (10.7)	2,019 (10.8)	.71
Troponin T	8,358 (44.7)	9,392 (50.3)	$<$.001
Brain natriuretic peptide	5,093 (27.3)	5,866 (31.4)	$<$.001
D-dimer	1,497 (8.0)	1,605 (8.6)	.04

A.2 Supplementary material for "Disparate Censorship and Undertesting: A Source of Label Bias in Clinical Machine Learning"

A.2.1 Proofs

A.2.1.1 Preliminaries

First, we restate the definition of a boundary-consistent noise (BCN) model from [191], which is useful for our proofs.

Definition 15 (Boundary-consistent noise (BCN) model.)**.** *Define class probability function $\eta(\mathbf{x}) = P(Y = 1 \mid \mathbf{x})$. Consider a data generating process in which $(\mathbf{x}, \tilde{y})$ is generated by drawing an instance $(\mathbf{x}, Y)$ and flipping Y with instance- and label-dependent probability $\rho_Y(\mathbf{x})$. Suppose that label flip-probability functions ρ_y can be written in the form $\rho_y = f_y \circ s$, where $f_y : \mathbb{R} \to [0, 1]$ for $y \in \{0, 1\}$ and $s : \mathcal{X} \to \mathbb{R}$, and $\rho_0(\mathbf{x}) + \rho_1(\mathbf{x}) < 1$ for all $\mathbf{x}$. Then, a noise model (f_0, f_1, s, η) is BCN-admissible if the following conditions are satisfied:*

- ***Feasible ranking:*** *s is order-preserving in $\mathbf{x}$ for η; that is, for any $(\mathbf{x}, \mathbf{x}') \in \mathcal{X}$, then $\eta(\mathbf{x}) < \eta(\mathbf{x}')$ implies $s(\mathbf{x}) < s(\mathbf{x}')$.*
- ***Piecewise-monotonicity:*** *f_0 and f_1 are non-decreasing where $\eta \leq \frac{1}{2}$, and non-increasing otherwise.*
- ***Flip-probability monotonicity:*** *$f_1(z) - f_0(z)$ is non-increasing in z.*

Note that, for our setting, since f_0 is constant, we can combine the two monotonicity constraints: it is sufficient that f_1 is non-increasing in z. Furthermore, we restate an important property of the BCN model shown in [191] with some notation adapted to our setting:

Theorem 16 (Theorem 2 of [191].)**.** *Pick any distribution D. Let $\bar{D}$ be a corrupted distribution. Suppose that for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, $\eta(\mathbf{x}) < \eta(\mathbf{x}')$ implies $\tilde{\eta}(\mathbf{x}) < \tilde{\eta}(\mathbf{x}')$, where $\tilde{\eta}$ is the analogue of η on noisy labels $\tilde{y}$ (i.e. $\tilde{\eta}(\mathbf{x}) - P(\tilde{Y} = 1 \mid X = \mathbf{x})$), and there exists a constant C such that $|\eta(\mathbf{x}) < \eta(\mathbf{x}')| \leq C \cdot |\tilde{\eta}(\mathbf{x}) < \tilde{\eta}(\mathbf{x}')|$. Then, for any scorer s,*

$$reg_{rank}(s; D) \leq C \cdot \frac{\bar{\pi}(1 - \bar{\pi})}{\pi(1 - \pi)} \cdot reg_{rank}(s; \bar{D}) \tag{A.1}$$

where reg_{rank} is the excess ranking risk of a scorer s, and $\pi = P(Y = 1), \bar{\pi} = P(\tilde{Y} = 1)$. In particular, if $\bar{D} = BCN(D, f_0, f_1, \eta)$ where (f_0, f_1, s, η) are BCN-admissible, then $C = (1 - 2 \cdot \rho_{max}) - 1$, where $\rho_{max} = \sup_{\mathbf{x} \in \mathcal{X}} P(\tilde{Y} = 0 \mid Y = 1, X = \mathbf{x})$ is sufficient.

We defer to [191] for the proof. This states that optimizing a model for AUC on noisy labels in a BCN model is consistent with optimizing a model for AUC on clean labels: both converge to the same Bayes-optimal scorer.

Lastly, we prove a Lemma relating the Bayes-optimal overall AUC to the Bayes-optimal within-group AUC and xAUC under perfect separability.

Lemma 17 (No ranking performance gap under perfect separation.)**.** *Let $\eta(\mathbf{x}) = P(Y = 1 \mid X = \mathbf{x})$. If Y is perfectly separable in $\mathbf{x}$; that is, there exists some $s : \mathcal{X} \to \{0, 1\}$ such that $s(\mathbf{x}) = y$ for all $\mathbf{x} \in \mathcal{X}$, and $P(X \mid Y = y, A = a) > 0$ for any $y, a \in \{0, 1\}$, then $\Delta AUC, \Delta xAUC = 0$.*

Proof. Under perfect separation, for the Bayes-optimal ranker η, we know that $P(\eta(\mathbf{x}) < \eta(\mathbf{x}')) = 1$ where $\mathbf{x} \in \{x \mid x \in S, Y = 1\}$, $\mathbf{x}' \in \{x \mid x \in S', Y = 0\}$ for any $S, S' \subseteq \mathcal{X}$.

First, we show that $\Delta\text{AUC} = 0$. Choose $S = \{x \mid x \in \mathcal{X}, Y = 1, A = a\}$ and $S' = \}x \mid x \in \mathcal{X}, Y = 0, A = a\}$ for arbitrary $a \in \{0, 1\}$. Then the Bayes-optimal within-group AUC is 1 for each a, so $\Delta\text{AUC} = 0$ as required. Now, we show that $\Delta\text{xAUC} = 0$. Choose $S = \{x \mid x \in \mathcal{X}, Y = 1, A = 1\}$ and $S' = \{x \mid x \in \mathcal{X}, Y = 0, A = 0\}$ (without loss of generality in assignment of A). Then the Bayes-optimal xAUC is 1 for each assignment of A, so $\Delta\text{xAUC} = 0$. We have shown that $\Delta\text{AUC}, \Delta\text{xAUC} = 0$, concluding the proof. □

A.2.1.2 Proof of Theorem 3

Proof. First, we show feasible ranking holds. As $y \in \{0, 1\}$, $\eta(\mathbf{x}) < \eta(\mathbf{x}')$ implies that $s(\mathbf{x}) \leq b$ and $s(\mathbf{x}') > b$, from which $s(\mathbf{x}) < s(\mathbf{x}')$ follows as required.

Now, define $\eta(\mathbf{x}) = P(Y = 1 \mid X = \mathbf{x})$, and $f_1(z) = P(\tilde{Y} = 0 \mid Y = 1, Z = z)$ for $z = s(\mathbf{x})$, and p_a as $P(A = 0)$. We show that f_1 is monotonically non-increasing on the interval where $\eta > 1/2$ if either no positives are tested or $\tau_0 \geq \tau_1$. If no positives are tested, then the theorem is vacuously true. Otherwise, we first define $b = \inf_{\mathbf{x} \in \mathcal{X}: y=1} (s(\mathbf{x}))$. In other words, b is the threshold that perfectly separates negative from positives examples. Then, choose some $\tau_1 > b$, and some $\tau_0 \geq \tau_1$. We can write f_1, which is the probability that a positive label is flipped, as

$$f_1(z) = \begin{cases} 1 - c & z < \tau_1 \\ \frac{p_a \cdot \exp\left(-\frac{(z-\mu_0)^2}{2\sigma^2}\right)}{p_a \cdot \exp\left(-\frac{(z-\mu_0)^2}{2\sigma^2}\right) + (1-p_a) \cdot (1-c) \cdot \exp\left(-\frac{(z-\mu_1)^2}{2\sigma^2}\right)} & \tau_1 \leq z < \tau_0 \\ 0 & z \geq \tau_0 \end{cases}. \tag{A.2}$$

This is clearly non-increasing on $(-\infty, \tau_1)$, $[\tau_0, \infty)$, so it suffices to show that $f_1(z)$ is non-increasing on $[\tau_1, \tau_0)]$, and that $f_1(\tau_1) \leq 1 - c, f_1(\tau_0) \geq 0$. Note that the portion of f_1 for $\tau_0 \leq z < \tau_1$ is simply $Pr[A = 1]/(Pr[A = 0] + Pr[A = 1])$, since group $A = 1$ is censored with

probability $1-c$ and group $A=0$ is not censored in that region. Denote this function as $a(z)$.[1] We rewrite the portion of f_1 for $\tau_1 \leq z < \tau_0$ as a sigmoid function:

$$a(z) = \frac{p_a \cdot \exp\left(-\frac{(z-\mu_0)^2}{2\sigma^2}\right)}{p_a \cdot \exp\left(-\frac{(z-\mu_0)^2}{2\sigma^2}\right) + (1-p_a)\cdot(1-c)\cdot\exp\left(-\frac{(z-\mu_1)^2}{2\sigma^2}\right)} \tag{A.3}$$

$$= \frac{1}{1+\exp\left(\log\frac{(1-p_a)(1-c)}{p_a} + \frac{(z-\mu_0)^2-(z-\mu_1)^2}{2\sigma^2}\right)} \tag{A.4}$$

$$= \frac{1}{1+\exp\left(\log\frac{(1-p_a)(1-c)}{p_a} + \frac{1}{2\sigma^2}\left[(2z-\mu_0-\mu_1)(\mu_1-\mu_0)\right]\right)}. \tag{A.5}$$

We further rewrite Eq. A.5 in the form $\sigma(g(z))$, where

$$g(z) = \log\frac{p_a}{(1-p_a)(1-c)} + \frac{(2z-\mu_0-\mu_1)(\mu_0-\mu_1)}{2\sigma^2}. \tag{A.6}$$

Then, taking derivatives:

$$\frac{d}{dz}a(z) = \sigma(g(z))(1-\sigma(g(z)))\frac{d}{dz}g(z) = \sigma(g(z))(1-\sigma(g(z)))\left((\mu_0-\mu_1)/2\sigma^2\right) \leq 0, \tag{A.7}$$

where the final inequality follows since $\mu_1 \geq \mu_0$ by assumption. Thus, $a(z)$ is monotonically non-increasing, which is what we wanted to show. To conclude, since $g:\mathbb{R}\to\mathbb{R}$ and $\sigma:\mathbb{R}\to(0,1)$, clearly $f_1(\tau_0) \geq 0$. Furthermore, rewriting the constraint $f_1(z) = \sigma(g(\tau_1)) \leq 1-c$ and simplifying yields

$$\tau_1 \leq \log\frac{(1-c)^2(1-p_a)}{c\cdot p_a}\cdot\frac{2\sigma^2}{\mu_0-\mu_1} + \frac{\mu_0+\mu_1}{2}, \tag{A.8}$$

i.e., some negative offset of the midpoint between the group-wise means μ_0, μ_1. So $f_1(\tau_1) \leq 1-c$ for such choices of τ_1. Thus, (f_0, f_1, s, η) is BCN admissible. Applying Theorem 5 and Lemma 2 concludes the proof. □

Remark 18. *Our simulation yields a risk score distribution in random variable $s_\alpha(\mathbf{x})$. Assuming the effect of clipping with respect to range $[0,1]$ is negligible, which is true for μ_a near 0.5, the distribution $s_\alpha(\mathbf{x})$ is approximately univariate Gaussian, allowing the application of Theorem 3. This is because the sum of independent Gaussians is Gaussian, so by the definition of $s_\alpha(\mathbf{x})$, the distribution of scores is approximately a discretized univariate Gaussian distribution.*

Remark 19. *We can apply the same argument in this proof to risk distributions beyond homoscedastic Gaussians, based on whether $Pr[A=0]/(Pr[A=0]+Pr[A=1])$ is non-decreasing on $[\tau_1, \tau_0)$.*

[1] If $\tau_1 \leq b$ instead, then only the portion of f_1 where $f_1(z)=0$ is observed, which is trivially non-increasing.

Let $R_0(z), R_1(z)$ be the probability density functions of risk scores z for group $a = 0, a = 1$, respectively. Clearly, if $R_0(z)/[R_0(z) + R_1(z)]$ is non-decreasing on $[\tau_1, \tau_0)$, that would satisfy boundary consistency. Alternately, applying quotient rule shows that $R_0'(z) \cdot R_1(z) - R_0(z) \cdot R_1'(z) > 0$ is also sufficient to violate boundary consistency.

A.2.1.3 Proof of Theorem 4

Proof. We prove that each BCN condition is satisfied. First, we show that feasible ranking holds. Choose any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Recall that $t = \mathbb{1}[\boldsymbol{\theta}^\top \mathbf{x} + \beta > 0 \vee p = 1]$ where $p \sim Bernoulli(c)$ and $y = \mathbb{1}[s_a(\mathbf{x}) > 0]$, where $s_a(\mathbf{x}) = \boldsymbol{\theta}_a^\top \mathbf{x} + b_a$. Since $y \in \{0, 1\}$, $\eta(\mathbf{x}) < \eta(\mathbf{x}')$ implies that $s_a(\mathbf{x}) \leq 0$ and $s_a(\mathbf{x}') > 0$, from which $s_a(\mathbf{x}) < s_a(\mathbf{x}')$ follows, as required.

Next, we show that the piecewise-monotonicity constraint holds. To do so, we need to show that the flip probability function f_1 is non-increasing where $\eta(\mathbf{x}) \geq 1/2$, or on $s(x) = \theta_a^\top x + b_a > 0$ for all a. Consider an arbitrary group a with corresponding θ_a, b_a. We can write f_1 as

$$f_1(\cdot) = \begin{cases} c & \boldsymbol{\theta}^\top \mathbf{x} + \beta \leq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{A.9}$$

Now, suppose that there exists some $\delta \in \mathbb{R}, \delta > 0$ such that $\boldsymbol{\theta} = \boldsymbol{\theta}_a \delta$. Choose any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ such that $0 < \boldsymbol{\theta}_a^\top \mathbf{x} + b_a < \boldsymbol{\theta}_a^\top \mathbf{x}' + b_a$.[2] Then:

$$\boldsymbol{\theta}_a^\top (\mathbf{x} - \mathbf{x}') > 0 \iff \delta \boldsymbol{\theta}_a^\top (\mathbf{x} - \mathbf{x}') > 0 \tag{A.10}$$

$$\iff \boldsymbol{\theta}^\top (\mathbf{x} - \mathbf{x}') > 0 \tag{A.11}$$

$$\iff \boldsymbol{\theta}^\top \mathbf{x} + \beta > \boldsymbol{\theta}^\top \mathbf{x}' + \beta. \tag{A.12}$$

Let $L = \boldsymbol{\theta}^\top \mathbf{x} + \beta, R = \boldsymbol{\theta}^\top \mathbf{x}' + \beta$. There are three possible value-pairs in the final inequality: (1) $L < 0, R < 0$, (2) $L < 0, R \geq 0$, and (3) $L \geq 0, R \geq 0$. We need to show that, in each setting, $f(L) \geq f(R)$. For the first case, $L < 0, R < 0$ implies that $f_1(L) = f_0(L) = c$, so $f(L) \geq f(R)$ clearly. For the second case, $L < 0, R \geq 0$ implies that $f_1(L) = c, f_0(L) = 0$, so $f(L) > f(R)$, from which $f(L) \geq f(R)$ is assured. Lastly, the third case is identical to the first as $f_1(L) = f_0(L) = 0$ if both $L, R \geq 0$. Thus, piecewise-monotonicity is satisfied.

Lastly, as f_0 is the zero function, flip-probability monotonicity follows for free from the preceding. As all three BCN conditions are satisfied, and the choice of a was arbitrary, (f_0, f_1, s, η) is BCN-admissible, which is what we wanted to show. Applying Theorem 16, Lemma 17 concludes the proof. □

[2]Note that, by the definition of y (which is $\eta(\cdot)$ in this setting), we only need to (and can only) show monotonicity for $\mathbf{x}$ such that $\theta_a^\top \mathbf{x} + b_a > 0$.

Remark 20. *The generalization to non-linear decision boundaries follows naturally from using reproducing kernel Hilbert spaces $\phi(\mathbf{x}), \phi(\mathbf{x}') \in \mathcal{H}$ where the kernel for $\mathcal{H}$, $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ corresponds to a feature map $\phi : \mathcal{X} \to \mathcal{H}$ (i.e. $\phi(\mathbf{x})^\top \phi(\mathbf{x}') = K(\mathbf{x}, \mathbf{x}')$). Then, since a Hilbert space is a complete metric space (i.e. inner product is well-defined), the proof proceeds identically.*

Remark 21. *The generalization to s of the form $s_a(\mathbf{x}) = \boldsymbol{\theta}_a^\top g(\mathbf{x}) + \beta$ where g is element-wise monotonic non-decreasing is also straightforward. We say that a function $g : \mathbb{R}^d \to \mathbb{R}$ is element-wise monotonic non-decreasing if for any $\mathbf{x} \prec \mathbf{x}'$, $g(\mathbf{x}) \succeq g(\mathbf{x}')$, where $\prec, \succeq$ are element-wise inequality operators. Then, substituting $g(\mathbf{x}), g(\mathbf{x}')$ for $\mathbf{x}, \mathbf{x}'$ in the proof, in Eq. A.12, since g is element-wise monotonic non-decreasing, then $\boldsymbol{\theta}^\top g(\mathbf{x}) > \boldsymbol{\theta}^\top g(\mathbf{x}')$ implies $\boldsymbol{\theta}^\top \mathbf{x} > \boldsymbol{\theta}^\top \mathbf{x}'$, from which the proof proceeds identically.*

A.2.2 Simulation Study Details

The simulation takes global parameters $\mu_a \in \mathbb{R}, \sigma^2 \in \mathbb{R}, \tau_a \in \mathbb{R}, c \in (0, 1]$, and functions $s_a : \mathcal{X} \to \mathbb{R}$. Note that parameters subscripted by $a \in \{0, 1\}$ may vary by group a. For each individual, the data generating process proceeds:

$$\mathbf{x} \sim \max(0, \min(1, \mathcal{N}(\mu_a \mathbf{1}, \sigma^2 \mathbf{I}))) \tag{A.13}$$

$$y = \mathbb{1}\left[s_a(\mathbf{x}) > 5\right] \tag{A.14}$$

$$t \sim \max\left(\mathbb{1}\left[s_a(\mathbf{x}) > \tau_a \vee p = 1\right]\right), \quad p \sim Bernoulli(c) \tag{A.15}$$

$$\tilde{y} = y \cdot t \tag{A.16}$$

Concretely, we generate $\mathbf{x} \in \mathbb{R}^{10}$ for each individual from a multivariate normal distribution. Each Gaussian is assigned mean $\mu_a \mathbf{1}$ based on a, where $\mathbf{1}$ is a 10-dimensional all-ones vector and $\mathbf{I}$ is the identity matrix of size 10×10. We clip all covariates between 0 and 1 (Statement A.13). The process for generating y follows directly from our theoretical setup. For $s_a(\mathbf{x})$, we use an 10-dimensional ceiling function (see Figure 3.8 for a 2D example):

$$s_0(\mathbf{x}) = s_1(\mathbf{x}) = \frac{1}{5}\left(\sum_{i=1}^{10} \lceil 5x_i \rceil\right). \tag{A.17}$$

This function discretizes each element of $\mathbf{x}$ (x_i) into 5 equally-spaced bins of size $1/5$. If the sum of these values exceeds 5, then $y = 1$ (Eq. A.14). Note that $s_a(\cdot)$ can be interpreted as a true risk function for y as a function of $\mathbf{x}$. Except where specified, $s_0(\cdot) = s_1(\cdot)$. We generate t similarly to y based on a threshold applied to $s_a(\cdot)$, but vary $\tau_a \in \mathbb{R}$ as an experimental threshold parameter to control the level of undertesting (Eq. A.15). If the risk for a patient in group a lies above τ_a, they

Setting	μ	$s_a(\mathbf{x})$
1	$\mu_0 = \mu_1 = 0.45$	$s_0(\mathbf{x}) = s_1(\mathbf{x}) = s(\mathbf{x})$
2	$\mu_0 = 0.35, \mu_1 = 0.55$	$s_0(\mathbf{x}) = s_1(\mathbf{x}) = s(\mathbf{x})$
3	$\mu_0 = \mu_1 = 0.45$	$s_0(\mathbf{x}) = s(\mathbf{x}), s_1(\mathbf{x}) = s(Rot(\mathbf{x}; \phi, d', \mathbf{x}_0))$

Table A.17: Simulation settings for each of the three distributional settings studied. Note that $\sigma_0^2 = \sigma_1^2 = 0.1$ for all settings. Scoring function $s(\mathbf{x})$ is defined in Eq. A.17.

are tested with probability 1; otherwise, they are tested with probability $c = 0.05$. Lastly, $\tilde{y} \triangleq y$ if $t = 1$ and is 0 otherwise (Eq. A.16). This models the fact that a test result is only observable when a test is ordered.

Simulating Distributional Differences. To induce differences in the marginal and conditional risk distributions for each patient subgroup, we vary the simulation settings following Table A.17. We define Rot as

$$Rot(\mathbf{x}; \phi, d', \mathbf{x}_0) = \begin{pmatrix} \overbrace{R(-\phi)}^{d'/2 \text{ times}} & & & \\ \vdots & \ddots & & \\ 0 & \dots & R(-\phi) & 0 \\ 0 & \dots & 0 & \mathbf{I}_{10-d'} \end{pmatrix} (\mathbf{x} - \mathbf{x}_0) + \mathbf{x}_0, \tag{A.18}$$

where the function r applies a 2×2 rotation matrix $R(-\phi)$ about $\mathbf{x}_0 = 0.4 \cdot \mathbf{1}$ to any number of pairs of dimensions (rotating the decision boundary by ϕ about the point $\mathbf{x}_0$ in orthogonal 2D subspace(s)). Setting $\phi, d' \neq 0$ satisfies $P_0(y \mid \mathbf{x}) \overset{d}{\neq} P_1(y \mid \mathbf{x})$. Note that Rot breaks the parallelism between the censorship and decision boundaries.

A.2.3 Evaluation Metrics

We provide further intuition for our usage of within-group AUC and xAUC. Recall that the AUC is the probability that a randomly chosen positive example, x_i, has a greater risk score, than a randomly chosen negative example, x_j (*i.e.*, $P(s(x_i) > s(x_j))$). The within-group AUC for group a, written as AUC_a, is thus the AUC considering only examples in group a (Figure 3.6, cyan). The xAUC has a similar interpretation: $\text{xAUC}_{a,a'}$ is the probability that a randomly chosen positive example from group a is scored above a randomly chosen negative example from group a' (Figure 3.6, magenta). We defer to [138] for details on xAUC.

This yields a decomposition of the overall AUC in terms of within-group AUC and xAUC

(Figure 3.6). Let $p_y(a) = P(A = a \mid Y = y)$ for $a \in \{0, 1\}$, $y \in \{0, 1\}$. By the law of total probability, and probabilistic definitions of AUC and xAUC, we have:

$$\begin{aligned}\text{Overall AUC} = p_0(0) \cdot p_1(0) \cdot \text{AUC}_0 + p_0(1) \cdot p_1(1) \cdot \text{AUC}_1 \\ + p_0(0) \cdot p_1(1) \cdot \text{xAUC}_{1,0} + p_0(1) \cdot p_1(0) \cdot \text{xAUC}_{0,1}.\end{aligned} \tag{A.19}$$

Hence, a perfectly separable problem guarantees that an AUC of 1 is possible. When an AUC of 1 is achieved, within-group AUC and xAUC for all groups must also be 1 for Eq. A.19 to hold.

A.2.4 Model Details

We provide all settings used for model training here.

A.2.4.1 Model Training

We train probabilistic kernel support vector machines (SVMs) [223], but any non-linear model suffices. For each setting of the simulation (*i.e.*, unique combination of simulation parameters), we train two SVMs with each model using one set of labels $y, \tilde{y}$ on the same 100 realizations of the simulation with 2,000 training data points and evaluate all models on a simulated sample of 20,000 test samples. As preprocessing, we apply one-hot encoding to $\mathbf{x}$ after discretization.

A.2.4.2 Model Hyperparameters

As the focus of this chapter is on evaluation, we keep all default parameters for the SVM; that is, regularization weight $C = 1$ and $\gamma = (d \cdot Var(\text{vec}(\mathbf{x})))^{-1}$ where $\mathbf{x} \in \mathbb{R}^{n \times d}$ for the radial basis function kernel, where vec is the matrix vectorization operator (*i.e.* `auto` setting in `scikit-learn`).

A.2.4.3 Software

We use `scikit-learn` for the SVM implementation, which is built on LIBSVM [41].

A.2.5 Full Results

For MIMIC results, we report the test names as well as the results of the hypothesis test. For the simulation study, we report the raw AUC and xAUC values by group with empirical 95% confidence intervals for all experiments.

A.2.5.1 Disparate Censorship in MIMIC-IV

Dataset description. We restrict the analysis to all hospital admissions involving a White ($n = 337630$) or Black/African-American patients ($n = 80293$; total: $n = 417923$). These categories were chosen as they represented the two most frequently-appearing racial/ethnic categories in the dataset. We selected the following set of common laboratory/diagnostic tests to investigate for disparate censorship: complete blood counts (CBC), with and without differential (CBC w/ diff.), base metabolic panels (BMP), Troponin T tests, D-dimer tests, arterial blood gas (ABG) tests, blood culture orders (for any organism), brain natriuretic peptide (BNP) tests, and chest X-ray (CXR) orders.[3] To obtain results for ICU and ED admissions, we cross-referenced hospital admission identifiers for ICU and ED stays from the relevant tables to obtain the relevant subset of patients.

CBC, CBC w/ diff., BMP, Troponin T, D-dimer, ABG, and BNP test results were directly available from the publicly available MIMIC concept SQL queries. Additionally, for CBC w/ diff., we excluded rows that did not contain any non-null values in the following columns: `"basophils_abs"`, `"eosinophils_abs"`, `"lymphocytes_abs"`, `"monocytes_abs"`, `"neutrophils_abs"`, `"basophils"`, `"eosinophils"`, `"lymphocytes"`, `"monocytes"`, `"neutrophils"`, `"atypical_lymphocytes"`, `"bands"`, `"immature_granulocytes"`, `"metamyelocytes"`, and `"nrbc"`. For the BMP, we excluded rows that did not contain any non-null values in the following columns: `"bicarbonate"`, `"bun"` (blood urea nitrogen), `"calcium"`, `"chloride"`, `"creatinine"`, `"glucose"`, `"sodium"`, and `"potassium"`. For CXR, we extracted information using the publicly-available MIMIC-CXR processing code at https://github.com/MIT-LCP/mimic-cxr/blob/master/dcm/create-mimic-cxr-jpg-metadata.ipynb.

We then extracted the testing rates ($P(T)$, % of admissions featuring at least one instance of the relevant test order) in each patient group (White vs. Black/African-American) and applied a two-sided z-test for equality of proportions. The null hypothesis is that testing rates are equal between groups (*i.e.*, the test in question does not exhibit disparate censorship in MIMIC-IV). Specifically, this tests the hypothesis that White and Black/African-American patients were equally likely to receive a particular lab test order at any point(s) during each admission. We use a 1% significance threshold with Bonferroni correction (9 tests total; $\alpha = 1.1 \times 10^{-3}$). All p-values below 10^{-4} are reported as "$< 10^{-4}$." In summary, significant disparate censorship with respect to race was identified in all tests examined except for Troponin T and d-dimer tests.

[3] For CXR only, since the MIMIC-CXR data is sourced from 2011-16, we limit ourselves to hospital admissions in that timeframe, yielding 122860 White patient admissions and 25968 Black/African-American admissions (total: 148828).

Test name	$P(T)$, White	$P(T)$, Black	z	p
CBC	73.71	68.20	30.46	$< 10^{-4}$
CBC w/ diff.	31.67	28.81	16.01	$< 10^{-4}$
BMP	71.26	63.72	40.42	$< 10^{-4}$
Blood cultures	15.20	13.01	16.36	$< 10^{-4}$
CXR	27.61	26.57	3.43	6.0×10^{-4}
ABG	13.75	10.42	27.10	$< 10^{-4}$
Troponin T	8.72	8.58	1.29	0.20
BNP	3.82	3.48	4.74	$< 10^{-4}$
D-dimer	0.21	0.25	-1.83	0.07

Table A.18: Disparate censorship in common laboratory/diagnostic tests in White vs. Black/African-American patients, MIMIC-IV v1.0, with testing rates by group, z-statistics, and p-values.

A.2.5.2 Setting 2: Difference in marginal risk distributions

We provide full results with empirical 95% confidence intervals for ΔAUC (Table A.19) and ΔxAUC (Table A.20) under marginal distributional differences. We report results for the full cross-product of $\tau_0, \tau_1 \in \{5, 5.4, 5.8, 6.2, 6.6, 7\}$ with their associated group noise rates.

A.2.5.3 Setting 3: Difference in conditional risk distributions

We provide full results with empirical 95% confidence intervals for ΔAUC and ΔxAUC under marginal distributional differences organized by d', the number of dimensions rotated for group $a = 1$ individuals. We plot heatmap visualizations of the performance gap as a function of noise rate (*i.e.* τ_1) and conditional shift (ϕ) across levels of d'. We index the figures with results for all choices of d' below:

- $d' = 2$: Figure A.5
- $d' = 4$: Figure A.6
- $d' = 6$: Figure A.7
- $d' = 8$: Figure A.8
- $d' = 10$: Figure A.9

We index the tables with all raw AUC and xAUC values below:

- $d' = 2$: AUC (Table A.21), xAUC (Table A.22)
- $d' = 4$: AUC (Table A.23), xAUC (Table A.24)

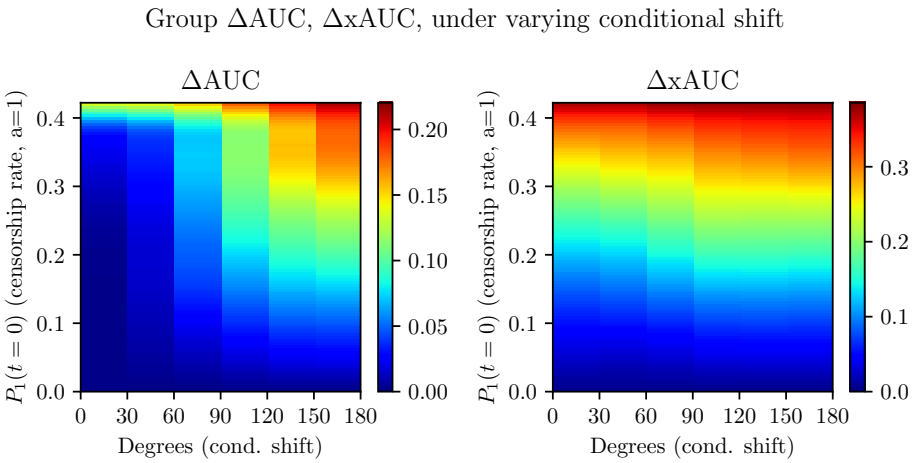


Figure A.5: Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (x-axis) and censorship rate in $a = 1$ (y-axis); 2 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger performance gaps are in dark red.

- $d' = 6$: AUC (Table A.25), xAUC (Table A.26)
- $d' = 8$: AUC (Table A.27), xAUC (Table A.28)
- $d' = 10$: AUC (Table A.29), xAUC (Table A.30)

In general, performance gaps worsen as d' or ϕ increase. Note the apparent duplicated AUC values at $d' = 10, \phi = 180$; as all 100 realizations of the simulation share the same set of random seeds across all experiments—since all dimensions are rotated, all points in group $a = 1$ flip across the decision boundary, yielding the exact same censorship pattern.

A.2.5.4 Additional Results for Varying Marginal Distributional Distance (Setting 2)

Increasing divergence between the marginal risk distributions correlates with larger performance gaps under disparate censorship. We set $\tau_0 = 5, \tau_1 = 6.6$, a setting in which we observed performance gaps under disparate censorship, and vary $\Delta\mu$ (defined as $\mu_1 - \mu_0$) and σ. We also fix $\mu_0 + \mu_1 = 0.35 + 0.55 = 0.9$. We are interested whether varying the KL divergence between the marginals, given by $\frac{1}{2\sigma}\|\Delta\mu\|_2^2$, impacts the magnitude of performance gaps. To that end, we first explore the impacts of changing σ for both groups on subgroup performance gaps. We then vary $\Delta\mu$, or the difference of the means between the marginal risk distributions, on subgroup performance gaps. In summary, we find that increasing the KL divergence by either decreasing the within-group variance or increasing the mean difference exacerbates the negative impacts of disparate censorship.

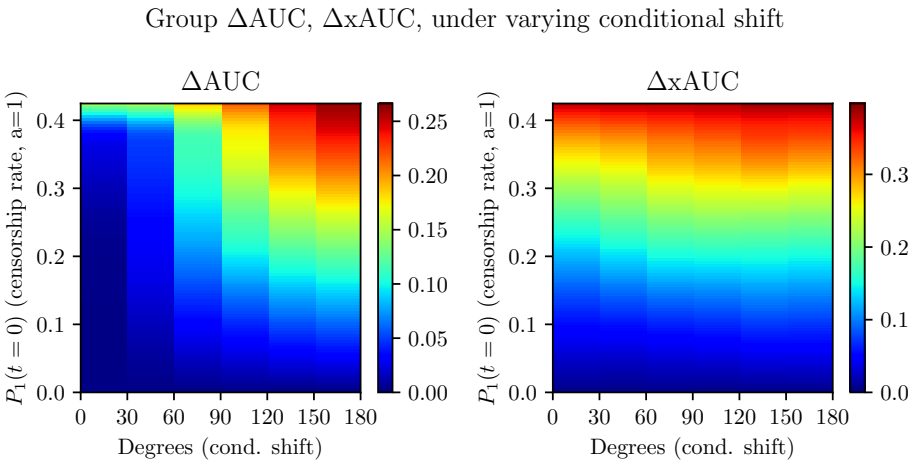


Figure A.6: Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (ϕ; x-axis) and censorship rate in $a = 1$ ($P_1(t = 0)$, y-axis); 4 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger performance gaps are in dark red.

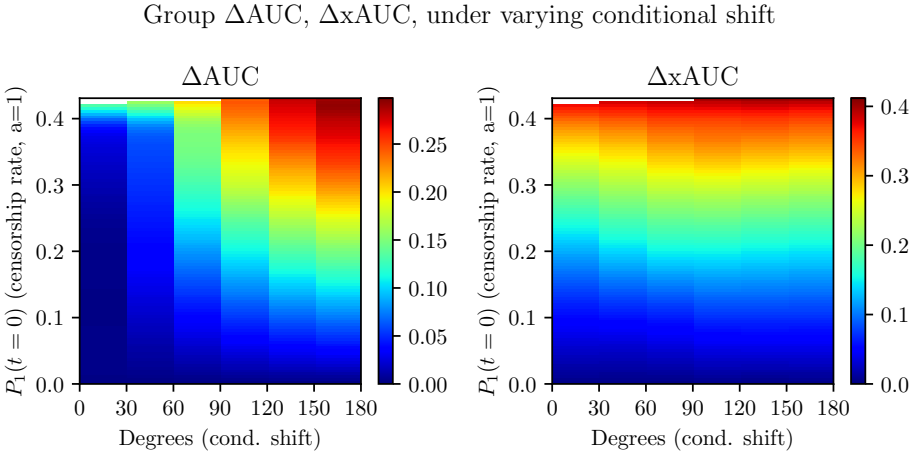


Figure A.7: Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (ϕ; x-axis) and censorship rate in $a = 1$ ($P_1(t = 0)$, y-axis); 6 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger performance gaps are in dark red.

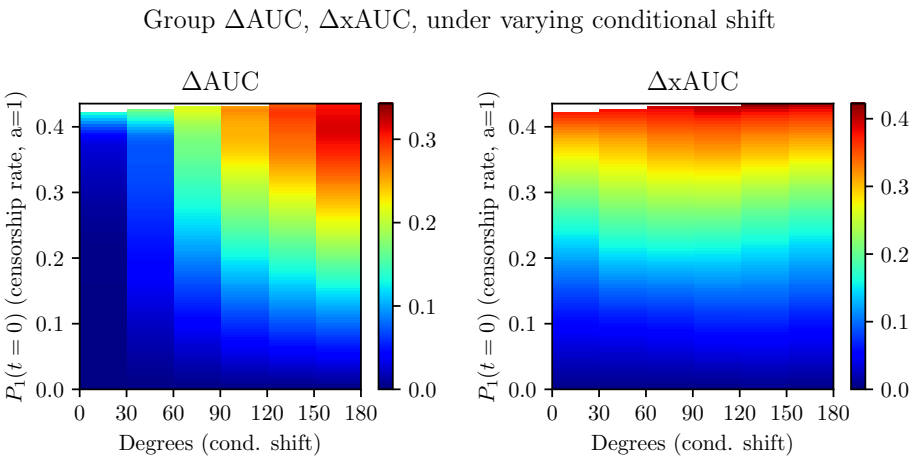


Figure A.8: Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (ϕ; x-axis) and censorship rate in $a = 1$ ($P_1(t = 0)$, y-axis); 8 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger performance gaps are in dark red.

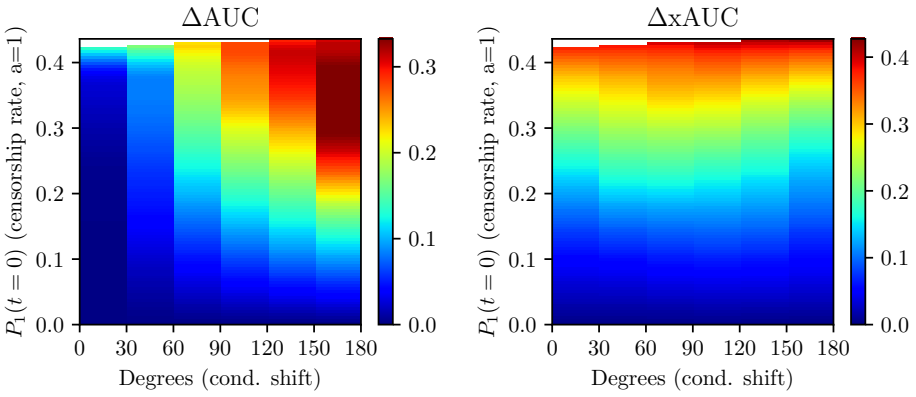


Figure A.9: Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (ϕ; x-axis) and censorship rate in $a = 1$ ($P_1(t = 0)$, y-axis); 10 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger performance gaps are in dark red.

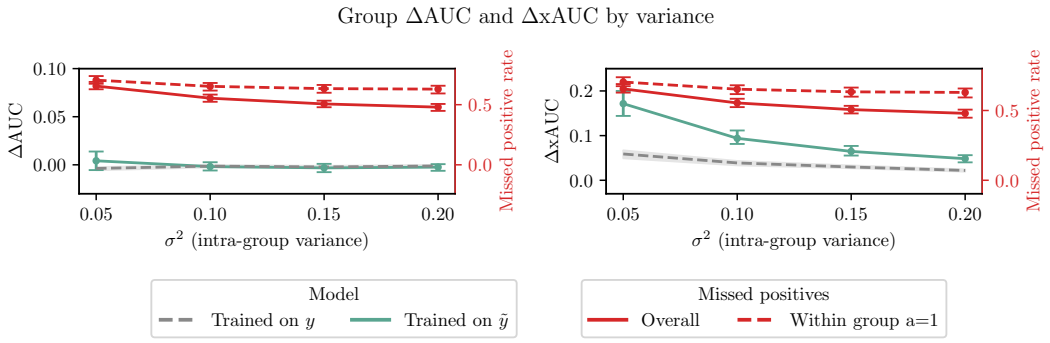


Figure A.10: ΔAUC (left) and ΔxAUC (right) with 95% empirical CIs for the model trained on y (gray) and on $\tilde{y}$ (green). As σ^2 increases to the right, increasing overlap between the distributions, the number of missed positives increases, and ΔAUC, ΔxAUC decrease.

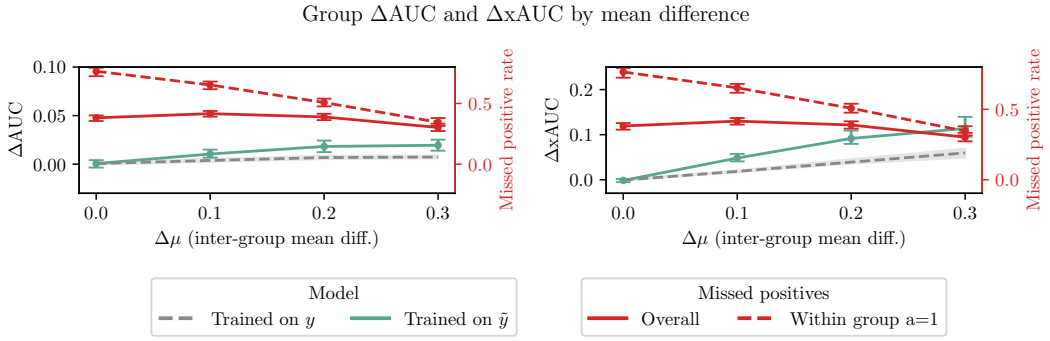


Figure A.11: ΔAUC (left) and ΔxAUC (right) with 95% empirical CIs for the model trained on y (gray) and on $\tilde{y}$ (green). As $\Delta\mu$ increases to the right, the AUC and xAUC gaps *widen*. Note that $P_0(\tilde{y} = 0 \mid y = 1) = 0$—there are no missed positives in group $a = 0$. The number of missed positives first increases, then decreases, since sufficiently high $\Delta\mu$ means that untested patients' risk largely lies in one tail of the risk distribution.

We first evaluate the effect of within-group variance. As we decrease σ^2 from 0.2 to 0.05, we observe increasing AUC and xAUC gaps. As seen in Figure A.10, at a variance of 0.05 (left side, all graphs), the median AUC and xAUC gaps are 0.09 and 0.22. However, as we increase the within-group variance to 0.2 (right side, all graphs), the median AUC and xAUC gaps narrow to 0.01 and 0.06. Decreasing σ^2 also decreases the overlap between the two distributions, leading to greater performance disparities.

We then evaluate the effect of mean difference on performance disparities. As between-group mean difference increases, disparities between groups generally worsen with respect to oracle performance. We increase the per-element mean difference between the distributions from 0 to 0.4, keeping the decision boundary constant. Figure A.11 shows that as mean difference $\Delta\mu$ increases, disparities in AUC, xAUC emerge. At $\Delta\mu = 0$, the performance is identical to the oracle, as

expected. At $\Delta\mu = 0$, we are in Setting 1: there are no marginal or conditional differences between groups. However, as $\Delta\mu$ grows to 0.3, the AUC and xAUC gaps increase to 0.03, 0.14, respectively.

Our results suggest that distributional distance between patient subgroups as measured using mean difference and variance is correlated with performance disparities. Characterizing distributional differences in patient subgroup risk could provide expected levels of performance disparities under disparate censorship. Full results are provided in Table A.31 (differences in $\Delta\mu$) and Table A.32 (differences in σ^2).

A.2.6 Code

For reproducibility, code used to generate all figures and experimental results after review is provided at the MLD3 Github.

		τ_1 (noise rate, a=1)					
τ_0 (noise rate, $a = 0$)	**Group**	**5.0** (0.0)	**5.4** (8.6)	**5.8** (14.5)	**6.2** (37.7)	**6.6** (46.2)	**7.0** (68.3)
5.0 (0.0)	a=0	98.3 (98.1, 98.5)	98.2 (98.0, 98.4)	98.0 (97.7, 98.3)	96.4 (95.8, 96.8)	95.1 (94.0, 95.7)	85.0 (83.4, 87.0)
	a=1	99.0 (98.8, 99.1)	98.9 (98.8, 99.1)	98.8 (98.6, 99.0)	97.7 (97.3, 98.1)	96.8 (96.2, 97.3)	90.0 (88.7, 91.6)
	ΔAUC	0.7 (0.5, 0.9)	0.7 (0.5, 0.9)	0.8 (0.6, 1.0)	1.4 (1.0, 1.8)	1.8 (1.3, 2.3)	5.0 (4.0, 6.2)
5.4 (11.8)	a=0	98.2 (98.0, 98.4)	98.4 (98.1, 98.5)	98.3 (98.0, 98.5)	97.3 (96.9, 97.7)	96.6 (95.9, 97.1)	90.7 (89.0, 92.1)
	a=1	98.9 (98.8, 99.1)	99.0 (98.9, 99.1)	99.0 (98.8, 99.1)	98.4 (98.1, 98.6)	97.9 (97.5, 98.2)	94.0 (92.9, 95.2)
	ΔAUC	0.7 (0.6, 0.9)	0.7 (0.5, 0.8)	0.7 (0.5, 0.9)	1.0 (0.8, 1.3)	1.3 (1.0, 1.7)	3.5 (2.7, 4.2)
5.8 (16.2)	a=0	98.1 (97.8, 98.3)	98.3 (98.1, 98.5)	98.3 (98.0, 98.5)	97.6 (97.2, 97.9)	97.1 (96.5, 97.6)	92.7 (90.8, 93.9)
	a=1	98.9 (98.7, 99.0)	99.0 (98.8, 99.1)	99.0 (98.8, 99.1)	98.5 (98.3, 98.7)	98.2 (97.8, 98.5)	95.4 (94.1, 96.3)
	ΔAUC	0.8 (0.6, 0.9)	0.7 (0.5, 0.9)	0.7 (0.5, 0.8)	0.9 (0.7, 1.2)	1.2 (0.8, 1.5)	2.7 (2.1, 3.6)
6.2 (23.6)	a=0	97.6 (97.3, 97.9)	98.0 (97.6, 98.2)	98.0 (97.7, 98.3)	97.7 (97.3, 98.0)	97.4 (96.9, 97.9)	95.0 (93.6, 96.2)
	a=1	98.6 (98.3, 98.8)	98.8 (98.6, 98.9)	98.8 (98.6, 99.0)	98.6 (98.3, 98.8)	98.4 (98.1, 98.7)	96.9 (95.9, 97.7)
	ΔAUC	0.9 (0.7, 1.2)	0.8 (0.6, 1.0)	0.8 (0.6, 1.0)	0.9 (0.7, 1.2)	1.0 (0.8, 1.3)	1.9 (1.3, 2.5)
6.6 (24.6)	a=0	97.5 (97.1, 97.9)	97.9 (97.5, 98.2)	98.0 (97.6, 98.2)	97.7 (97.2, 98.0)	97.4 (96.9, 97.8)	95.1 (93.6, 96.4)
	a=1	98.5 (98.2, 98.7)	98.7 (98.5, 98.9)	98.8 (98.5, 99.0)	98.6 (98.3, 98.8)	98.4 (98.1, 98.7)	97.0 (96.1, 97.7)
	ΔAUC	1.0 (0.8, 1.2)	0.8 (0.6, 1.1)	0.8 (0.6, 1.0)	0.9 (0.7, 1.2)	1.0 (0.8, 1.3)	1.9 (1.4, 2.7)
7.0 (26.1)	a=0	97.4 (97.0, 97.8)	97.8 (97.4, 98.1)	97.9 (97.5, 98.1)	97.6 (97.1, 97.9)	97.3 (96.7, 97.7)	95.2 (93.3, 96.2)
	a=1	98.4 (98.1, 98.6)	98.7 (98.4, 98.8)	98.7 (98.4, 98.9)	98.5 (98.3, 98.8)	98.4 (98.0, 98.6)	97.0 (95.9, 97.6)
	ΔAUC	1.0 (0.8, 1.2)	0.9 (0.7, 1.1)	0.8 (0.7, 1.1)	1.0 (0.7, 1.3)	1.1 (0.8, 1.3)	1.9 (1.4, 2.6)

Table A.19: Group-wise AUC and ΔAUC under marginal distributional distance for all values of τ_0, τ_1.

		τ_1 (noise rate, $a = 1$)					
τ_0 (noise rate, $a = 0$)	**Group**	**5.0** (0.0)	**5.4** (8.6)	**5.8** (14.5)	**6.2** (37.7)	**6.6** (46.2)	**7.0** (68.3)
5.0 (0.0)	a=0	99.6 (99.5, 99.7)	99.6 (99.5, 99.6)	99.5 (99.4, 99.6)	99.0 (98.8, 99.1)	98.6 (98.3, 98.8)	93.8 (92.8, 95.0)
	a=1	95.7 (95.0, 96.1)	95.5 (95.0, 95.9)	95.0 (94.5, 95.7)	91.5 (90.4, 92.6)	89.5 (87.8, 90.7)	77.0 (75.1, 79.6)
	ΔxAUC	3.9 (3.6, 4.6)	4.1 (3.7, 4.6)	4.5 (3.9, 5.0)	7.5 (6.4, 8.5)	9.2 (8.0, 10.6)	16.6 (15.3, 18.2)
5.4 (11.8)	a=0	99.6 (99.5, 99.6)	99.6 (99.6, 99.7)	99.6 (99.6, 99.7)	99.4 (99.2, 99.4)	99.1 (98.9, 99.3)	96.8 (96.1, 97.5)
	a=1	95.5 (94.8, 95.9)	95.7 (95.1, 96.1)	95.5 (94.9, 96.1)	93.5 (92.6, 94.4)	92.1 (90.8, 93.1)	83.4 (81.0, 85.7)
	ΔxAUC	4.1 (3.7, 4.7)	3.9 (3.5, 4.5)	4.1 (3.6, 4.6)	5.9 (5.1, 6.7)	7.0 (6.1, 8.2)	13.4 (11.7, 15.2)
5.8 (16.2)	a=0	99.6 (99.5, 99.6)	99.6 (99.6, 99.7)	99.6 (99.5, 99.7)	99.4 (99.3, 99.5)	99.3 (99.1, 99.4)	97.7 (97.0, 98.2)
	a=1	95.1 (94.5, 95.8)	95.5 (95.0, 96.1)	95.5 (94.9, 96.0)	94.0 (93.1, 94.8)	92.9 (91.6, 93.9)	85.9 (83.2, 87.7)
	ΔxAUC	4.4 (3.8, 5.0)	4.1 (3.5, 4.6)	4.1 (3.6, 4.7)	5.4 (4.7, 6.2)	6.4 (5.5, 7.5)	11.8 (10.3, 13.9)
6.2 (23.6)	a=0	99.4 (99.3, 99.5)	99.5 (99.4, 99.6)	99.6 (99.5, 99.6)	99.5 (99.3, 99.6)	99.4 (99.2, 99.5)	98.6 (98.1, 99.0)
	a=1	94.2 (93.1, 94.9)	94.8 (94.0, 95.5)	95.0 (94.2, 95.5)	94.2 (93.3, 95.0)	93.6 (92.3, 94.6)	89.1 (87.0, 91.4)
	ΔxAUC	5.3 (4.6, 6.3)	4.7 (4.0, 5.4)	4.5 (4.1, 5.3)	5.3 (4.5, 6.1)	5.8 (4.9, 6.9)	9.5 (7.6, 11.3)
6.6 (24.6)	a=0	99.4 (99.3, 99.5)	99.5 (99.4, 99.6)	99.5 (99.4, 99.6)	99.5 (99.3, 99.5)	99.4 (99.2, 99.5)	98.7 (98.2, 99.0)
	a=1	94.0 (92.8, 94.7)	94.7 (93.8, 95.3)	94.8 (93.9, 95.4)	94.0 (93.2, 94.9)	93.5 (92.2, 94.5)	89.3 (86.8, 91.6)
	ΔxAUC	5.4 (4.7, 6.5)	4.9 (4.2, 5.7)	4.7 (4.2, 5.5)	5.4 (4.6, 6.2)	5.9 (5.0, 6.9)	9.3 (7.4, 11.5)
7.0 (26.1)	a=0	99.4 (99.2, 99.5)	99.5 (99.4, 99.5)	99.5 (99.4, 99.6)	99.4 (99.3, 99.5)	99.3 (99.2, 99.5)	98.7 (98.2, 99.0)
	a=1	93.7 (92.5, 94.5)	94.4 (93.5, 95.1)	94.7 (93.7, 95.2)	93.8 (92.9, 94.6)	93.3 (92.0, 94.3)	89.4 (86.4, 91.1)
	ΔxAUC	5.7 (5.0, 6.8)	5.0 (4.4, 5.9)	4.8 (4.4, 5.7)	5.6 (4.8, 6.4)	6.1 (5.2, 7.2)	9.3 (7.8, 11.8)

Table A.20: xAUC metrics and ΔxAUC under marginal distributional distance for all values of τ_0, τ_1.

		τ_1					
ϕ	**Group**	**5.0**	**5.4**	**5.8**	**6.2**	**6.6**	**7.0**
0	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	97.6 (97.3, 97.9)	97.5 (97.1, 97.8)	97.3 (96.8, 97.6)	97.7 (97.3, 98.1)	93.7 (92.1, 94.9)	83.4 (77.2, 86.9)
	ΔAUC	0.2 (0.0, 0.6)	0.2 (0.0, 0.7)	0.4 (0.0, 1.0)	2.4 (1.6, 3.6)	4.0 (2.9, 5.7)	14.3 (10.8, 20.5)
30	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	97.2 (96.8, 97.7)	96.3 (95.6, 96.9)	96.0 (95.3, 96.7)	94.1 (92.6, 95.0)	92.7 (90.7, 93.8)	83.0 (77.1, 87.3)
	ΔAUC	0.5 (0.0, 1.0)	1.4 (0.8, 2.2)	1.7 (1.0, 2.6)	3.6 (2.6, 5.0)	5.0 (3.7, 7.1)	14.7 (10.1, 20.5)
60	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	96.0 (95.2, 96.6)	93.3 (92.3, 94.3)	92.6 (91.2, 93.8)	90.7 (89.1, 92.0)	89.9 (87.5, 91.4)	81.6 (76.8, 85.7)
	ΔAUC	1.7 (1.1, 2.5)	4.3 (3.4, 5.5)	5.1 (3.9, 6.5)	7.0 (5.6, 8.7)	7.8 (6.4, 10.2)	16.0 (11.9, 20.8)
90	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	93.9 (92.9, 94.8)	90.6 (89.5, 92.0)	88.9 (87.3, 90.5)	86.5 (84.8, 88.1)	86.0 (83.2, 88.2)	79.5 (74.4, 83.6)
	ΔAUC	3.7 (2.9, 4.8)	7.1 (5.9, 8.2)	8.7 (7.2, 10.3)	11.3 (9.5, 13.1)	11.7 (9.6, 14.4)	18.3 (13.9, 23.1)
120	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	92.1 (90.5, 93.2)	88.2 (86.7, 89.4)	86.4 (84.5, 88.2)	82.2 (80.4, 84.5)	82.1 (78.8, 84.4)	77.6 (72.0, 81.8)
	ΔAUC	5.6 (4.5, 7.3)	9.5 (8.1, 11.2)	11.3 (9.5, 13.2)	15.5 (13.1, 17.5)	15.6 (13.4, 19.1)	20.2 (15.9, 25.7)
150	a=0	97.4 (97.0, 97.8)	97.8 (97.4, 98.1)	97.9 (97.5, 98.1)	97.6 (97.1, 97.9)	97.3 (96.7, 97.7)	95.2 (93.3, 96.2)
	a=1	91.2 (89.7, 92.4)	86.9 (85.3, 88.2)	84.9 (83.4, 86.3)	80.1 (77.9, 82.5)	79.7 (76.7, 82.2)	76.2 (70.5, 80.2)
	ΔAUC	6.6 (5.3, 8.1)	10.8 (9.2, 12.3)	12.8 (11.0, 14.3)	17.5 (15.0, 19.9)	17.9 (15.1, 21.2)	21.5 (17.4, 27.3)
180	a=0	97.4 (97.0, 97.8)	97.8 (97.4, 98.1)	97.9 (97.5, 98.1)	97.6 (97.1, 97.9)	97.3 (96.7, 97.7)	95.2 (93.3, 96.2)
	a=1	91.0 (89.5, 92.4)	86.8 (85.0, 88.2)	85.0 (83.0, 86.7)	80.0 (77.6, 82.8)	79.6 (76.6, 82.1)	75.5 (70.0, 79.7)
	ΔAUC	6.7 (5.2, 8.1)	11.0 (9.1, 12.8)	12.7 (10.8, 14.4)	17.6 (14.9, 20.2)	18.1 (15.3, 21.1)	22.0 (18.0, 27.8)

Table A.21: Group-wise AUC and ΔAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 2$.

		τ_1					
ϕ	**Group**	**5.0**	**5.4**	**5.8**	**6.2**	**6.6**	**7.0**
0	a=0	97.6 (97.0, 98.1)	90.9 (89.5, 92.4)	85.8 (83.6, 88.3)	72.3 (68.5, 77.0)	69.0 (64.6, 73.6)	97.6 (97.0, 98.1)
	a=1	97.6 (97.2, 98.2)	99.7 (99.5, 99.8)	99.9 (99.8, 99.9)	100.0 (99.9, 100.0)	100.0 (99.9, 100.0)	99.9 (99.8, 100.0)
	ΔxAUC	0.4 (0.1, 1.1)	8.8 (7.1, 10.2)	14.1 (11.6, 16.3)	27.6 (23.0, 31.4)	30.9 (26.3, 35.4)	37.6 (32.0, 44.2)
30	a=0	96.5 (95.5, 97.2)	89.9 (88.1, 91.8)	85.1 (82.6, 87.5)	71.6 (66.3, 76.7)	68.9 (63.2, 73.6)	62.4 (56.6, 68.0)
	a=1	98.2 (97.8, 98.6)	99.5 (99.2, 99.6)	99.8 (99.6, 99.9)	99.9 (99.9, 100.0)	100.0 (99.9, 100.0)	99.9 (99.8, 100.0)
	ΔxAUC	1.8 (0.8, 2.9)	9.5 (7.6, 11.4)	14.6 (12.2, 17.2)	28.3 (23.2, 33.7)	31.0 (26.3, 36.8)	37.5 (31.9, 43.4)
60	a=0	94.0 (92.8, 95.1)	87.3 (85.1, 89.1)	82.7 (80.1, 85.8)	70.9 (66.5, 76.1)	68.1 (62.4, 73.5)	61.8 (56.5, 67.8)
	a=1	98.6 (98.1, 98.9)	99.2 (98.9, 99.4)	99.5 (99.2, 99.7)	99.9 (99.8, 99.9)	99.9 (99.8, 100.0)	99.9 (99.8, 100.0)
	ΔxAUC	4.5 (3.2, 5.9)	11.9 (9.7, 14.2)	16.7 (13.5, 19.5)	28.9 (23.8, 33.4)	31.8 (26.4, 37.6)	38.2 (32.0, 43.5)
90	a=0	91.4 (89.8, 93.1)	83.7 (81.2, 87.3)	79.6 (76.3, 83.5)	69.4 (64.7, 75.2)	67.0 (61.0, 73.1)	61.8 (55.9, 68.6)
	a=1	98.6 (98.1, 98.9)	99.1 (98.8, 99.4)	99.3 (99.0, 99.5)	99.8 (99.6, 99.9)	99.9 (99.7, 99.9)	99.9 (99.8, 100.0)
	ΔxAUC	7.3 (5.4, 9.1)	15.4 (11.6, 18.1)	19.7 (15.5, 23.1)	30.3 (24.5, 35.1)	32.8 (26.6, 38.9)	38.1 (31.2, 44.1)
120	a=0	89.4 (87.7, 91.3)	81.7 (78.8, 85.5)	77.8 (74.1, 81.6)	68.5 (62.8, 73.8)	66.3 (61.1, 72.5)	61.8 (56.3, 67.6)
	a=1	98.6 (98.1, 98.9)	99.1 (98.8, 99.3)	99.3 (99.0, 99.5)	99.7 (99.5, 99.8)	99.8 (99.7, 99.9)	99.9 (99.7, 100.0)
	ΔxAUC	9.2 (7.0, 10.9)	17.5 (13.4, 20.4)	21.4 (17.4, 25.2)	31.2 (25.7, 37.0)	33.5 (27.2, 38.8)	38.1 (32.2, 43.7)
150	a=0	88.3 (86.7, 90.6)	80.9 (77.8, 84.4)	77.1 (73.5, 81.4)	67.6 (62.7, 73.9)	66.1 (60.7, 72.1)	61.7 (55.8, 67.3)
	a=1	98.6 (98.2, 99.0)	99.0 (98.6, 99.3)	99.2 (98.9, 99.5)	99.6 (99.4, 99.8)	99.8 (99.6, 99.9)	99.9 (99.7, 100.0)
	ΔxAUC	10.3 (7.8, 12.1)	18.2 (14.3, 21.1)	22.2 (17.5, 25.8)	31.9 (25.6, 37.0)	33.7 (27.5, 39.1)	38.2 (32.4, 44.1))
180	a=0	88.5 (86.5, 90.6)	81.1 (78.3, 84.4)	77.0 (72.9, 81.2)	67.5 (63.2, 72.9)	65.3 (60.4, 71.1)	61.3 (56.0, 67.8)
	a=1	98.6 (98.2, 98.9)	99.0 (98.6, 99.2)	99.2 (98.9, 99.4)	99.6 (99.4, 99.8)	99.8 (99.6, 99.9)	99.9 (99.7, 100.0)
	ΔxAUC	10.1 (7.7, 12.1)	17.8 (14.4, 20.5)	22.3 (17.6, 26.4)	32.2 (26.6, 36.6)	34.5 (28.3, 39.5)	38.5 (32.0, 44.0)

Table A.22: xAUC and ΔxAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 2$.

		τ_1					
ϕ	**Group**	**5.0**	**5.4**	**5.8**	**6.2**	**6.6**	**7.0**
0	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	97.6 (97.3, 97.9)	97.5 (97.1, 97.8)	97.3 (96.8, 97.6)	97.7 (97.3, 98.1)	93.7 (92.1, 94.9)	83.4 (77.2, 86.9)
	ΔAUC	0.2 (0.0, 0.6)	0.2 (0.0, 0.7)	0.4 (0.0, 1.0)	2.4 (1.6, 3.6)	4.0 (2.9, 5.7)	14.3 (10.8, 20.5)
30	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	97.0 (96.5, 97.4)	95.4 (94.4, 96.0)	94.8 (93.9, 95.6)	93.0 (91.9, 94.1)	91.7 (90.0, 93.1)	82.9 (77.2, 86.6)
	ΔAUC	0.7 (0.1, 1.3)	2.3 (1.6, 3.3)	2.9 (1.9, 3.8)	4.7 (3.5, 5.8)	5.9 (4.5, 8.0)	14.7 (11.2, 20.7)
60	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	94.5 (93.5, 95.3)	91.2 (89.5, 92.2)	89.5 (87.7, 91.0)	86.6 (84.7, 88.3)	85.8 (83.5, 87.7)	79.6 (75.7, 83.8)
	ΔAUC	3.2 (2.3, 4.2)	6.5 (5.4, 8.0)	8.2 (6.7, 10.0)	11.2 (9.4, 13.1)	11.8 (9.9, 14.2)	18.2 (14.1, 22.4)
90	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	91.4 (90.1, 92.6)	87.0 (85.5, 88.4)	85.2 (83.2, 87.0)	80.7 (78.7, 83.3)	80.3 (77.7, 83.1)	76.3 (71.9, 79.5)
	ΔAUC	6.2 (5.0, 7.8)	10.6 (9.2, 12.2)	12.5 (10.6, 14.6)	17.0 (14.4, 19.0)	17.6 (14.5, 19.9)	21.5 (18.1, 25.8)
120	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	87.9 (85.9, 89.7)	82.9 (81.0, 84.7)	80.7 (78.4, 83.0)	75.8 (73.6, 79.0)	75.0 (72.3, 79.0)	73.0 (68.8, 77.3)
	ΔAUC	9.8 (8.1, 11.6)	14.7 (12.7, 16.9)	16.9 (14.7, 19.3)	21.9 (18.8, 24.2)	22.7 (18.6, 25.5)	24.8 (20.4, 28.9)
150	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	85.4 (82.9, 87.4)	80.3 (77.6, 82.5)	77.9 (75.4, 80.7)	73.5 (70.7, 76.8)	72.7 (69.9, 76.4)	71.6 (66.1, 76.1)
	ΔAUC	12.3 (10.4, 14.7)	17.4 (15.1, 20.2)	19.8 (17.1, 22.0)	24.3 (20.9, 27.2)	25.1 (21.2, 28.0)	26.2 (21.7, 31.7)
180	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	84.4 (82.4, 85.8)	79.1 (76.7, 81.2)	77.1 (75.1, 79.9)	72.6 (69.2, 75.2)	72.3 (69.1, 76.0)	71.0 (66.0, 75.4)
	ΔAUC	13.4 (11.9, 15.4)	18.5 (16.7, 21.2)	20.6 (18.0, 22.6)	25.1 (22.4, 28.4)	25.3 (21.7, 28.6)	26.6 (22.4, 31.9)

Table A.23: Group-wise AUC and ΔAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 4$.

		τ_1					
ϕ	**Group**	**5.0**	**5.4**	**5.8**	**6.2**	**6.6**	**7.0**
0	a=0	97.6 (97.0, 98.1)	90.9 (89.5, 92.4)	85.8 (83.6, 88.3)	72.3 (68.5, 77.0)	69.0 (64.6, 73.6)	62.3 (55.8, 67.7)
	a=1	97.6 (97.2, 98.2)	99.7 (99.5, 99.8)	99.9 (99.8, 99.9)	100.0 (99.9, 100.0)	100.0 (99.9, 100.0)	99.9 (99.8, 100.0)
	ΔxAUC	0.4 (0.1, 1.1)	8.8 (7.1, 10.2)	14.1 (11.6, 16.3)	27.6 (23.0, 31.4)	30.9 (26.3, 35.4)	37.6 (32.0, 44.2)
30	a=0	95.7 (94.8, 96.6)	89.1 (87.0, 91.0)	83.9 (81.6, 86.7)	71.5 (67.0, 76.6)	68.3 (63.7, 74.6)	62.1 (56.4, 68.2)
	a=1	98.5 (98.0, 98.8)	99.4 (99.1, 99.5)	99.7 (99.5, 99.8)	99.9 (99.9, 100.0)	99.9 (99.9, 100.0)	99.9 (99.8, 100.0)
	ΔxAUC	2.7 (1.7, 3.7)	10.2 (8.1, 12.4)	15.7 (12.7, 18.1)	28.4 (23.2, 33.0)	31.6 (25.3, 36.3)	37.9 (31.6, 43.5)
60	a=0	92.3 (90.9, 93.7)	85.2 (82.3, 87.7)	80.7 (77.7, 83.7)	69.9 (64.3, 74.9)	67.5 (62.2, 72.7)	62.4 (56.7, 68.3)
	a=1	98.6 (98.1, 98.9)	99.1 (98.7, 99.3)	99.3 (99.0, 99.5)	99.9 (99.7, 99.9)	99.8 (99.6, 99.9)	99.9 (99.8, 100.0)
	ΔxAUC	6.3 (4.6, 7.9)	14.1 (11.2, 16.9)	18.6 (15.6, 21.6)	29.9 (24.7, 35.6)	32.3 (27.1, 37.7)	37.6 (31.5, 43.3)
90	a=0	89.0 (86.9, 90.8)	81.6 (79.4, 84.7)	77.3 (73.7, 81.0)	68.3 (63.5, 72.7)	66.0 (61.0, 72.2)	62.0 (56.1, 67.9)
	a=1	98.6 (98.2, 99.0)	99.0 (98.6, 99.2)	99.2 (99.0, 99.4)	99.7 (99.5, 99.8)	99.8 (99.6, 99.9)	99.9 (99.7, 100.0)
	ΔxAUC	9.7 (7.7, 11.8)	17.4 (14.1, 19.6)	22.0 (18.1, 25.6)	31.4 (26.8, 36.3)	33.7 (27.5, 38.9)	37.9 (31.9, 43.7)
120	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	92.1 (90.5, 93.2)	88.2 (86.7, 89.4)	86.4 (84.5, 88.2)	82.2 (80.4, 84.5)	82.1 (78.8, 84.4)	77.6 (72.0, 81.8)
	ΔxAUC	5.6 (4.5, 7.3)	9.5 (8.1, 11.2)	11.3 (9.5, 13.2)	15.5 (13.1, 17.5)	15.6 (13.4, 19.1)	20.2 (15.9, 25.7)
150	a=0	85.5 (83.2, 87.8)	78.4 (75.1, 81.7)	74.7 (70.5, 78.6)	66.0 (62.3, 71.5)	64.9 (59.8, 70.8)	61.0 (56.0, 66.8)
	a=1	98.6 (98.2, 98.9)	99.0 (98.6, 99.2)	99.2 (98.9, 99.4)	99.6 (99.4, 99.8)	99.7 (99.6, 99.9)	99.9 (99.7, 100.0)
	ΔxAUC	13.3 (10.7, 15.5)	20.6 (17.3, 24.0)	24.6 (20.7, 28.7)	33.6 (28.0, 37.5)	34.9 (28.8, 40.0)	38.9 (33.0, 43.8)
180	a=0	83.2 (80.8, 85.9)	76.3 (72.2, 80.2)	72.7 (68.4, 76.7)	65.4 (60.7, 70.9)	64.2 (58.7, 69.4)	60.1 (54.7, 65.6)
	a=1	98.5 (98.0, 98.9)	98.9 (98.5, 99.2)	99.2 (98.8, 99.5)	99.6 (99.4, 99.8)	99.8 (99.6, 99.9)	99.9 (99.8, 100.0)
	ΔxAUC	15.5 (12.2, 17.9)	22.6 (18.7, 26.8)	26.6 (22.2, 30.7)	34.3 (28.7, 39.0)	35.6 (30.3, 41.1)	39.8 (34.2, 45.3)

Table A.24: Group-wise xAUC and ΔxAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 4$.

		τ_1					
ϕ	**Group**	**5.0**	**5.4**	**5.8**	**6.2**	**6.6**	**7.0**
0	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	97.6 (97.3, 97.9)	97.5 (97.1, 97.8)	97.3 (96.8, 97.6)	97.7 (97.3, 98.1)	93.7 (92.1, 94.9)	83.4 (77.2, 86.9)
	ΔAUC	0.2 (0.0, 0.6)	0.2 (0.0, 0.7)	0.4 (0.0, 1.0)	2.4 (1.6, 3.6)	4.0 (2.9, 5.7)	14.3 (10.8, 20.5)
30	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	96.5 (95.8, 97.1)	94.6 (93.4, 95.4)	93.7 (92.4, 94.6)	91.9 (90.2, 93.0)	90.7 (88.6, 92.0)	82.1 (77.1, 86.8)
	ΔAUC	1.2 (0.6, 1.9)	3.1 (2.3, 4.2)	3.9 (2.9, 5.3)	5.8 (4.6, 7.5)	7.1 (5.7, 9.1)	15.6 (11.2, 20.6)
60	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	92.9 (91.4, 93.8)	89.0 (87.6, 90.3)	87.0 (85.5, 88.7)	83.1 (81.0, 85.2)	89.9 (87.5, 91.4)	77.9 (72.6, 81.6)
	ΔAUC	4.7 (3.9, 6.3)	8.6 (7.3, 10.0)	10.7 (8.9, 12.2)	14.6 (12.4, 16.8)	15.1 (13.2, 18.0)	19.8 (16.1, 25.2)
90	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	88.1 (86.4, 89.7)	83.4 (81.6, 85.4)	80.9 (78.8, 83.2)	76.4 (74.1, 79.3)	75.3 (71.9, 78.7)	79.5 (74.4, 83.6)
	ΔAUC	9.7 (8.0, 11.3)	14.4 (12.3, 16.0)	16.9 (14.4, 18.9)	11.3 (9.5, 13.1)	21.3 (18.4, 23.6)	22.5 (18.6, 25.9)
120	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	82.7 (81.0, 84.9)	77.6 (75.7, 80.4)	75.6 (72.9, 77.8)	71.2 (68.0, 75.3)	70.8 (67.4, 75.5)	69.8 (65.1, 74.7)
	ΔAUC	14.9 (12.6, 16.7)	20.0 (17.1, 21.9)	22.1 (19.9, 24.8)	26.5 (22.5, 29.6)	26.9 (22.2, 30.6)	27.8 (22.8, 32.9)
150	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	78.6 (76.4, 80.9)	73.2 (70.5, 76.2)	71.4 (68.4, 74.3)	68.9 (64.7, 72.9)	68.5 (64.4, 73.5)	69.0 (32.8, 74.5)
	ΔAUC	19.3 (16.7, 21.2)	24.5 (21.3, 27.3)	26.2 (23.4, 29.2)	28.7 (24.8, 33.0)	29.1 (24.1, 33.4)	28.5 (23.4, 64.9)
180	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	77.0 (74.2, 79.3)	71.4 (68.4, 74.3)	69.8 (66.6, 72.8)	67.8 (63.0, 73.2)	68.0 (63.3, 72.9)	67.9 (33.0, 72.9)
	ΔAUC	20.8 (18.3, 23.6)	26.4 (23.5, 29.3)	27.8 (24.9, 31.0)	29.8 (24.6, 34.8)	29.6 (24.7, 34.5)	29.8 (24.9, 64.7)

Table A.25: Group-wise AUC and ΔAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 6$.

		τ_1					
ϕ	**Group**	**5.0**	**5.4**	**5.8**	**6.2**	**6.6**	**7.0**
0	a=0	97.6 (97.0, 98.1)	90.9 (89.5, 92.4)	85.8 (83.6, 88.3)	72.3 (68.5, 77.0)	69.0 (64.6, 73.6)	62.3 (55.8, 67.7)
	a=1	97.6 (97.2, 98.2)	99.7 (99.5, 99.8)	99.9 (99.8, 99.9)	100.0 (99.9, 100.0)	100.0 (99.9, 100.0)	99.9 (99.8, 100.0)
	ΔxAUC	0.4 (0.1, 1.1)	8.8 (7.1, 10.2)	14.1 (11.6, 16.3)	27.6 (23.0, 31.4)	30.9 (26.3, 35.4)	37.6 (32.0, 44.2)
30	a=0	95.1 (94.1, 96.1)	88.4 (86.4, 90.6)	83.5 (81.2, 86.4)	71.1 (66.4, 75.9)	68.1 (63.8, 73.7)	62.1 (56.5, 67.9)
	a=1	98.5 (98.0, 98.8)	99.3 (99.0, 99.4)	99.6 (99.3, 99.7)	99.9 (99.8, 99.9)	99.9 (99.8, 100.0)	99.9 (99.8, 100.0)
	ΔxAUC	3.3 (2.3, 4.5)	10.9 (8.7, 13.0)	16.1 (12.9, 18.4)	28.8 (23.9, 33.5)	31.9 (26.2, 36.2)	37.8 (31.9, 43.4)
60	a=0	90.4 (88.7, 92.0)	82.9 (80.1, 85.7)	79.3 (76.3, 81.7)	69.9 (64.3, 74.9)	66.6 (60.5, 72.3)	61.6 (56.1, 67.8)
	a=1	98.5 (98.1, 99.0)	99.0 (98.6, 99.3)	99.2 (98.9, 99.4)	99.7 (99.5, 99.8)	99.8 (99.6, 99.9)	99.9 (99.7, 100.0)
	ΔxAUC	8.2 (6.3, 10.5)	16.1 (13.0, 19.1)	20.0 (17.4, 23.1)	30.8 (25.5, 36.8)	33.3 (27.3, 39.4)	38.3 (31.9, 43.9)
90	a=0	89.0 (86.9, 90.8)	81.6 (79.4, 84.7)	75.4 (71.3, 78.9)	66.8 (59.3, 71.6)	65.5 (58.9, 70.8)	61.3 (54.3, 66.9)
	a=1	98.6 (98.2, 99.0)	99.0 (98.6, 99.2)	99.1 (98.8, 99.4)	99.7 (99.4, 99.8)	99.7 (99.5, 99.9)	99.9 (99.7, 100.0)
	ΔxAUC	9.7 (7.7, 11.8)	17.4 (14.1, 19.6)	23.8 (20.1, 28.0)	32.8 (27.8, 40.3)	34.2 (28.8, 41.0)	38.6 (32.7, 45.6)
120	a=0	81.8 (79.1, 85.0)	75.2 (71.3, 79.1)	71.9 (66.7, 76.1)	65.1 (59.1, 70.9)	63.6 (57.3, 70.2)	60.0 (53.4, 66.2)
	a=1	98.5 (98.1, 98.9)	99.0 (98.5, 99.3)	99.2 (98.8, 99.4)	99.6 (99.4, 99.8)	99.8 (99.5, 99.9)	99.9 (99.7, 100.0)
	ΔxAUC	16.8 (13.3, 19.6)	23.9 (19.7, 27.7)	27.4 (23.0, 32.6)	34.6 (28.5, 40.7)	36.3 (29.4, 42.5)	39.9 (33.5, 46.5)
150	a=0	78.8 (76.3, 82.8)	72.5 (67.9, 77.0)	69.5 (65.1, 74.6)	63.7 (57.9, 69.4)	63.7 (57.9, 69.4)	59.0 (52.2, 65.8)
	a=1	98.6 (98.1, 99.0)	99.1 (98.6, 99.4)	99.3 (98.9, 99.6)	99.7 (99.5, 99.9)	99.7 (99.5, 99.9)	99.9 (99.7, 100.0)
	ΔxAUC	19.6 (15.7, 22.3)	26.6 (21.9, 31.1)	29.9 (24.6, 34.3)	36.1 (30.1, 41.9)	36.1 (30.1, 41.9)	40.9 (33.8, 47.8)
180	a=0	78.4 (75.4, 81.6)	71.6 (67.6, 76.7)	68.8 (64.1, 73.0)	63.4 (57.7, 68.5)	61.9 (55.8, 67.4)	58.7 (52.6, 64.9)
	a=1	98.5 (98.1, 98.9)	99.1 (98.6, 99.3)	99.4 (99.0, 99.6)	99.8 (99.5, 99.9)	99.8 (99.6, 100.0)	99.9 (99.7, 100.0)
	ΔxAUC	20.3 (16.7, 23.2)	27.5 (22.2, 31.6)	30.6 (26.3, 35.2)	36.4 (30.9, 42.2)	38.0 (32.2, 44.1)	41.2 (35.0, 47.4)

Table A.26: Group-wise xAUC and ΔxAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 6$.

		τ_1					
ϕ	**Group**	**5.0**	**5.4**	**5.8**	**6.2**	**6.6**	**7.0**
0	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	97.6 (97.3, 97.9)	97.5 (97.1, 97.8)	97.3 (96.8, 97.6)	97.7 (97.3, 98.1)	93.7 (92.1, 94.9)	83.4 (77.2, 86.9)
	ΔAUC	0.2 (0.0, 0.6)	0.2 (0.0, 0.7)	0.4 (0.0, 1.0)	2.4 (1.6, 3.6)	4.0 (2.9, 5.7)	14.3 (10.8, 20.5)
30	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	96.2 (95.7, 96.9)	93.9 (92.8, 94.8)	92.8 (91.1, 94.1)	90.8 (89.2, 92.0)	89.8 (87.5, 91.3)	81.4 (76.4, 85.6)
	ΔAUC	1.5 (0.8, 2.0)	3.8 (2.6, 4.9)	4.9 (3.3, 6.6)	7.0 (5.7, 8.6)	7.1 (5.7, 9.1)	16.4 (12.1, 21.4)
60	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	91.5 (90.2, 92.7)	87.3 (85.5, 88.8)	85.2 (83.0, 87.1)	81.1 (78.4, 83.5)	80.3 (77.3, 82.6)	76.5 (72.5, 79.9)
	ΔAUC	6.1 (5.0, 7.3)	10.4 (8.8, 12.4)	12.4 (10.4, 14.6)	16.5 (14.2, 19.5)	17.5 (14.7, 20.3)	21.2 (17.9, 25.1)
90	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	84.9 (83.2, 86.9)	79.9 (77.5, 81.9)	77.7 (74.9, 79.9)	73.2 (70.0, 76.7)	72.9 (69.6, 75.9)	71.6 (66.3, 75.1)
	ΔAUC	12.8 (10.9, 14.5)	17.8 (15.8, 20.0)	20.0 (17.7, 22.9)	24.5 (21.2, 27.6)	24.7 (21.8, 28.0)	26.1 (22.6, 31.4)
120	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	78.2 (75.7, 80.9)	73.1 (69.4, 76.1)	71.1 (67.7, 73.6)	69.0 (64.0, 73.0)	68.5 (63.8, 73.1)	69.5 (58.7, 73.5)
	ΔAUC	19.5 (17.0, 22.0)	24.6 (21.6, 28.4)	26.5 (24.0, 30.0)	28.7 (24.7, 33.7)	29.1 (24.7, 34.0)	28.1 (24.4, 38.7)
150	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	71.0 (68.3, 74.8)	73.2 (70.5, 76.2)	66.2 (61.9, 70.6)	67.0 (61.0, 72.6)	67.3 (31.5, 72.4)	68.1 (32.4, 73.7)
	ΔAUC	26.7 (23.0, 29.5)	30.6 (26.9, 34.7)	31.4 (26.9, 36.0)	30.7 (25.0, 36.5)	30.4 (25.2, 66.4)	29.5 (24.0, 65.5)
180	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	67.7 (63.4, 70.7)	63.6 (59.9, 68.4)	63.8 (59.0, 69.0)	65.8 (35.2, 70.8)	66.3 (31.7, 72.3)	67.2 (31.0, 72.7)
	ΔAUC	30.0 (27.0, 34.3)	34.3 (29.3, 38.0)	33.8 (28.7, 38.6)	31.7 (26.6, 62.6)	31.6 (25.4, 66.0)	30.5 (24.8, 66.6)

Table A.27: Group-wise AUC and ΔAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 8$.

		τ_1					
ϕ	**Group**	**5.0**	**5.4**	**5.8**	**6.2**	**6.6**	**7.0**
0	a=0	97.6 (97.0, 98.1)	90.9 (89.5, 92.4)	85.8 (83.6, 88.3)	72.3 (68.5, 77.0)	69.0 (64.6, 73.6)	62.3 (55.8, 67.7)
	a=1	97.6 (97.2, 98.2)	99.7 (99.5, 99.8)	99.9 (99.8, 99.9)	100.0 (99.9, 100.0)	100.0 (99.9, 100.0)	99.9 (99.8, 100.0)
	ΔxAUC	0.4 (0.1, 1.1)	8.8 (7.1, 10.2)	14.1 (11.6, 16.3)	27.6 (23.0, 31.4)	30.9 (26.3, 35.4)	37.6 (32.0, 44.2)
30	a=0	94.6 (93.6, 95.5)	87.9 (85.6, 89.8)	82.8 (80.7, 85.9)	70.9 (66.2, 75.6)	68.2 (62.9, 73.9)	62.4 (56.7, 68.1)
	a=1	98.5 (98.1, 98.9)	99.2 (98.9, 99.4)	99.5 (99.3, 99.7)	99.9 (99.8, 99.9)	99.9 (99.8, 100.0)	99.9 (99.8, 100.0)
	ΔxAUC	3.9 (2.9, 5.2)	11.5 (9.3, 13.6)	16.6 (13.6, 18.9)	29.0 (24.2, 33.7)	31.7 (25.9, 37.0)	37.5 (31.7, 43.2)
60	a=0	89.1 (86.8, 90.6)	81.8 (78.7, 84.4)	77.9 (74.5, 80.3)	68.4 (61.9, 72.9)	66.2 (60.5, 71.1)	62.0 (55.6, 67.4)
	a=1	98.6 (98.1, 98.9)	99.0 (98.6, 99.3)	99.2 (98.8, 99.4)	99.7 (99.4, 99.8)	99.8 (99.6, 99.9)	99.9 (99.7, 100.0)
	ΔxAUC	9.5 (7.6, 11.9)	17.2 (14.4, 20.5)	21.2 (18.6, 24.9)	31.3 (26.6, 37.7)	33.5 (28.5, 39.3)	38.0 (32.4, 44.3)
90	a=0	83.6 (80.8, 86.0)	76.5 (72.6, 79.7)	73.1 (67.8, 76.2)	66.1 (59.3, 70.5)	64.6 (58.2, 69.7)	60.5 (54.2, 66.1)
	a=1	98.5 (98.1, 98.9)	98.9 (98.5, 99.2)	99.1 (98.8, 99.4)	99.6 (99.4, 99.8)	99.7 (99.5, 99.9)	99.9 (99.7, 100.0)
	ΔxAUC	15.0 (12.2, 17.6)	22.4 (19.0, 26.4)	26.1 (22.2, 31.5)	33.5 (28.9, 40.3)	35.2 (29.8, 41.5)	39.4 (33.6, 45.6)
120	a=0	79.0 (75.1, 81.9)	72.3 (67.0, 76.5)	69.4 (63.4, 74.0)	63.8 (57.9, 69.5)	62.1 (55.9, 67.5)	58.9 (52.6, 65.0)
	a=1	98.6 (98.1, 99.0)	99.1 (98.7, 99.4)	99.3 (98.9, 99.6)	99.8 (99.5, 99.9)	99.8 (99.6, 100.0)	99.9 (99.7, 100.0)
	ΔxAUC	19.7 (16.3, 23.6)	26.9 (22.5, 32.0)	29.8 (25.2, 36.2)	35.9 (30.1, 42.1)	36.3 (29.4, 42.5)	41.1 (34.8, 47.4)
150	a=0	75.1 (71.4, 79.3)	68.6 (63.4, 73.3)	66.3 (61.0, 71.5)	60.7 (54.6, 66.0)	59.4 (52.3, 65.4)	58.0 (51.9, 63.8)
	a=1	98.8 (98.3, 99.1)	99.4 (99.1, 99.6)	99.6 (99.3, 99.8)	99.9 (99.7, 100.0)	99.9 (99.7, 100.0)	99.9 (99.8, 100.0)
	ΔxAUC	19.6 (15.7, 22.3)	30.7 (25.9, 35.8)	33.4 (27.9, 38.8)	39.2 (33.7, 45.3)	40.5 (34.4, 47.7)	41.9 (36.1, 48.1)
180	a=0	74.0 (70.3, 77.8)	67.5 (62.1, 72.2)	64.4 (58.9, 69.7)	59.4 (53.0, 65.2)	58.3 (51.6, 63.5)	57.7 (50.7, 63.7)
	a=1	98.9 (98.4, 99.3)	99.5 (99.2, 99.7)	99.7 (99.5, 99.9)	99.9 (99.7, 100.0)	99.9 (99.7, 100.0)	99.9 (99.8, 100.0)
	ΔxAUC	24.9 (20.7, 29.0)	31.9 (27.1, 37.5)	35.2 (29.8, 40.9)	40.5 (34.5, 46.9)	41.7 (36.3, 48.3)	42.2 (36.1, 49.3)

Table A.28: Group-wise xAUC and ΔxAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 8$.

		τ_1					
ϕ	**Group**	**5.0**	**5.4**	**5.8**	**6.2**	**6.6**	**7.0**
0	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	97.6 (97.3, 97.9)	97.5 (97.1, 97.8)	97.3 (96.8, 97.6)	97.7 (97.3, 98.1)	93.7 (92.1, 94.9)	83.4 (77.2, 86.9)
	ΔAUC	0.2 (0.0, 0.6)	0.2 (0.0, 0.7)	0.4 (0.0, 1.0)	2.4 (1.6, 3.6)	4.0 (2.9, 5.7)	14.3 (10.8, 20.5)
30	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	95.9 (95.0, 96.6)	93.2 (91.9, 94.4)	92.0 (90.3, 93.2)	989.4 (87.7, 91.2)	88.7 (86.1, 90.3)	81.2 (76.0, 84.5)
	ΔAUC	1.8 (1.1, 2.7)	4.5 (3.2, 5.9)	5.7 (4.5, 7.4)	8.2 (6.6, 10.0)	9.0 (7.3, 11.5)	16.5 (13.1, 21.7))
60	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	89.7 (88.5, 91.4)	85.3 (83.1, 87.1)	83.1 (80.5, 85.5)	78.8 (76.0, 81.0)	78.1 (75.1, 80.5)	74.9 (71.0, 78.5)
	ΔAUC	8.0 (6.3, 9.2)	12.4 (10.5, 14.5)	14.6 (12.3, 17.1)	18.7 (16.7, 21.5)	19.6 (17.3, 22.6)	22.8 (19.3, 26.6)
90	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	81.3 (79.2, 83.1)	76.0 (73.4, 78.3)	73.8 (70.8, 76.2)	71.1 (67.1, 73.2)	70.3 (65.2, 72.9)	70.0 (65.0, 73.9)
	ΔAUC	16.4 (14.6, 18.6)	21.6 (19.2, 24.2)	24.0 (21.5, 26.6)	26.7 (24.2, 30.7)	27.4 (24.7, 32.5)	27.7 (24.0, 32.4)
120	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	72.4 (69.2, 75.7)	68.1 (65.2, 71.9)	67.6 (63.0, 71.8)	68.4 (61.7, 71.8)	68.5 (63.8, 73.1)	68.9 (31.4, 72.8)
	ΔAUC	25.4 (22.2, 28.5)	29.6 (25.7, 32.6)	30.4 (26.7, 33.9)	30.0 (26.1, 34.6)	29.3 (26.0, 35.7)	28.9 (24.9, 66.1)
150	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	64.4 (60.8, 68.5)	66.2 (57.1, 70.2)	66.8 (32.9, 71.1)	67.8 (31.0, 71.6)	67.8 (31.0, 71.6)	67.8 (31.0, 71.6)
	ΔAUC	33.2 (28.9, 37.0)	31.4 (27.1, 40.0)	31.0 (26.4, 64.7)	30.1 (26.0, 66.7)	30.1 (26.0, 66.7)	30.1 (26.0, 66.7)
180	a=0	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)	97.7 (97.4, 98.0)
	a=1	66.0 (31.4, 72.1)	66.0 (31.4, 72.1)	63.8 (59.0, 69.0)	65.8 (35.2, 70.8)	66.3 (31.7, 72.3)	67.2 (31.0, 72.7)
	ΔAUC	31.7 (25.6, 66.2)	31.7 (25.6, 66.2)	31.7 (25.6, 66.2)	31.7 (25.6, 66.2)	31.7 (25.6, 66.2)	31.7 (25.6, 66.2)

Table A.29: Group-wise AUC and ΔAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 10$.

		τ_1					
ϕ	**Group**	**5.0**	**5.4**	**5.8**	**6.2**	**6.6**	**7.0**
0	a=0	97.6 (97.0, 98.1)	90.9 (89.5, 92.4)	85.8 (83.6, 88.3)	72.3 (68.5, 77.0)	69.0 (64.6, 73.6)	62.3 (55.8, 67.7)
	a=1	97.6 (97.2, 98.2)	99.7 (99.5, 99.8)	99.9 (99.8, 99.9)	100.0 (99.9, 100.0)	100.0 (99.9, 100.0)	99.9 (99.8, 100.0)
	ΔxAUC	0.4 (0.1, 1.1)	8.8 (7.1, 10.2)	14.1 (11.6, 16.3)	27.6 (23.0, 31.4)	30.9 (26.3, 35.4)	37.6 (32.0, 44.2)
30	a=0	94.0 (93.0, 95.0)	87.1 (85.1, 89.2)	82.2 (79.6, 85.1)	70.6 (65.8, 75.3)	67.8 (62.4, 73.5)	62.2 (56.7, 67.9)
	a=1	98.5 (98.1, 98.9)	99.1 (98.8, 99.4)	99.4 (99.2, 99.6)	99.9 (99.7, 99.9)	99.9 (99.8, 100.0)	99.9 (99.8, 100.0)
	ΔxAUC	4.5 (3.4, 5.8)	12.0 (9.8, 14.0)	17.3 (14.2, 19.8)	29.2 (24.4, 34.1)	32.1 (26.3, 37.5)	37.6 (31.8, 43.2)
60	a=0	87.4 (85.4, 89.5)	80.4 (77.2, 83.0)	77.9 (74.5, 80.3)	67.7 (61.4, 72.3)	66.2 (59.3, 70.8)	61.7 (55.9, 67.9)
	a=1	98.6 (98.1, 98.9)	98.9 (98.6, 99.2)	99.1 (98.8, 99.4)	99.6 (99.5, 99.8)	99.8 (99.5, 99.9)	99.9 (99.7, 100.0)
	ΔxAUC	11.1 (9.1, 13.3)	18.8 (15.5, 22.1)	22.4 (18.9, 26.1)	31.9 (27.2, 38.4)	33.5 (28.8, 40.5)	38.1 (31.9, 44.0)
90	a=0	81.0 (77.5, 83.8)	74.1 (70.1, 77.7)	71.2 (66.4, 75.6)	64.7 (57.9, 70.0)	63.2 (57.0, 69.3)	59.8 (52.5, 65.5)
	a=1	98.5 (98.1, 98.9)	99.0 (98.5, 99.2)	99.2 (98.8, 99.4)	99.7 (99.5, 99.9)	99.8 (99.6, 99.9)	99.9 (99.7, 100.0)
	ΔxAUC	17.6 (14.2, 21.1)	22.4 (19.0, 26.4)	27.9 (23.5, 32.8)	35.0 (29.5, 41.8)	36.5 (30.2, 42.9)	40.1 (34.3, 47.4)
120	a=0	75.5 (71.5, 79.6)	69.2 (63.2, 74.1)	66.4 (58.9, 71.5)	60.7 (53.8, 66.7)	59.5 (51.8, 65.5)	58.2 (49.7, 63.4)
	a=1	98.8 (98.2, 99.2)	99.4 (98.9, 99.6)	99.6 (99.2, 99.8)	99.9 (99.6, 100.0)	99.9 (99.7, 100.0)	99.9 (99.7, 100.0)
	ΔxAUC	23.2 (18.8, 27.9)	30.1 (24.9, 36.5)	33.1 (28.0, 40.7)	39.2 (32.8, 46.1)	40.3 (34.3, 48.1)	41.7 (36.3, 50.3)
150	a=0	67.9 (62.8, 72.4)	61.0 (54.3, 67.3)	58.7 (50.6, 64.8)	60.7 (54.6, 66.0)	59.4 (52.3, 65.4)	57.3 (50.5, 63.9)
	a=1	99.5 (99.1, 99.7)	99.9 (99.6, 100.0)	99.6 (99.3, 99.8)	99.9 (99.7, 100.0)	99.9 (99.8, 100.0)	99.9 (99.8, 100.0)
	ΔxAUC	31.6 (26.9, 37.0)	38.8 (32.4, 45.7)	33.4 (27.9, 38.8)	41.2 (35.0, 49.4)	42.5 (35.9, 49.5)	42.5 (35.9, 49.5)
180	a=0	57.2 (48.3, 63.9)	57.2 (48.3, 63.9)	57.2 (48.3, 63.9)	57.2 (48.3, 63.9)	57.2 (48.3, 63.9)	57.2 (48.3, 63.9)
	a=1	99.9 (99.7, 100.0)	99.9 (99.7, 100.0)	99.9 (99.7, 100.0)	99.9 (99.7, 100.0)	99.9 (99.7, 100.0)	99.9 (99.7, 100.0)
	ΔxAUC	42.7 (35.8, 51.7)	42.7 (35.8, 51.7)	42.7 (35.8, 51.7)	42.7 (35.8, 51.7)	42.7 (35.8, 51.7)	42.7 (35.8, 51.7)

Table A.30: Group-wise xAUC and ΔxAUC under conditional distributional difference for all values of τ_1, ϕ, at $d' = 10$.

Intergroup mean diff. ($\Delta\mu$)	Group	AUC	xAUC
0	$a = 0$	95.8 (95.2, 96.2)	95.7 (95.0, 96.2)
	$a = 1$	95.8 (95.2, 96.3)	95.9 (95.2, 96.3)
	Δ	0.1 (0.0, 0.4)	0.2 (0.0, 0.5)
0.1	$a = 0$	94.7 (93.8, 95.4)	97.1 (96.5, 97.5)
	$a = 1$	95.8 (95.0, 96.3)	92.4 (91.1, 93.2)
	Δ	1.0 (0.7, 1.5)	4.8 (4.2, 5.6)
0.2	$a = 0$	95.1 (94.0, 95.7)	98.6 (98.3, 98.8)
	$a = 1$	96.8 (96.2, 97.3)	89.5 (87.8, 90.7)
	Δ	1.8 (1.3, 2.3)	9.2 (8.0, 10.6)
0.3	$a = 0$	96.5 (95.8, 97.1)	99.6 (99.5, 99.7)
	$a = 1$	98.5 (98.1, 98.7)	88.2 (86.0, 89.9)
	Δ	1.9 (1.5, 2.5)	11.4 (9.8, 13.4)

Table A.31: AUC, xAUC with empirical 95% CIs at varying settings of $\Delta\mu$, $\tau_0 = 5, \tau_1 = 6.6$.

Within-group var. (σ^2)	Group	AUC	xAUC
0.05	$a = 0$	93.7 (91.6, 95.5)	98.2 (97.5, 98.7)
	$a = 1$	94.1 (92.7, 95.3)	81.0 (78.3, 83.5)
	Δ	0.5 (0.0, 1.3)	17.2 (15.2, 19.2)
0.1	$a = 0$	95.9 (95.1, 96.6)	98.4 (98.1, 98.7)
	$a = 1$	95.7 (95.0, 96.4)	89.0 (87.3, 90.4)
	Δ	0.2 (0.0, 0.5)	9.4 (8.2, 10.7)
0.15	$a = 0$	96.4 (95.9, 96.9)	98.3 (98.0, 98.6)
	$a = 1$	96.1 (95.5, 96.6)	91.8 (90.5, 92.9)
	Δ	0.3 (0.0, 0.7)	6.5 (5.6, 7.5)
0.2	$a = 0$	96.7 (96.3, 97.2)	98.2 (97.9, 98.6)
	$a = 1$	96.4 (95.9, 97.0)	93.4 (92.5, 94.4)
	Δ	0.2 (0.0, 0.6)	4.8 (4.1, 5.4)

Table A.32: AUC, xAUC with empirical 95% CIs at varying settings of σ^2, $\tau_0 = 5, \tau_1 = 6.6$.

A.3 Supplementary material for "From Biased Selective Labels to Pseudo-Labels: An Expectation-Maximization Framework for Learning from Biased Decisions"

A.3.1 Selective labeling in the literature

We enumerate domains in which our literature review found instances of selective label problems in the ML methods and applications literature:

- **Healthcare**: Laboratory/diagnostic testing [42, 197] and diagnosis [78, 253, 14]
- **Social & public policy:** Child welfare assessment [244, 145], urban planning/policy [151, 163, 177], hiring pipelines [219, 265], student placement [22], and bias in policing [233, 222]
- **Finance:** Credit repayment [29] and financial auditing [112]
- **Others/miscellaneous:** Wildlife conservation [93], social media content moderation [27]

We note that this is not an exhaustive list of all papers in the selective labeling literature or related problem settings. However, this list illustrates the broad applicability and relevance of our problem setting.

A.3.2 Omitted Proofs

For convenience, we restate all theorems and propositions here.

A.3.2.1 Theorem 3.1

Theorem (E-step derivation)**.** *The posterior conditional mean of* $y^{(i)} = 1$ *given the observed data,* $Q(y^{(i)}) \triangleq \mathbb{E}[y^{(i)} = 1 \mid t^{(i)}, \tilde{y}^{(i)}, a^{(i)}, x^{(i)}]$*, is equal to*

$$Q(y^{(i)}) = \begin{cases} \tilde{y}^{(i)} & t^{(i)} = 1 \\ \mathbb{E}[y^{(i)} = 1 \mid x^{(i)}] & \textit{otherwise} \end{cases}. \tag{A.20}$$

Proof. We drop superscripts $(\cdot)^{(i)}$ in the proof for clarity. Denote $Q(y)$ as posterior distribution of y given the observed data, $\mathbb{E}[y = 1 \mid t, \tilde{y}, a, x]$ (*i.e.*, the E-step estimate). First, we can write

$$Q(y) \triangleq \mathbb{E}[y = 1 \mid t, \tilde{y}, a, x] = \mathbb{E}[y = 1 \mid \tilde{y}, x, t] = P(y = 1 \mid \tilde{y}, x, t) \tag{A.21}$$

for simplicity, where we use the fact that $Y \perp\!\!\!\perp A \mid (T, X)$ and the fact that $\mathbb{E}[y = 1 \mid \tilde{y}, x, t]$ is binary. Proceeding, we can write:

$$= t \cdot P(y = 1 \mid \tilde{y}, x, t = 1) + (1 - t) \cdot P(y = 1 \mid \tilde{y}, x, t = 0) \tag{A.22}$$

$$= t\tilde{y} + (1 - t)[\tilde{y}P(y = 1 \mid \tilde{y} = 1, x, t = 0) + (1 - \tilde{y})P(y = 1 \mid \tilde{y} = 0, x, t = 0)] \tag{A.23}$$

$$= t\tilde{y} + (1 - t)(1 - \tilde{y})P(y = 1 \mid \tilde{y} = 0, x, t = 0) \tag{A.24}$$

$$= t\tilde{y} + (1 - t)(1 - \tilde{y}) \cdot \frac{P(\tilde{y} = 0 \mid y = 1, x, t = 0)P(y = 1 \mid x, t = 0)}{P(\tilde{y} = 0 \mid x, t = 0)} \tag{A.25}$$

$$= t\tilde{y} + (1 - t)(1 - \tilde{y})P(y = 1 \mid x). \tag{A.26}$$

The second equality follows since $t = 1 \implies \tilde{y} = y$. The third equality holds since $P(y = 1 \mid \tilde{y} = 1, x, t = 0) = 0$ by construction, since $\tilde{y} = yt$ under disparate censorship. The final step follows from three facts: (1) $P(\tilde{y} = 0 \mid y = 1, x, t = 0) = 1$ for all x, (2) $P(\tilde{y} = 0 \mid x, t = 0) = 1$ for all x, and (3) $Y \perp\!\!\!\perp T \mid X$. This is more succinctly rewritten as E-step is:

$$Q(y) = \begin{cases} \tilde{y} & t = 1 \\ \mathbb{E}[y = 1 \mid x] & \text{otherwise (i.e., } \tilde{y} = 0 \wedge t = 0) \end{cases}, \tag{A.27}$$

which is what we wanted to show. □

Remark 22. *Since $y^{(i)}$ is binary by assumption, this result fully determines the posterior distribution since $\mathbb{E}[y = 1 \mid \tilde{y}, x, t] = 1 - \mathbb{E}[y = 0 \mid \tilde{y}, x, t]$.*

A.3.2.2 Theorem 3.2

Theorem (M-step derivation)**.** *Let $P(U, A, X, T, \tilde{Y}; \theta)$ be a model for the joint data distribution parameterized by some arbitrary $\theta \in \Theta$ in some parameter space Θ, which factorizes according to the disparate censorship DAG =(Fig. 5.1). Let $Q(y) \triangleq \mathbb{E}[y = 1 \mid t, \tilde{y}, a, x]$ be the posterior expectation that $y = 1$ given the observed data. Then (replacing random variables with their realized counterparts), we have*

$$\begin{aligned} \max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log P(u^{(i)}, a^{(i)}, x^{(i)}, t^{(i)}, \tilde{y}^{(i)}; \theta) \geq \max_{\theta} \frac{1}{N} \sum_{i=1}^{N} & Q(y^{(i)}) \log P(y^{(i)} \mid x^{(i)}; \theta_{Y|X}) \\ & + (1 - Q(y^{(i)})) \log(1 - P(y^{(i)} \mid x^{(i)}; \theta_{Y|X})) \\ & + Q(y^{(i)}) \log P(\tilde{y}^{(i)} \mid y^{(i)}, t^{(i)}; \theta_{\tilde{Y}|Y,T}) \end{aligned} \tag{A.28}$$

where $\theta = [\theta_{Y|X} \ \theta_{\tilde{Y}|Y,T}]^\top$.

Proof. We first construct the evidence-based lower bound (ELBO) of the LHS in the theorem statement. First, for a single example indexed by i, we can write:

$$\log P(u^{(i)}, a^{(i)}, x^{(i)}, t^{(i)}, \tilde{y}^{(i)}; \theta) = \log \sum_{y^{(i)} \in \{0,1\}} Q(y^{(i)}) \frac{P(u^{(i)}, a^{(i)}, x^{(i)}, t^{(i)}, \tilde{y}^{(i)}, y^{(i)}; \theta)}{Q(y^{(i)})} \tag{A.29}$$

$$\geq \sum_{y^{(i)} \in \{0,1\}} Q(y^{(i)}) \log \frac{P(u^{(i)}, a^{(i)}, x^{(i)}, t^{(i)}, \tilde{y}^{(i)}, y^{(i)}; \theta)}{Q(y^{(i)})} \tag{A.30}$$

via Jensen's inequality. Then, we note that

$$\max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{y^{(i)} \in \{0,1\}} Q(y^{(i)}) \log \frac{P(u^{(i)}, a^{(i)}, x^{(i)}, t^{(i)}, \tilde{y}^{(i)}, y^{(i)}; \theta)}{Q(y^{(i)})}$$

$$= \max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{y^{(i)} \in \{0,1\}} Q(y^{(i)}) \log P(u^{(i)}, a^{(i)}, x^{(i)}, t^{(i)}, \tilde{y}^{(i)}, y^{(i)}; \theta), \tag{A.31}$$

dropping $Q(y^{(i)}) \log Q(y^{(i)})$, which is constant with respect to θ, after expanding the $\log$ term. We can then use the DAG to factorize the joint distribution of all variables (including latent variable Y), which is given by

$$P(\tilde{Y}, Y, T, X, A, U) = P(\tilde{Y} \mid Y, T) P(Y \mid X) P(T \mid X, A) P(X, A, U). \tag{A.32}$$

Note that we need only model the first two terms for estimation of $y^{(i)}$. The first two terms do not involve $y^{(i)}$, are not parameterized, and can be dropped from the maximization problem. Hence, we proceed to write

$$= \max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{y^{(i)} \in \{0,1\}} Q(y^{(i)}) \log P(y^{(i)} \mid x^{(i)}; \theta_{\tilde{Y}|X}) P(\tilde{y}^{(i)} \mid t^{(i)}, y^{(i)}; \theta_{\tilde{Y}|Y,T}) \tag{A.33}$$

where $\theta = [\theta_{Y|X} \ \theta_{\tilde{Y}|Y,T}]^\top$. This can be rewritten as

$$= \max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{y^{(i)} \in \{0,1\}} Q(y^{(i)}) \log P(y^{(i)} \mid x^{(i)}; \theta_{\tilde{Y}|X}) + Q(y^{(i)}) P(\tilde{y}^{(i)} \mid t^{(i)}, y^{(i)}; \theta_{\tilde{Y}|Y,T}), \tag{A.34}$$

at which point we note that it is sufficient to show that

$$(1 - Q(y^{(i)})) \log(1 - P(\tilde{y}^{(i)} \mid y^{(i)}, t^{(i)}; \theta_{\tilde{Y}|Y,T})) \tag{A.35}$$

is constant in θ. We can rewrite the above as

$$(1 - Q(y^{(i)}))[\tilde{y}^{(i)} \log(P(\tilde{y}^{(i)} = 1 \mid y^{(i)} = 0, t^{(i)}; \theta_{\tilde{Y}|Y,T})) \\ + (1 - \tilde{y}^{(i)}) \log(P(\tilde{y}^{(i)} = 0 \mid y^{(i)} = 0, t^{(i)}; \theta_{\tilde{Y}|Y,T}))]. \tag{A.36}$$

First, note that the event $\{\tilde{y}^{(i)} = 1 \mid y^{(i)} = 0\}$ occurs with probability zero by definition (recall $\tilde{y}^{(i)} = y^{(i)}t^{(i)}$). Thus, $P(\tilde{y}^{(i)} = 1 \mid y^{(i)} = 0, t^{(i)}; \theta_{\tilde{Y}|Y,T}))$ cannot change with respect to θ; we drop it from the maximization problem. Similarly, $P(\tilde{y}^{(i)} = 0 \mid y^{(i)} = 0, t^{(i)}) = 0$ by definition, so $(1 - Q(y^{(i)})) \log(1 - P(\tilde{y}^{(i)} \mid y^{(i)}, t^{(i)})) = 0$ which is constant as needed, from which the theorem follows. □

Remark 23. *In the theorem statement, replacing $P(y^{(i)} \mid x^{(i)}; \theta_{Y|X})$ with $\hat{y}^{(i)}$ and $P(\tilde{y}^{(i)} \mid y^{(i)}, t^{(i)}; \theta_{\tilde{Y}|Y,T})$ with $h_\phi(\cdot)$, assuming $y^{(i)}$ and $\tilde{y}^{(i)}$ are binary, and writing the explicit form of negative binary cross-entropy (*e.g., *$y \log y + (1-y) \log \hat{y}$*) recovers the form of the M-step objective seen in Eq. 3.9. Note that the optimization problem flips from a maximization to a minimization due to the relationship between maximizing log-likelihood of binary variable(s) and minimizing cross-entropy loss.*

A.3.2.3 Theorem 8

Theorem (Strength of the causal regularizer in $\hat{t}^{(i)}$)**.** *For an example indexed by i, $Q(y^{(i)}) \in [0, 1)$, and $J^{(i)}$ defined as in Eq. A.95, $R(J^{(i)})$ is monotonically increasing in $\hat{t}^{(i)}$ on $(0, 1]$.*

Proof. As a proof outline, we first show the closed-form of $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ by solving the first-order optimality condition of Eq. A.95. Then, we show that $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ is decreasing in $\hat{t}^{(i)}$, and attains a maximum of $Q(y^{(i)})$ as $\hat{t}^{(i)} \to 0^+$. We conclude by showing that the preceding implies that $R(J^{(i)})$ is monotonically increasing in $\hat{t}^{(i)}$ on $(0, 1]$, as desired.

The first-order optimality condition of Eq. A.95 is

$$-\left(\frac{Q(y^{(i)})}{\hat{y}^{(i)}} - \frac{1 - Q(y^{(i)})}{1 - \hat{y}^{(i)}}\right) + \frac{\hat{t}^{(i)} Q(y^{(i)})}{1 - \hat{y}^{(i)}\hat{t}^{(i)}} = 0. \tag{A.37}$$

By assumption (convexity of $\mathcal{L}$), the minimizer is unique. Some algebra yields

$$- Q(y^{(i)})(1 - \hat{y}^{(i)})(1 - \hat{y}^{(i)}\hat{t}^{(i)}) + (1 - Q(y^{(i)}))\hat{y}^{(i)}(1 - \hat{y}^{(i)}\hat{t}^{(i)}) \\ + \hat{t}^{(i)} Q(y^{(i)})\hat{y}^{(i)}(1 - \hat{y}^{(i)}) = 0 \tag{A.38}$$

$$\iff (\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})\hat{y}^{(i)2} - (2Q(y^{(i)})\hat{t}^{(i)} + 1)\hat{y}^{(i)} + Q(y^{(i)}) = 0, \tag{A.39}$$

from which we can apply the quadratic formula. Define $B(Q(y^{(i)}), \hat{t}^{(i)}) \triangleq 2Q(y^{(i)})\hat{t}^{(i)} + 1$ and

$D(Q(y^{(i)}), \hat{t}^{(i)}) \triangleq (-B(Q(y^{(i)}), \hat{t}^{(i)}))^2 - 4Q(y^{(i)})(Q(y^{(i)})\hat{t}^{(i)} + \hat{t}^{(i)})$.[4] The quadratic formula yields solutions

$$\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) = \frac{B(Q(y^{(i)}), \hat{t}^{(i)}) \pm \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}}{2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})}. \tag{A.40}$$

We use the fact that $\hat{y}^{(i)}$ must be in [0, 1] and the constraints that $\hat{t}^{(i)} \in (0, 1]$ and $Q(y^{(i)}) \in [0, 1)$ to determine which branch of Eq. A.40 yields real solutions in $[0, 1]$. By Lemma 27, $D(Q(y^{(i)}, \hat{t}^{(i)}) \geq 0$, so the solutions are real. Then, by Lemma 28,

$$\frac{B(Q(y^{(i)}), \hat{t}^{(i)}) + \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}}{2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})} \geq 1, \tag{A.41}$$

eliminating that branch. By elimination,

$$\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) = \frac{B(Q(y^{(i)}), \hat{t}^{(i)}) - \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}}{2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})}. \tag{A.42}$$

Lemma 29 verifies that the resulting $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ are in [0, 1], as needed. To proceed, it suffices to show that $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ is decreasing in $\hat{t}^{(i)}$ and attains a maximum of $Q(y^{(i)})$ as $\hat{t}^{(i)} \to 0^+$.

Applying Lemma 30, to prove that $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ decreases in $\hat{t}^{(i)}$, it is sufficient to show

$$1 - 2\hat{t}^{(i)}Q(y^{(i)})^2 - \sqrt{D(Q(y^{(i)}), \hat{t}^{(i)}))} < 0 \tag{A.43}$$

because $1 - 2\hat{t}^{(i)}Q(y^{(i)})^2 - \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}$ has the same sign as the derivative of $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ for the values of $(\hat{t}^{(i)}, Q(y^{(i)}))$ of interest.

For values of $\hat{t}^{(i)} \in (0, 1]$ and $Q(y^{(i)}) \in [0, 1)$ such that $1 - 2\hat{t}^{(i)}Q(y^{(i)})^2 < 0$, Lemma 27 yields the desired result. For the remaining values of $(\hat{t}^{(i)}, Q(y^{(i)}))$, we can write

$$1 - 4Q(y^{(i)})^2\hat{t}^{(i)} + 4Q(y^{(i)})^4\hat{t}^{(i)2} < D(Q(y^{(i)}), \hat{t}^{(i)}) \tag{A.44}$$

$$\iff 1 - 4Q(y^{(i)})^2\hat{t}^{(i)} + 4Q(y^{(i)})^4\hat{t}^{(i)2} < 1 - 4Q(y^{(i)})^2\hat{t}^{(i)} + 4Q(y^{(i)})^2\hat{t}^{(i)2} \tag{A.45}$$

$$\iff Q(y^{(i)})^2 < 1 \tag{A.46}$$

[4] We use letter $B(\cdot)$ because it corresponds to coefficient b in the conventional quadratic formula:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

for a quadratic polynomial $ax^2 + bx + c = 0$. We choose letter $D(\cdot)$ since it corresponds to the discriminant.

which holds for all feasible values of $Q(y^{(i)}) \in [0, 1)$. Lastly, due to the monotonicity of $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ in $\hat{t}^{(i)}$, the following one-sided limit is the maximum:

$$\lim_{\hat{t}^{(i)} \to 0^+} \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) = \max_{\hat{t}^{(i)} \in (0,1]} \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}). \tag{A.47}$$

We want to show that the limit is $Q(y^{(i)})$. Note that, since $\mathcal{L}$ is finite and convex, it is continuous (Corollary 10.1.1, [240]); hence, this limit exists. Since substituting $\hat{t}^{(i)} = 0$ yields the indeterminate form $0/0$, we appeal to L'Hôpital's rule:

$$\lim_{\hat{t}^{(i)} \to 0^+} \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) = \lim_{\hat{t}^{(i)} \to 0^+} \frac{2Q(y^{(i)}) - \frac{4\hat{t}^{(i)}Q(y^{(i)})^2 - 2Q(y^{(i)})^2}{\sqrt{D(Q(y^{(i)}), \hat{t}^{(i)})}}}{2(Q(y^{(i)}) + 1)} \tag{A.48}$$

$$= \frac{2Q(y^{(i)}) + 2Q(y^{(i)})^2}{2(Q(y^{(i)}) + 1)} = Q(y^{(i)}). \tag{A.49}$$

Note that $\sqrt{D(Q(y^{(i)}), \hat{t}^{(i)})} \Big|_{\hat{t}^{(i)}=0} = 1$. Since $Q(y^{(i)}) - \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) > 0$,

$$R(J^{(i)}) = |Q(y^{(i)}) - \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})| = Q(y^{(i)}) - \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}). \tag{A.50}$$

Furthermore, since $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ is decreasing in $\hat{t}^{(i)}$, $R(J^{(i)})$ must increase in $\hat{t}^{(i)}$, from which the theorem follows. □

Remark 24. *We comment on the potential for DCEM to improve robustness to low overlap. To do so, we analyze the sensitivity of the M-step optimum to extreme $\hat{t}^{(i)}$. While analyzing the asymptotic variance of consistent estimators is a common approach, asymptotic guarantees for DCEM are unclear due to the inherently non-convex (with respect to the parameters) objective function. Thus, we analyze the Lipschitzness of the M-step optimum versus other causal effect estimators. First, note that*

$$\frac{d}{d\hat{t}} \hat{y}_{OPT}(Q(y), \hat{t}) = \frac{1 - 2Q(y)^2\hat{t}^2 - \sqrt{C}}{2(Q(y) + 1)\hat{t}^2\sqrt{C}} \tag{A.51}$$

*where $C = 4Q(y)^2\hat{t}^2 - 4Q(y)^2\hat{t} + 1$. For all $Q(y^{(i)}) < 1$ and all $\hat{t^{(i)}}$, this derivative is bounded (*e.g.*, see Figure A.12), and is $O(1)-$Lipschitz in $\hat{t}^{(i)}$. However, consider the expression for an inverse-propensity-weighted estimator, which sums terms of the form*

$$\frac{y^{(i)}t^{(i)}}{\hat{t}^{(i)}} - \frac{y^{(i)}(1 - t^{(i)})}{1 - \hat{t}^{(i)}} \tag{A.52}$$

to obtain a final estimate. Eq. A.52 has $O(\hat{t}^{(i)2})$-Lipschitz terms with respect to $\hat{t}^{(i)}$. Thus, in a Lipschitz sense, DCEM may be less sensitive to extreme propensity scores than causal effect

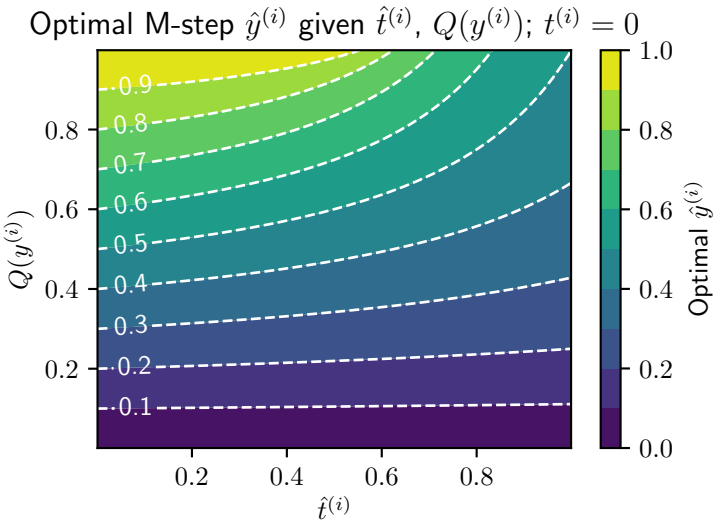


Figure A.12: Contour plot of $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ with respect to $\hat{t}^{(i)}$ (x-axis) and $Q(y^{(i)})$ (y-axis). $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$ scales with $Q(y^{(i)})$ but decreases in $\hat{t}^{(i)}$.

estimators such as IPW.

Corollary 25. *For an example indexed by i, $Q(y^{(i)}) = 1$, and $J^{(i)}$ defined as in Eq. A.95, $R(J^{(i)})$ is monotonically non-decreasing in $\hat{t}^{(i)}$ on $(0, 1]$.*

Proof. The proof is identical to that of Theorem 3.4, except we find that

$$\frac{\partial}{\partial \hat{t}^{(i)}} \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) \leq 0, \tag{A.53}$$

instead of being strictly less than zero, from which the corollary follows. □

Remark 26. *For intuition, we show a contour plot of $\hat{y}^{(i)}_{OPT}(Q(y^{(i)}), \hat{t}^{(i)})$ in Fig. A.12. We verify the result in CVXPY.*

A.3.2.4 Proposition 9

Proposition (Minimizer of M-step when $t^{(i)} = 1$)**.** *Suppose that $t^{(i)} = 1$, let $Q(y^{(i)}) \triangleq \mathbb{E}[y = 1 \mid t, \tilde{y}, a, x]$, and let $\hat{y}^{(i)}$ be some estimate of $y^{(i)}$. Use $\mathcal{L} : [0, 1]^2 \to \mathbb{R}_+$ be shorthand for binary cross-entropy loss. Then, the minimization problem*

$$\min_{\hat{y}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)}) + Q(y^{(i)}) \mathcal{L}(\tilde{y}^{(i)}, \hat{y}^{(i)} \hat{t}^{(i)}) \tag{A.54}$$

admits the solution $\hat{y}^{(i)} = y^{(i)}$ for all $i \in \{1, \ldots, N\}$.

Proof. We briefly verify the convexity of the objective, which follows from the convexity of binary cross-entropy loss and the closure of convexity under addition and positive scalar multiplication ($Q(y^{(i)}) \geq 0$). Thus, any minimizer of the objective is unique.

We proceed by cases. First, suppose that $y^{(i)} = 1$. Substituting the definition of $Q(y^{(i)})$, and using the fact that $t^{(i)} = 1 \implies y^{(i)} = \tilde{y}^{(i)}$), the objective function for a single example becomes

$$\mathcal{L}(1, \hat{y}^{(i)}) + \mathcal{L}(1, \hat{y}^{(i)}\hat{t}^{(i)}), \tag{A.55}$$

which, by inspection, is maximized for $\hat{y}^{(i)} = 1$. Similarly, for $y^{(i)} = 0$, the objective function for a single example is

$$\mathcal{L}(0, \hat{y}^{(i)}) \tag{A.56}$$

which reduces to binary cross-entropy loss, and $\hat{y}^{(i)} = 0$ minimizes the objective. Combining the two cases, the minimizer of the M-step objective when $t^{(i)} = 1$ is $\hat{y}^{(i)} = y^{(i)}$ as desired. □

A.3.2.5 Causal identifiability

For completeness, we provide the derivation of the causal identifiability results, though it follows directly from existing results [124].

Proposition. *Suppose that conditional exchangeability, or $\tilde{Y}(t) \perp\!\!\!\perp T \mid X$, holds. Then $\mathbb{E}[Y \mid X] = \mathbb{E}[\tilde{Y} \mid X, do(T = 1)]$, which is identifiable as $\mathbb{E}[\tilde{Y} \mid X, T = 1]$.*

Proof. We can write

$$\mathbb{E}[Y \mid X] = \mathbb{E}[Y \mid X, T = 1] = \mathbb{E}[\tilde{Y} \mid X, T = 1] = \mathbb{E}[\tilde{Y} \mid X, do(T = 1)] \tag{A.57}$$

where the first equality is due to $Y \perp\!\!\!\perp T \mid X$, the second equality results from $T = 1 \implies Y = \tilde{Y}$, and the final equality applies conditional exchangeability. Since $\mathbb{E}[Y \mid X] = \mathbb{E}[\tilde{Y} \mid X, do(T = 1)] = \mathbb{E}[\tilde{Y} \mid X, T = 1]$, the theorem follows. □

A.3.2.6 Lemmas used

Below are the lemmas and proofs referenced in the preceding theorem and proposition proofs.

Lemma 27. *Define $B(Q(y^{(i)}), \hat{t}^{(i)}) \triangleq 2Q(y^{(i)})\hat{t}^{(i)} + 1$ and $D(Q(y^{(i)}), \hat{t}^{(i)}) \triangleq (-B(Q(y^{(i)}), \hat{t}^{(i)}))^2 - 4Q(y^{(i)})(Q(y^{(i)})\hat{t}^{(i)} + \hat{t}^{(i)})$ on $Q(y^{(i)}) \in [0, 1]$ and $\hat{t}^{(i)} \in (0, 1]$. Then, $D(Q(y^{(i)}), \hat{t}^{(i)}) \geq 0$.*

Proof. Choose any $Q(y^{(i)}) \in [0,1]$ and $\hat{t}^{(i)} \in (0,1]$. We can write:

$$D(Q(y^{(i)}, \hat{t}^{(i)}) \geq 0 \tag{A.58}$$
$$\iff B(Q(y^{(i)}), \hat{t}^{(i)})^2 \geq 4Q(y^{(i)})(Q(y^{(i)})\hat{t}^{(i)} + \hat{t}^{(i)}) \tag{A.59}$$
$$\iff 4Q(y^{(i)})^2\hat{t}^{(i)2} + 4Q(y^{(i)})\hat{t}^{(i)} + 1 \geq 4Q(y^{(i)})(Q(y^{(i)})\hat{t}^{(i)} + \hat{t}^{(i)}) \tag{A.60}$$
$$\iff 4Q(y^{(i)})^2\hat{t}^{(i)2} - 4Q(y^{(i)})^2\hat{t}^{(i)} + 1 \geq 0. \tag{A.61}$$

The final LHS is convex (by inspection) in $\hat{t}^{(i)}$, such that

$$\min_{\hat{t}^{(i)}} 4Q(y^{(i)})^2\hat{t}^{(i)2} - 4Q(y^{(i)})^2\hat{t}^{(i)} + 1 = 1 - Q(y^{(i)}) \geq 0 \tag{A.62}$$

where the minimum is attained at $\hat{t}^{(i)} = \frac{1}{2}$, and concave (by inspection) in $Q(y^{(i)})$, such that it suffices to evaluate the final LHS at $Q(y^{(i)}) \in \{0,1\}$:[5]

$$4Q(y^{(i)})^2\hat{t}^{(i)2} - 4Q(y^{(i)})^2\hat{t}^{(i)} + 1\big|_{Q(y^{(i)})=0} = 1 \geq 0 \tag{A.63}$$
$$4Q(y^{(i)})^2\hat{t}^{(i)2} - 4Q(y^{(i)})^2\hat{t}^{(i)} + 1\big|_{Q(y^{(i)})=1} = 4(\hat{t}^{(i)2} - \hat{t}^{(i)}) + 1 \geq 4 \cdot \left(-\frac{1}{4}\right) + 1 = 0, \tag{A.64}$$

such that for all other $Q(y^{(i)}) \in (0,1)$, $4Q(y^{(i)})^2\hat{t}^{(i)2} - 4Q(y^{(i)})^2\hat{t}^{(i)} + 1 \geq 0$ as needed.

□

Lemma 28. *Define $B(Q(y^{(i)}), \hat{t}^{(i)})$ and $D(Q(y^{(i)}, \hat{t}^{(i)})$ as in Lemma 27. Then, for $Q(y^{(i)}) \in [0,1]$ and $\hat{t}^{(i)} \in (0,1]$,*

$$\frac{B(Q(y^{(i)}), \hat{t}^{(i)}) + \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}}{2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})} \geq 1. \tag{A.65}$$

Proof. Choose any $Q(y^{(i)}) \in [0,1]$ and $\hat{t}^{(i)} \in (0,1]$. First, we rewrite

$$\frac{B(Q(y^{(i)}), \hat{t}^{(i)}) + \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}}{2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})} \geq 1 \tag{A.66}$$
$$\iff 2Q(y^{(i)})\hat{t}^{(i)} + 1 + \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})} \geq 2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)}) \tag{A.67}$$
$$\iff \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})} \geq 2\hat{t}^{(i)} - 1. \tag{A.68}$$

[5]Recall that, for a concave function f, $f(\alpha x + (1-\alpha)y) \geq \alpha f(x) + (1-\alpha)f(y)$ for $\alpha \in [0,1]$ with equality for $\alpha = 0$ or 1. Thus, via the extreme value theorem, the minimum of f on $[x,y]$ is achieved at x or y.

For $\hat{t}^{(i)} \in (0, \frac{1}{2})$, Lemma 27 yields the desired conclusion. For $\hat{t}^{(i)} \in [\frac{1}{2}, 1]$, we can write

$$\sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})} \geq 2\hat{t}^{(i)} - 1 \tag{A.69}$$

$$\iff 4Q(y^{(i)2})\hat{t}^{(i)2} - 4Q(y^{(i)}2)\hat{t}^{(i)} + 1 \geq 4\hat{t}^{(i)2} - 4\hat{t}^{(i)} + 1 \tag{A.70}$$

$$\iff Q(y^{(i)2})(\hat{t}^{(i)} - 1) \geq \hat{t}^{(i)} - 1 \tag{A.71}$$

$$\iff Q(y^{(i)2}) \leq 1 \tag{A.72}$$

which all $Q(y^{(i)2}) \in [0, 1]$ satisfy. This completes the proof. □

Lemma 29. *Define $B(Q(y^{(i)}), \hat{t}^{(i)})$ and $D(Q(y^{(i)}, \hat{t}^{(i)})$ as in Lemma 27. Then, for $Q(y^{(i)}) \in [0, 1]$ and $\hat{t}^{(i)} \in (0, 1]$,*

$$0 \leq \frac{B(Q(y^{(i)}), \hat{t}^{(i)}) - \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}}{2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})} \leq 1. \tag{A.73}$$

Proof. Choose any $Q(y^{(i)}) \in [0, 1]$ and $\hat{t}^{(i)} \in (0, 1]$. Equivalently, we can show

$$0 \leq B(Q(y^{(i)}), \hat{t}^{(i)}) - \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})} \leq 2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)}). \tag{A.74}$$

For the first inequality, note that

$$\sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})} = \sqrt{(-B(Q(y^{(i)}), \hat{t}^{(i)}))^2 - 4Q(y^{(i)})(Q(y^{(i)})\hat{t}^{(i)} + \hat{t}^{(i)})} \leq \sqrt{(-B(Q(y^{(i)}), \hat{t}^{(i)}))^2} \tag{A.75}$$

$$= |B(Q(y^{(i)}), \hat{t}^{(i)}))| = B(Q(y^{(i)}), \hat{t}^{(i)})) \tag{A.76}$$

which rearranges to $0 \leq B(Q(y^{(i)}), \hat{t}^{(i)}) - \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}$ as desired. For the second inequality, note that

$$B(Q(y^{(i)}), \hat{t}^{(i)}) - \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})} \leq 2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)}) \tag{A.77}$$

$$\iff 1 - \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})} \leq 2\hat{t}^{(i)} \tag{A.78}$$

$$\iff 1 - 2\hat{t}^{(i)} \leq \sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})}. \tag{A.79}$$

For $\hat{t}^{(i)} \in [\frac{1}{2}, 1]$, Lemma 27 yields the desired conclusion. For $\hat{t}^{(i)} \in (0, \frac{1}{2})$, the proof proceeds

similarly to Lemma 28:

$$\sqrt{D(Q(y^{(i)}, \hat{t}^{(i)})} \geq 1 - 2\hat{t}^{(i)} \tag{A.80}$$

$$\iff 4Q(y^{(i)2})\hat{t}^{(i)2} - 4Q(y^{(i)})^2\hat{t}^{(i)} + 1 \geq 4\hat{t}^{(i)2} - 4\hat{t}^{(i)} + 1 \tag{A.81}$$

$$\iff Q(y^{(i)2})(\hat{t}^{(i)} - 1) \geq \hat{t}^{(i)} - 1 \tag{A.82}$$

$$\iff Q(y^{(i)2}) \leq 1 \tag{A.83}$$

which all $Q(y^{(i)2}) \in [0, 1]$ satisfy. This completes the proof. □

Lemma 30. *Define $B(Q(y^{(i)}), \hat{t}^{(i)})$ and $D(Q(y^{(i)}, \hat{t}^{(i)})$ as in Lemma 27, and define $\hat{y}^{(i)}_{OPT}(Q(y^{(i)}), \hat{t}^{(i)})$ as in Definition A.95. Then,*

$$\textit{sign}\left(\frac{\partial}{\partial \hat{t}^{(i)}}\, \hat{y}^{(i)}_{OPT}(Q(y^{(i)}), \hat{t}^{(i)})\right) = \textit{sign}\left(1 - 2\hat{t}^{(i)}Q(y^{(i)})^2 - \sqrt{D(Q(y^{(i)}), \hat{t}^{(i)}))}\right) \tag{A.84}$$

where $\textit{sign}(x) : \mathbb{R} \to \{-1, 0, 1\}$ is the sign function:

$$\textit{sign}(x) \triangleq \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}. \tag{A.85}$$

Proof. The proof is largely algebraic simplification based on sign-preserving operations. Taking derivatives:

$$\frac{\partial}{\partial \hat{t}^{(i)}}\, \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) = \frac{Q(y^{(i)}) - \frac{2\hat{t}^{(i)}Q(y^{(i)})^2 - Q(y^{(i)})^2}{\sqrt{D(Q(y^{(i)}), \hat{t}^{(i)})}}}{\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)}} - \frac{\left(B(Q(y^{(i)}), \hat{t}^{(i)}) - \sqrt{D(Q(y^{(i)}), \hat{t}^{(i)})}\right)(Q(y^{(i)}) + 1)}{2(\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)})^2} \tag{A.86}$$

via the quotient rule of derivatives and cancelling terms. We can apply sign-preserving operations, namely, positive scalar multiplication, canceling additive zeroes, and commuting additive terms, as

follows:

$$\frac{\partial}{\partial \hat{t}^{(i)}}\ \hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)}) \propto (\hat{t}^{(i)} + Q(y^{(i)})\hat{t}^{(i)}) \left(2Q(y^{(i)}) - \frac{4\hat{t}^{(i)}Q(y^{(i)})^2 - 2Q(y^{(i)})^2}{\sqrt{D(Q(y^{(i)}), \hat{t}^{(i)})}} \right)$$

$$- \left(B(Q(y^{(i)}), \hat{t}^{(i)}) - \sqrt{D(Q(y^{(i)}), \hat{t}^{(i)}))} \right) (Q(y^{(i)}) + 1) \tag{A.87}$$

$$\propto \hat{t}^{(i)} \left(2Q(y^{(i)}) - \frac{4\hat{t}^{(i)}Q(y^{(i)})^2 - 2Q(y^{(i)})^2}{\sqrt{D(Q(y^{(i)}), \hat{t}^{(i)})}} \right)$$

$$- B(Q(y^{(i)}), \hat{t}^{(i)}) + \sqrt{D(Q(y^{(i)}), \hat{t}^{(i)}))} \tag{A.88}$$

$$= \left(\frac{2\hat{t}^{(i)}Q(y^{(i)})^2 - 4\hat{t}^{(i)2}Q(y^{(i)})^2}{\sqrt{D(Q(y^{(i)}), \hat{t}^{(i)})}} \right) - 1 + \sqrt{D(Q(y^{(i)}), \hat{t}^{(i)}))} \tag{A.89}$$

$$\propto 2\hat{t}^{(i)}Q(y^{(i)})^2 - 4\hat{t}^{(i)2} - \sqrt{D(Q(y^{(i)}), \hat{t}^{(i)}))} + D(Q(y^{(i)}), \hat{t}^{(i)}) \tag{A.90}$$

$$= 1 - 2\hat{t}^{(i)}Q(y^{(i)})^2 - \sqrt{D(Q(y^{(i)}), \hat{t}^{(i)}))} \tag{A.91}$$

which completes the proof. □

A.3.2.7 Definition 7: causal regularization strength

We expand on our definition of causal regularization strength here. Conventionally, regularization strength is operationalized in terms of a regularization parameter $\lambda \in \mathbb{R}_+$, given a loss $\ell(\theta)$ and a regularizer $r(\theta)$ (*e.g.*, $r(\theta) = \|\theta\|_2^2$ for some objective $J(\theta)$ of the form

$$J(\theta) \triangleq \ell(\theta) + \lambda r(\theta). \tag{A.92}$$

Eq. A.92 is an instance of *regularized risk minimization* [252]. It is also identical to *linear scalarization*, a technique for characterizing tradeoffs in multi-objective optimization. The equivalence between regularized risk minimization and linear scalarization simply reflects that regularization can impose tradeoffs in optimizing $J(\theta)$ between minimizing $\ell(\theta)$ versus $r(\theta)$. Regularized risk minimization treats $r(\theta)$ as a "penalty" term, while linear scalarization treats $r(\theta)$ as merely another objective. As λ increases, the tradeoff increasingly favors $r(\theta)$, and vice versa.

Now, consider our M-step objective example-wise:

$$\mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)}) + Q(y^{(i)})\mathcal{L}(\tilde{y}^{(i)}, \hat{y}^{(i)}\hat{t}^{(i)}). \tag{A.93}$$

The M-step objective can similarly be interpreted as variation of a regularized risk minimziation

problem, where $\ell(\theta) = \mathcal{L}(Q(y^{(i)}, \hat{y}^{(i)})$, and $Q(y^{(i)})\mathcal{L}(\tilde{y}^{(i)}, \hat{y}^{(i)}, \hat{t}^{(i)}) = r(\theta), \lambda = 1$. However, $\hat{t}^{(i)}$ is a constant that can affect regularization strength, but is not a multiplier like λ. The purpose of our result is to characterize the impact of $\hat{t}^{(i)}$ on the tradeoff between the two terms of the M-step objective.

Thus, motivated by the tradeoff/multi-objective perspective of regularization, we define regularization strength in terms of a tradeoff between optimizing $\mathcal{L}(Q(y^{(i)}, \hat{y}^{(i)})$ and $Q(y^{(i)})\mathcal{L}(\tilde{y}^{(i)}, \hat{y}^{(i)}\hat{t}^{(i)})$. We observe that

$$Q(y^{(i)}) = \arg\min_{\hat{y}^{(i)}} \mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)}) \tag{A.94}$$

and define causal regularization strength as the absolute distance between $Q(y^{(i)})$, the minimizer of $\mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)})$, and the optimum of the example-wise M-step objective.

Definition 31 (Causal regularization strength)**.** *Given an example indexed by i, and a finite loss function $\mathcal{L} : [0, 1]^2 \to \mathbb{R}$ convex in $\hat{y}^{(i)}$ on [0, 1] for all i, define*

$$\hat{y}^{(i)}_{OPT}(Q(y^{(i)}), \hat{t}^{(i)}) \triangleq \arg\min_{\hat{y}^{(i)}} J^{(i)}(\hat{y}^{(i)}, \dots) \triangleq \arg\min_{\hat{y}^{(i)}} \mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)}) + Q(y^{(i)})\mathcal{L}(\tilde{y}^{(i)}, \hat{y}^{(i)}\hat{t}^{(i)}). \tag{A.95}$$

The causal regularization strength of objective $J^{(i)}$ is defined as $R(J^{(i)}) = |Q(y^{(i)}) - \hat{y}^{(i)}_{OPT}(Q(y^{(i)}), \hat{t}^{(i)})|$.

Intuitively, we define causal regularization strength in terms of the absolute distance between the optimum of each term of the M-step objective, which captures some notion of a tradeoff between the two terms. Note that this definition does not relate to convergence to $\hat{y}^{(i)}_{\text{OPT}}(Q(y^{(i)}), \hat{t}^{(i)})$; we are largely interested in how much the solution to $\min\ \mathcal{L}(Q(y^{(i)}), \hat{y}^{(i)})$ shifts after adding the causal regularization term.

A.3.3 Additional experimental setup

For both settings, we set random seeds to 42 to facilitate reproducibility.

A.3.3.1 Additional details for fully synthetic dataset

We choose s_Y as follows:

$$S_Y(x) = (s_{Y1} \circ s_{Y2})(x)$$
$$s_{Y1}(x) = x_1 - \frac{1}{4}\sin(8\pi x_0 + \psi) \quad s_{Y2}(x) = \mathbf{R}_{\pi/6}x + 0.5$$

where ψ is a simulation parameter, and $\mathbf{R}_{\pi/6}$ is a 2D rotation matrix. Intuitively, $s_Y(x)$ rotates and translates x, then applies a sinewave-based function that yields a similarly rotated, sinewave-shaped decision boundary.

We choose s_T as follows:

$$s_T(x^{(i)}, a^{(i)}) = \mathbf{1}^\top \mathbf{x}^{(i)} - \tau_{a^{(i)}}$$

where τ_a is a simulation parameter. For demonstration, we set c_y such that $P(Y = 1) = 0.25$ to allow for sufficiently-sized performance gaps across groups to emerge.[6] As a sensitivity analysis, we also replicate all experiments on fully synthetic data across $\psi \in [0, \pi/6, \pi/3, \ldots, 11\pi/6]$, representing the "phase" of the decision boundary.

Computing simulation parameters. We discuss how we find simulation parameters for each value of q_y, q_t, and k. Given q_y and $P(A = 0) = P(A = 1) = 0.5$, we have:

$$q_y = \frac{P(Y = 1 \mid A = 0)}{P(Y = 1 \mid A = 1)}$$
$$P(Y = 1) = \frac{1}{4} = \frac{P(Y = 1 \mid A = 0) + P(Y = 1 \mid A = 1)}{2}$$

which yields, by substitution,

$$\frac{q_y}{2(q_y + 1)} = P(Y = 1 \mid A = 0)$$
$$\frac{1}{2(q_y + 1)} = P(Y = 1 \mid A = 1),$$

from which we use a binary search algorithm (bisection) evaluated using simulated versions of $X \mid A = a$ with the current estimate of the mean μ_a to solve for the requisite values of μ_a. Given values of μ_a, we can then solve for τ_a using q_t and k identically:

$$q_t = \frac{P(T = 1 \mid A = 0)}{P(T = 1 \mid A = 1)}$$
$$P(T = 1) = kP(Y = 1) = \frac{k}{4} = \frac{P(T = 1 \mid A = 0) + P(T = 1 \mid A = 1)}{2}$$

[6] Empirically, at extreme values of $P(Y = 1)$, we found artificially small performance gaps. This is because model errors tend to concentrate near the true decision boundary, which lies in the tails of the covariate distributions defining $X \mid A = a$. In those tail regions, the difference between the densities $X \mid A = a$ across values of A is smaller in our two-Gaussian simulation design.

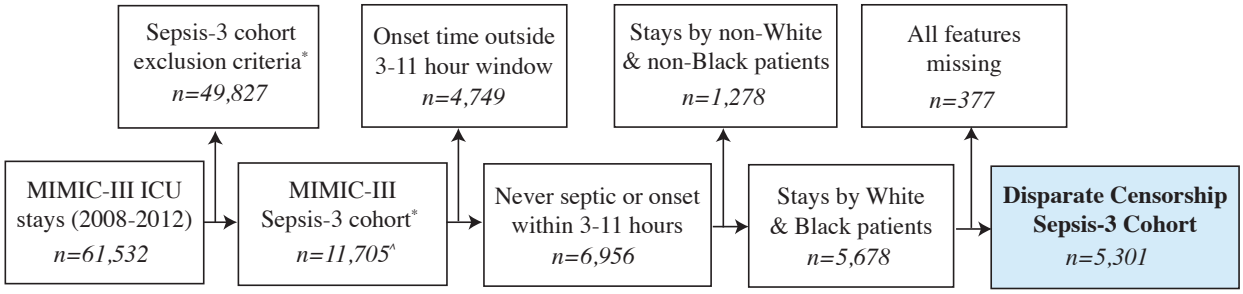


Figure A.13: Cohort diagram for our Sepsis-3 cohort ($N = 5301$). *: Further details are provided in [133]. ^: Our cohort size differs slightly from that reported in [133] due to an apparent pre-processing error in defining Sepsis-3 [260]; our reported cohort size applies the relevant correction.

which yields, again by substitution,

$$\frac{q_t k}{2(q_t + 1)} = P(T = 1 \mid A = 0)$$
$$\frac{k}{2(q_t + 1)} = P(T = 1 \mid A = 1),$$

and we can again use binary search to solve for τ_a.

A.3.3.2 Additional details for pseudo-synthetic sepsis risk-stratification task

Cohort description. Our cohort follows from the MIMIC-III Sepsis-3 cohort [133]. Their cohort exclusion criteria are publicly available.[7] We corrected an apparent Sepsis-3 definition bug that erroneously labeled individuals with suspicion of infection if they received a blood culture at any time after an antibiotic before re-running their pipeline. In contrast, the Sepsis-3 [260] definition requires the blood culture to occur *within 24 hours* of the antibiotic time for suspicion of infection.[8] In practice, this stricter condition affects $< 1\%$ of rows in their original cohort: their cohort size is $N = 11,791$, while ours is $N = 11,705$.

Feature extraction. Following the Risk of Sepsis model [62], we extract the following 13 summary statistics over the initial 3-hour observation period:

1. Maximum lactic acid measurement,

2. first shock index times age (years),

[7] https://github.com/alistairewj/sepsis3-mimic

[8] There are multiple "paths" for meeting the criteria for suspicion of infection; for a full enumeration, see Table 2 of [260].

3. last shock index times age (years),
4. maximum white blood cell count,
5. change in lactic acid (last - first),
6. maximum neutrophil count,
7. maximum blood glucose,
8. maximum blood urea nitrogen,
9. maximum respiratory rate,
10. last albumin measurement,
11. minimum systolic blood pressure,
12. maximum creatinine, and
13. maximum body temperature (Fahrenheit).

The shock index is defined as the ratio of heart rate (beats per minute) and systolic blood pressure. Missing features are replaced with -9999 following the original manuscript.

Testing decision boundary. We define s_T as

$$\beta \cdot \frac{RR_{\max} - 22}{RR_\sigma} + (1-\beta) \cdot \frac{SBP_{\min} - 100}{SBP_\sigma} - \tau_a, \tag{A.96}$$

where $RR_{\max}$ and $SBP_{\min}$ are maximum respiratory rate and minimum systolic blood pressure, respectively, and RR_σ, SBP_σ are their corresponding standard deviations on the training split ($RR_\sigma = 9.8, SBP_\sigma = 21.8$). The parameter β allows us to examine different testing decisions. Thus, we replicate all experiments over $\beta \in \{0, 0.1, \ldots, 1\}$.

A.3.4 Hyperparameters & additional model details

All hyperparameters were selected using a validation set of examples. Hyperparameters for the sepsis simulation task were chosen such that all approaches attained similar performance when using y. We reimplement all existing methods following the original papers, using the code repository as a reference if applicable. We set random seeds to 42 for all models (used for initialization), unless otherwise noted.

A.3.4.1 Default hyperparameters

Fully synthetic. All models use a two-layer neural network with layer sizes $(64, 64)$, trained for 1000 epochs via Adam [149] with learning rate 10^{-3} and no weight decay unless specified. EM approaches are trained up to 50 iterations with early stopping on validation loss (patience 3) and warm starts (initialized with solution from the previous iteration).

Sepsis classification. All predictors are three-layer neural networks with sizes $(128, 128, 16)$ trained for 10000 epochs using Adam with learning rate 10^{-5} and weight decay 10^{-3}. The DCEM propensity model (g_ζ) is trained for 20000 epochs with learning rate 10^{-5} and early stopping with patience 1000, and the DCEM model (f_θ) uses learning rate and weight decay 5×10^{-7} and 10^{-6}, respectively.

A.3.4.2 Simulation study

Peer loss & group peer loss: Both peer loss methods depend on a hyperparameter α, for which the optimal value depends on y. To show the peer loss methods in the best light, we manually calculate the optimal value for usage in training.

ITE corrected model (DragonNet): Our estimand of interest is the conditional average treatment effect (CATE) of the sensitive attribute A on testing T, which is identifiable via

$$\begin{aligned} \text{CATE}_{A\to T}(x) &\triangleq \mathbb{E}[T(1) \mid X = x] - \mathbb{E}[T(0) \mid X = x] \\ &= \mathbb{E}[T \mid X = x, A = 1] - \mathbb{E}[T \mid X = x, A = 0] \end{aligned} \tag{A.97}$$

under assumptions of consistency ($T(a) = T$) and conditional exchangeability ($T(a) \perp\!\!\!\perp A \mid X$). We then apply the CATE as a correction factor to the default model:

$$\hat{y} \triangleq \hat{\tilde{y}} - \text{CATE}_{A\to T}(x); \tag{A.98}$$

i.e., counterbalancing disparate censorship by "subracting out" the labeling bias. Note that this is an alternative to the counterbalancing approach of DCEM. We train and conduct inference with targeted regularization.

Truncated LQ: We searched across $k \in \{0.1, 0.2, \ldots, 1\}$ and $q \in \{0.1, 0.2, \ldots, 1\}$ (using the notation of the original paper), using $k = 0.1, q = 0.1$ for the final results.

SELF: We were unable to obtain convergence with Adam, so we used SGD with learning rate 0.01, momentum 0.9, noise parameter 0.05 (for input augmentation), consistency regularization parameter 1, and weight decay 1×10^{-6} as used for one of the experiments in the original paper. Weight decay was selected from $\{0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$. The ensembling/mean teacher parameters were chosen from $\{0.9, 0.99, 0.999\}$. The noise was chosen from $\{0, 0.005, 0.01, 0.05, 0.1\}$. The regularization parameter was chosen from $\{1, 5, 10, 50\}$. SELF proceeds for a maximum of 50 iterations with patience 1 with respect to validation AUC. We set ensembling momentum to 0.9 and the mean teacher moving average parameter to 0.9. To show SELF in the best light, we prevented SELF from filtering tested positive individuals.

DivideMix: We use 20 warmup epochs, with $\alpha = 4$ as the Beta parameter for the MixMatch step, $T = 0.5$, $\lambda_u = 50$, $\lambda_r = 1$, and $\tau = 0.5$, and weight decay 5×10^{-4}. We also experimented with preventing DivideMix from filtering tested positive individuals, but DivideMix was unstable in both settings. Ultimately, we did not prevent DivideMix from filtering tested positive individuals.

EM-based methods (SAREM, DCEM): We tested SAREM and DCEM with and without the usage of warm starts in the M-step.

A.3.4.3 Sepsis risk-stratification

For all baselines, the setup matches the fully synthetic setting except as specified below.

DCEM: The learning rates under consideration were $[10^{-7}, 5 \times 10^{-7}, 10^{-6}, 5 \times 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}]$. The weight decay was selected from $[0, 10^{-6}, 2 \times 10^{-6}]$.

SELF: For the sepsis classification experiments, weused SGD and set the learning rate to 10^{-8}, the highest learning rate tested that did not result in NaN loss. We tested learning rates of the form $\{10^{-d}, 5 \times 10^{-d}\}$ for $d \in \{2, 3, 4, 5, 6, 7, 8\}$.

A.3.5 Additional empirical results and discussion

A.3.5.1 Full results

Here, we report empirical results for all baselines and settings. Due to the large number of empirical settings tested (simulation: 224, sepsis classification: 45), we include a representative subset of the figures, and report the raw numbers used for these results *and* results not shown in the Appendix via CSV files in the code appendix.

For the simulated task, we show results for $k \in [1/3, 3]$, $q_t \in [0.5, 2]$, and $q_y = 0.5$. Empirically, changing q_y did not affect the general trends, but amplified/dampened the scale. Increasing q_t beyond the selected range has similar impacts. For lower values of k, all methods perform poorly. For the sepsis classification task, we show results for $k \in [1/3, 5]$ and $q_t = 1.5$.

Summary of results. We summarize when our method (DCEM) empirically performed the best, when it performed similarly to baselines, and when it underperformed baselines.

DCEM is best where...

- (Both metrics) The higher-prevalance group is undertested ($q_y < 1$ but $q_t > 1$) *and*
- (Both metrics) testing rates are sufficiently high ($k \geq 0.5$).

DCEM is similar to baselines when...

- (Both metrics) Testing rates are moderately low ($1/3 \leq k \leq 1/2$), or sufficiently high that it is easier to extrapolate from labeled data ($k \geq 3$).

DCEM underperforms baselines when...

- (ROC gap only) when testing rates are low ($k \leq 1/2$) *and*
- (ROC gap only) the testing disparity aligns with the prevalence disparity (*e.g.*, q_t and $q_y < 1$ such that learning to predict $\tilde{y}$ preserves ranking in y), *or*
- (both metrics) testing rates are extremely low ($k < 1/3$).

The strongest alternatives to DCEM in our experiments were SELF (both datasets, bias mitigation), DragonNet (sepsis only, both metrics), and the tested-only model (simulation only, discriminative performance).

Index of figures. We provide here a list of all result figures in the Appendix, indexed by problem parameters k (overall testing rate multiplier), q_t (testing disparity), and q_y (prevalence disparity; simulation only).

Fully-synthetic data

- Figure A.18: $q_y = 1/2, k = 1/3, q_t = 1/2$
- Figure A.19: $q_y = 1/2, k = 1/3, q_t = 1$
- Figure A.20: $q_y = 1/2, k = 1/3, q_t = 2$
- Figure A.21: $q_y = 1/2, k = 1/2, q_t = 1/2$
- Figure A.22: $q_y = 1/2, k = 1/2, q_t = 1$
- Figure A.23: $q_y = 1/2, k = 1/2, q_t = 2$
- Figure A.24: $q_y = 1/2, k = 1, q_t = 1/2$
- Figure A.25: $q_y = 1/2, k = 1, q_t = 1$
- Figure A.26: $q_y = 1/2, k = 1, q_t = 2$
- Figure A.27: $q_y = 1/2, k = 2, q_t = 1/2$
- Figure A.28: $q_y = 1/2, k = 2, q_t = 1$
- Figure A.29: $q_y = 1/2, k = 2, q_t = 2$
- Figure A.30: $q_y = 1/2, k = 3, q_t = 1/2$
- Figure A.31: $q_y = 1/2, k = 3, q_t = 1$
- Figure A.32: $q_y = 1/2, k = 3, q_t = 2$

Sepsis classification

- Figure A.33: $k = 1/4, q_t = 3/2$
- Figure A.34: $k = 1/3, q_t = 3/2$
- Figure A.35: $k = 1/2, q_t = 3/2$
- Figure A.36: $k = 1, q_t = 3/2$
- Figure A.37: $k = 2, q_t = 3/2$
- Figure A.38: $k = 3, q_t = 3/2$
- Figure A.39: $k = 4, q_t = 3/2$
- Figure A.40: $k = 5, q_t = 3/2$

Table A.33: Sensitivity analysis of DCEM components with respect to AUC and ROC gap (min, max across s_T in parentheses) for $q_y = 0.5, k = 1, q_t = 2$. Maximum (minimum) median AUC (ROC gap) in bold.

Method	↑ AUC	↓ ROC gap
Imputation-only	.676 [.644, .715]	.063 [.036, .086]
No causal regularization	.767 [.733, .813]	.056 [.016, .086]
DCEM (ours)	**.791 [.763, .820]**	**.031 [.019, .072]**

A.3.5.2 DCEM ablation study

To understand how DCEM design choices impact performance, we conduct an ablation study of repeated iterations and causal regularization:

- **Imputation-only**: This approach trains a model on the tested-only (labeled) examples, imputes pseudo-labels for the remaining, then trains a model on both the pseudo-labeled and labeled data. This is equivalent to a single EM iteration without causal regularization.
- **No causal regularization**: This approach runs multiple EM iterations, but without causal regularization.

The results (Table A.33) suggest that both repeated iterations and causal regularization are essential to the bias mitigation and discriminative capabilities of DCEM. The imputation-only approach fails due to low overlap between the tested vs. untested regions. Consequently, the imputed outcomes could be arbitrarily inaccurate. If we keep imputing and retraining (without causal regularization), we recover a form of pseudo-labeling [167]. The empirical improvement in performance suggests that repeated supervision from reliably labeled examples helps improve discriminative performance. However, this approach does not adjust for labeling bias (e.g., by using A), and indeed the ROC gap does not improve. Incorporating causal regularization recovers the DCEM M-step. Adding causal regularization guarantees that DCEM locally maximizes log-likelihood, and allows it to mitigate labeling bias by incorporating A into a propensity score-like term (causal regularization; see Theorem A.3.2.3).

A.3.5.3 Sensitivity analysis of causally-motivated approaches

Here, we conduct a sensitivity analysis of causally-motivated approaches under disparate censorship. The causally-motivated approaches are theoretically consistent estimators of $P(Y \mid X)$, which we can interpret at the conditional average treatment effect of testing (T) on the observed outcome ($\tilde{Y}$; see Appendix A.3.2.5). We examine the following causally-motivated estimators:

Table A.34: Sensitivity analysis of causal effect estimators for estimating $P(Y \mid X)$ compared to DCEM with respect to AUC and ROC gap (min, max across s_T in parentheses) for $q_y = 0.5, k = 1, q_t = 2$. Maximum (minimum) median AUC (ROC gap) in bold.

Method	↑ AUC	↓ ROCGap
Tested-only	.808 [.623, .876]	.052 [.020, .093]
Tested-only + group	.764 [.675, .863]	.078 [.025, .278]
IPW	**.829 [.598, .874]**	.048 [.020, .104]
DR-Learner	.643 [.558, .769]	.117 [.080, .216]
DCEM (ours)	.791 [.763, .820]	**.031 [.019, .072]**

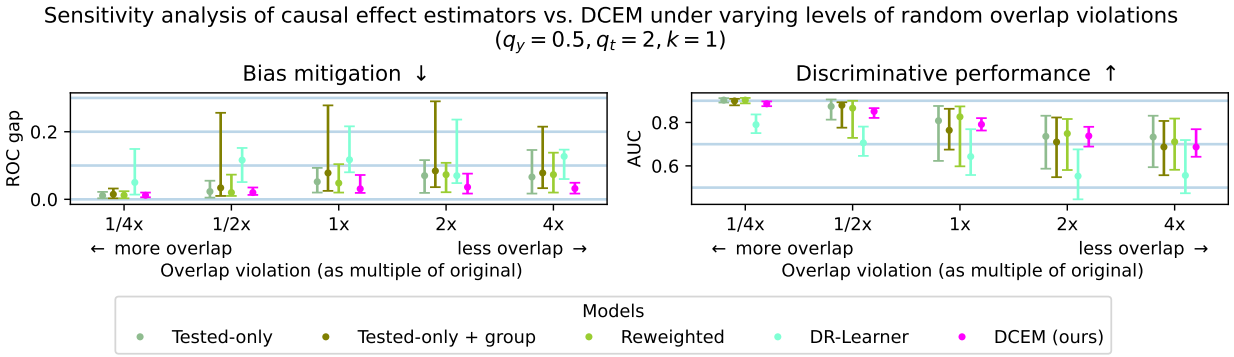


Figure A.14: Sensitivity analysis of causal effect estimators with respect to AUC and ROC gap (min, max across s_T in parentheses) for $q_y = 0.5, k = 1, q_t = 2$ across varying levels of overlap. Causally-motivated methods are shown in green, while DCEM is shown in magenta. Empirically, DCEM improves robustness to overlap violations.

- **Tested-only**: training models on tested individuals only, using X as covariates,
- **Tested-only + group**: training models on tested individuals only, using X and A as covariates,
- **Inverse propensity weighting (IPW)**: an IPW-based [242] version of the tested-only approach, and
- **Doubly-robust estimator (DR-Learner)**: a doubly-robust estimator of $P(Y \mid X)$ [142].

Models are evaluated for $q_y = 0.5, k = 1, q_t = 2$ (*i.e.*, same setting as Fig. 3.13). Under disparate censorship, low overlap is common due to the "sharpness" of the testing boundary. To validate this hypothesis, we also evaluate causal effect estimators versus DCEM at varying levels of overlap ($1/4$x, $1/2$x, 1x, 2x, and 4x of the original setting). Overlap is controlled by the coefficient inside the sigmoid for generating $t^{(i)}$ (*i.e.*, 30 in the original experiments).[9] For the DR-learner, we trimmed

[9]Recall that $t^{(i)}$ is generated as a Bernoulli random variable with parameters of the form $\sigma(ax + b)$.

propensity scores (threshold: 0.05) to obtain estimates that were in [0, 1] (the possible values of $P(Y \mid X)$).

DCEM has better bias mitigation capabilities than causal approaches, and a tighter range of discriminative performance. Table A.34 shows that, empirically, DCEM exhibits lower variance under overlap violations than causally-motivated approaches. Notably, DCEM achieves the lowest median ROC gap, and maintains competitive (but not necessarily best) median AUC. Causally-motivated methods generally have good median discriminative performance, but poor bias mitigation properties. Furthermore, the wide performance ranges of causally-motivated approaches may be unacceptable for safety-critical/high-stakes domains. We note that the DR-learner may underperform in this setting due if the propensity score trimming introduces sufficient bias: recall that, although double-robustness only requires one correctly-specified model, the asymptotic properties may still depend on the asymptotics of each model (*e.g.*, as shown in [287]).

DCEM is empirically more robust to overlap violations. Figure A.14 shows that, empirically, as overlap violations increase, DCEM degrades more slowly than causally-motivated approaches in terms of both bias mitigation and discriminative performance. Furthermore, DCEM maintains similarly tight performance ranges across levels of overlap, while the performance ranges of causal approaches widens as overlap violations increase. At low overlap, causally-motivated approaches have similarly tight performance ranges as DCEM.

A.3.5.4 Sensitivity analysis of softmax temperature scaling

We can further tune the smoothness of $\hat{t}^{(i)}$ via the softmax temperature τ of the binary classifier for $\hat{t}^{(i)}$:

$$\hat{t}^{(i)} := \frac{\exp(z_t/\tau)}{\exp(z_t/\tau) + \exp(z_{1-t}/\tau)} \tag{A.99}$$

where z_t is the logit outputted by g_ζ for each $t \in \{0, 1\}$. Lower values of τ sharpen $\hat{t}^{(i)}$ towards $\{0, 1\}$, while larger values smooth $\hat{t}^{(i)}$ toward $\frac{1}{2}$. Note that $\tau = 1$ recovers the standard softmax function. Thus, adjusting τ allows us to control the smoothness of the $\tilde{y}^{(i)} = t^{(i)}y^{(i)}$ constraint.

Table A.35 shows full results (median AUC and ROC gap, plus minima and maxima across s_Y) for DCEM across various values of temperature scaling parameter τ. Empirically, our results suggest that temperature scaling does not significantly change the AUC, and may trade off with bias mitigation since $\hat{t}^{(i)}$ may no longer be calibrated. Furthermore, even though median AUC improves in one case ($\tau = 10$), the range of AUC is much larger (0.057 vs. 0.128), and $\tau = 1$ still yields the maximum empirical worst-case AUC (0.763).

Table A.35: Sensitivity analysis of softmax temperature scaling parameter (T) with respect to DCEM AUC and ROC gap (min, max across s_T in parentheses) for $q_y = 0.5, k = 1, q_t = 2$. Maximum (minimum) median AUC (ROC gap) in bold.

τ	$\uparrow$ AUC	$\downarrow$ ROC gap
0.01	.778 [.737, .815]	.051 [.020, .104]
0.1	.791 [.762, .818]	**.025 [.014, .057]**
1 (default)	.791 [.763, .820]	.031 [.019, .072]
10	**.800 [.730, .858]**	.051 [.021, .096]
100	.762 [.667, .835]	.071 [.032, .097]

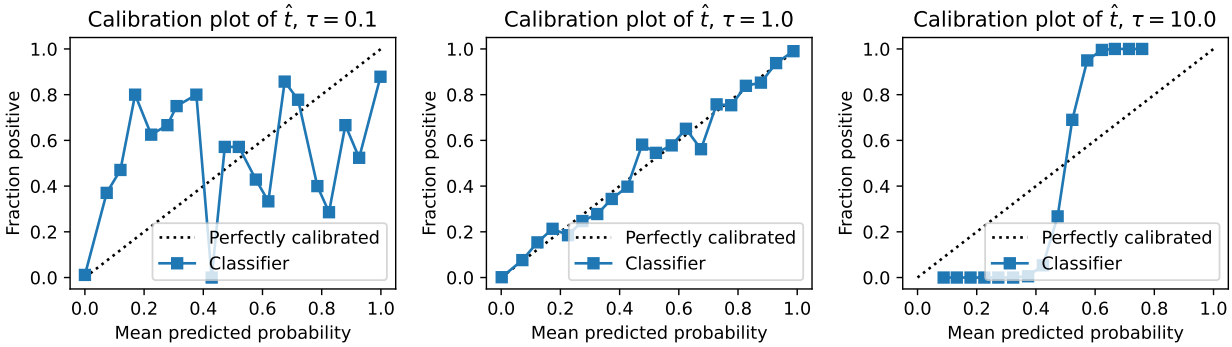


Figure A.15: Calibration plot of $\hat{t}$ for (from left to right) $\tau \in \{0.1, 1, 10\}$. While $\hat{t}$ is well-calibrated for $\tau = 1$, changing τ in either direction (< 1 vs. > 1) induces miscalibration error.

Values of τ away from 1 tend to yield larger ROC gaps. We find that $\hat{t}^{(i)}$ is well-calibrated for $\tau = 1$, but not so for values of τ (Figure A.15). Since $\hat{t}^{(i)}$ is critical to counterbalancing disparate censorship, miscalibration error in $\hat{t}^{(i)}$ could result in larger ROC gaps by reducing the effectiveness/correctness of the causal regularization term. Thus, we opt to maintain $\tau = 1$.

A.3.5.5 Sensitivity analysis of E-step initialization

We compare random initialization to using a tested-only model as initialization (the final approach). Empirically, Table A.36 shows trivial changes to performance when using a model trained on labeled data to initialize the E-step. This suggests that DCEM is able to overcome poor initialization in the settings studied; *i.e.*, the gains from tested-only initialization may be marginal, if nonzero.

A.3.5.6 Tradeoffs between bias mitigation and discriminative performance: SELF

We compare instances of DCEM to SELF, controlling for ROC gap. We find that DCEM optimizes discriminative performance more effectively than SELF. Fig. A.16 shows a histogram of AUC for

Table A.36: Sensitivity analysis of E-step initialization with respect to DCEM AUC and ROC gap (min, max across s_Y in parentheses) for $q_y = 0.5, k = 1, q_t = 2$. Maximum (minimum) median AUC (ROC gap) in bold.

Initialization scheme	AUC	ROC gap
random	.787 [.768, .822]	**.031 [.011, .060]**
tested-only	**.791 [.763, .820]**	**.031 [.019, .072]**

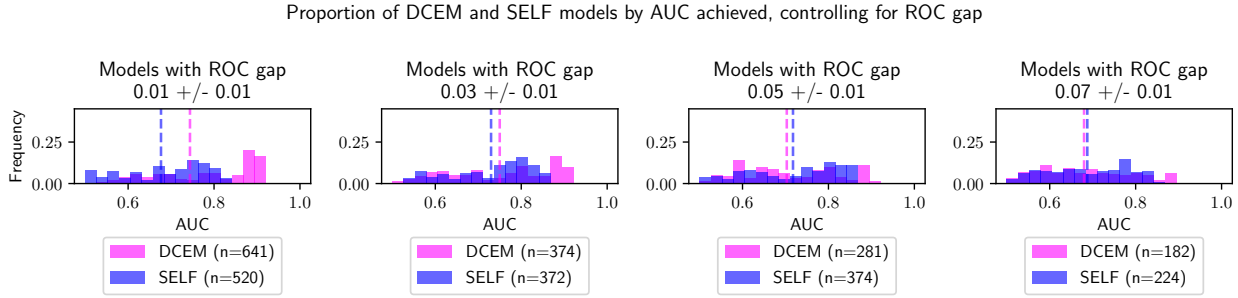


Figure A.16: Relative frequencies of AUC for DCEM vs. SELF at similar ROC gaps, pooling models across all k, q_y, q_t tested. Dashed lines = mean AUC by model. DCEM improves AUC among models with similar ROC gaps when the ROC gap is below 0.04.

SELF and DCEM models with similar ROC gaps across q_t, q_y, k and s_Y, increasing in ROC gap to the right. For models with ROC gaps < 0.04 (Fig. A.16, 1st and 2nd from left), DCEM improves AUC compared to instances of SELF with similar ROC gaps. At larger ROC gaps, DCEM and SELF obtain similar AUCs (Fig. A.16, 1st and 2nd from right). Similarly to the comparison with tested-only models, the results suggest that DCEM is not simply trading improved bias mitigation for performance, but is also able to optimize discriminative performance. Since SELF is a filtering approach that does not account for the causal structure of disparate censorship, its estimates of label bias are likely skewed. In contrast, DCEM explicitly uses the causal structure of disparate censorship to counterbalance label bias.

A.3.5.7 Sepsis classification and robustness to shifts in labeling decisions

Fig. A.17 shows the performance of DCEM vs. models with bimodal behavior across different s_T, indexed by different feature weightings in s_T. Our results suggest that the baselines require specific s_T to perform above random. The baselines catastrophically underperform (AUC below 0.5) otherwise. Trends are analogous for the ROC gap.

Specifically, baseline performance improves when one feature is more heavily weighted than the other in the labeling decision (x-axis near 0 or 1). However, when both features feature in labeling

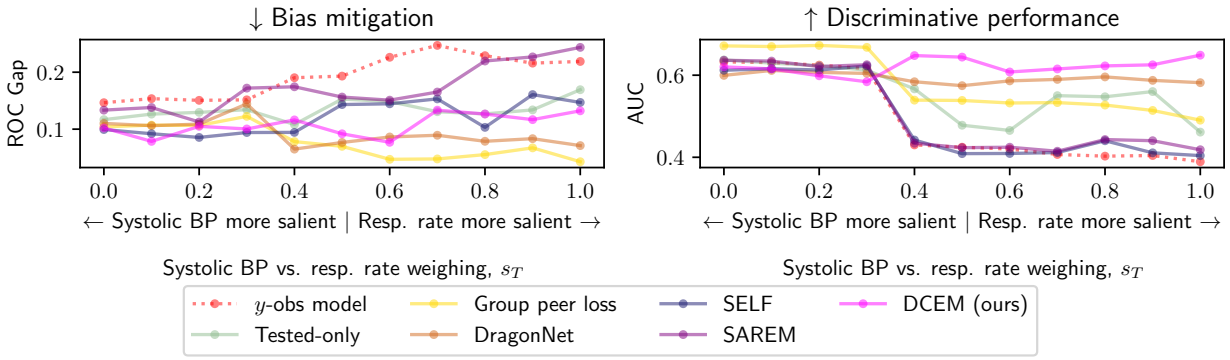


Figure A.17: ROC gaps (left) and AUC (right) of selected models on sepsis classification task as weighting of systolic blood pressure (BP) and respiratory (resp.) rate (s_T) for testing changes ("0.0"/left: consider systolic BP only; "1.0"/right: consider resp. rate only) at $q_t = 1.5, k = 4$. If a feature is "more salient," it is weighted higher than the other in the testing decision function s_T.

decisions (x-axis near 0.5), the baselines catastrophically fail, while DCEM performance stays high. As seen in Fig. 3.15, DCEM AUC and ROC gap also exhibit less variation across the different s_T.

Determining which s_T is appropriate is a clinical problem that requires domain expertise, and we make no claims as to the clinical appropriateness of s_T. Thus, ML practitioners should not assume that their data will be representative of any particular decision-making pattern. DCEM is an alternative approach that is more robust than baselines to shifts in s_T, and thus warrants consideration when narrow assumptions about labeling biases are undesirable.

A.3.6 Computing Infrastructure

Hardware. We parallelize experiments across 4 A6000 GPUs and 256 AMD CPU cores (4x AMD EPYC 7763 64-Core processors), though the memory requirements of each model are under 2GB of VRAM.

Software. All experiments are run on a distribution of Ubuntu 20.04.5 LTS (Focal Fossa) with Python 3.9.16 managed by conda 23.3.1. We use Pytorch 1.13.1 with CUDA 11.6 for all experiments [211], with scikit-learn 1.2.2 [218], scipy 1.10.1 [283], numpy 1.25.0 [107] and pandas 1.5.3 [274] for data processing/analysis. Matplotlib 3.7.1 was used to generate figures. Additionally, torch_ema 0.3 was used in our implementation of SELF. For the simulation study, we use a modified version of the official disparate censorship repository at https://github.com/MLD3/disparate_censorship [42], which is included with our code repository.

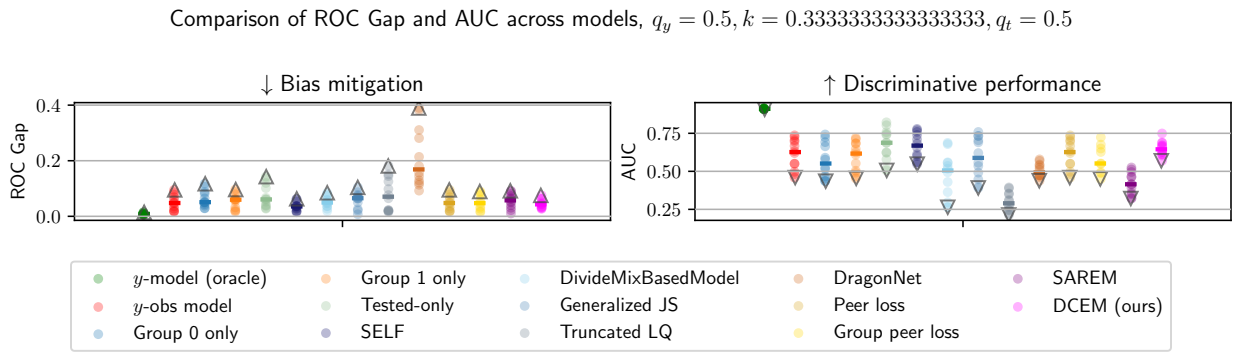


Figure A.18: ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/3, q_t = 1/2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

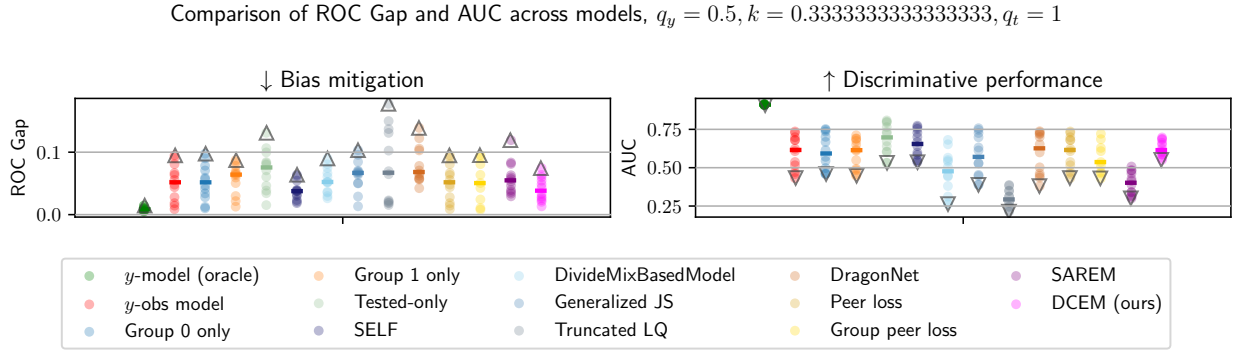


Figure A.19: ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/3, q_t = 1$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

A.3.7 Code

Code will be released at the MLD3 Github repository at https://github.com/MLD3/DCEM. We redact the data-processing code for the sepsis task only where necessary to ensure compliance with the terms of use for MIMIC-III [134].

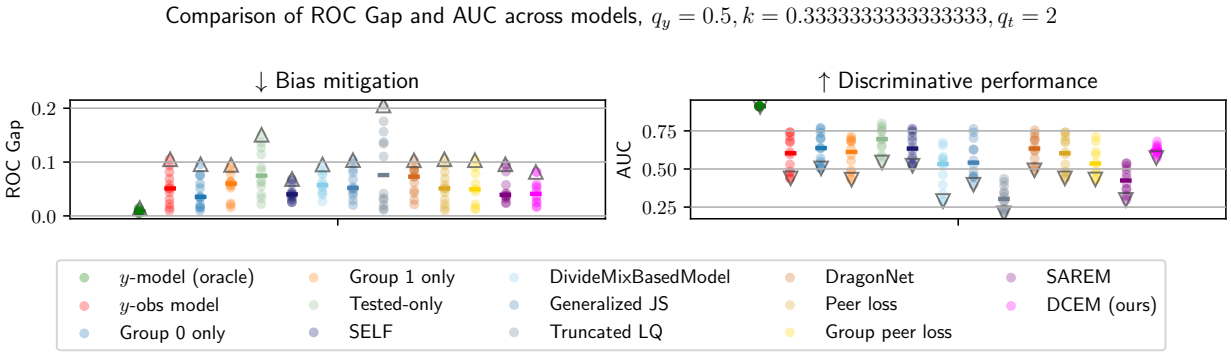


Figure A.20: ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/3, q_t = 2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

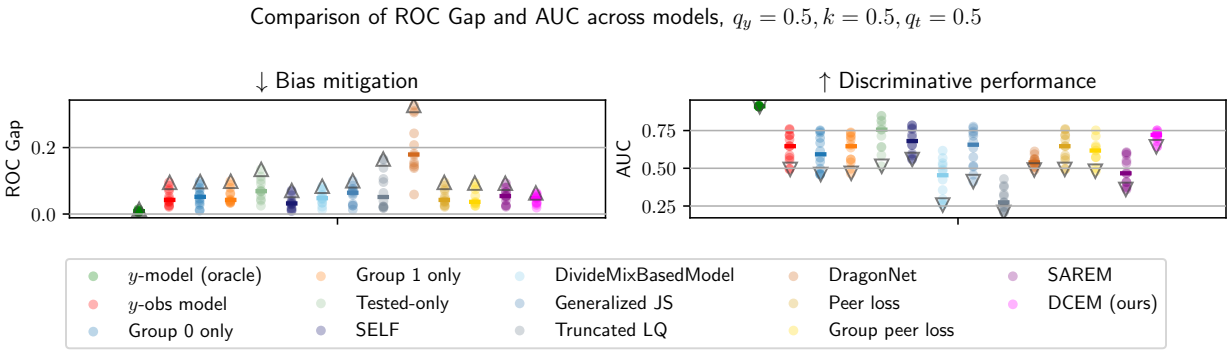


Figure A.21: ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/2, q_t = 1/2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

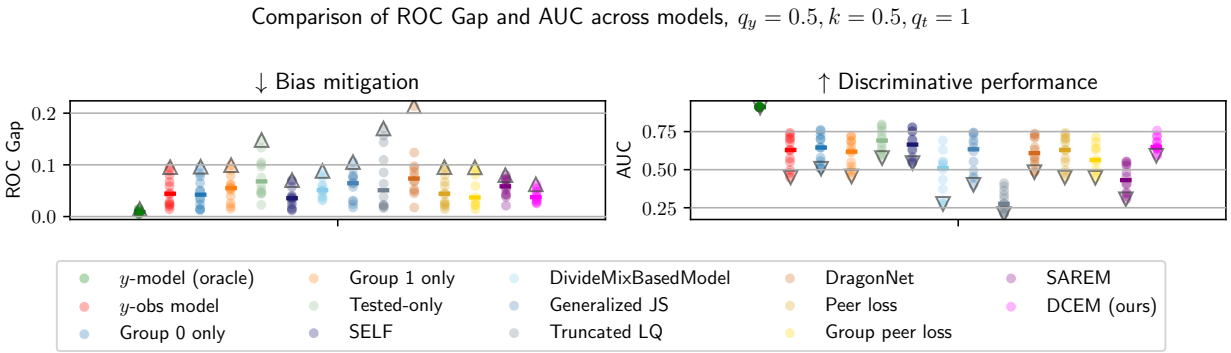


Figure A.22: ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/2, q_t = 1$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

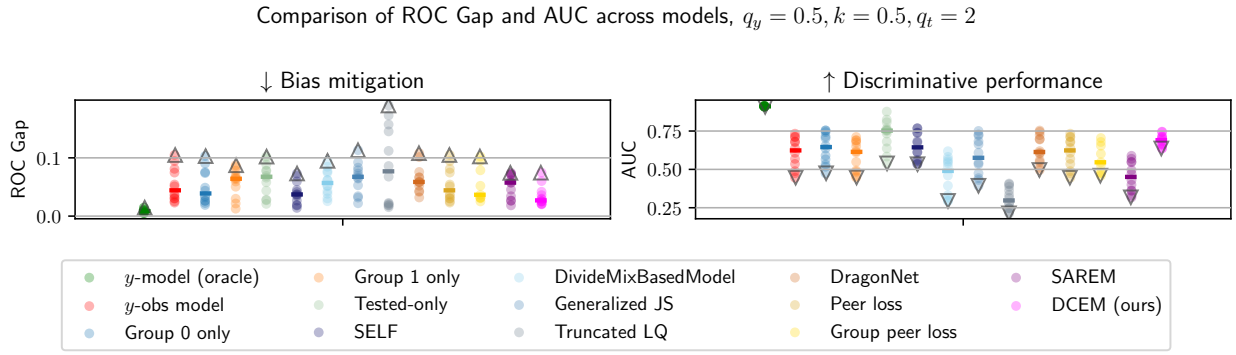


Figure A.23: ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1/2, q_t = 2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

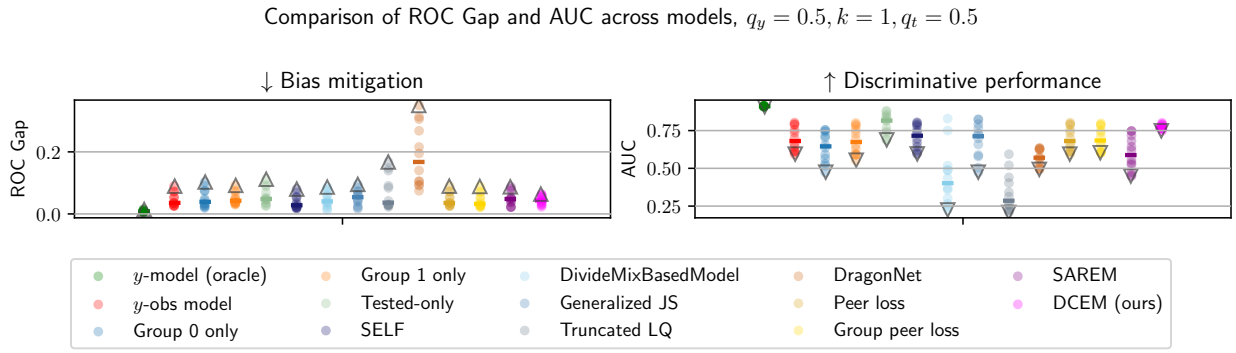


Figure A.24: ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1, q_t = 1/2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

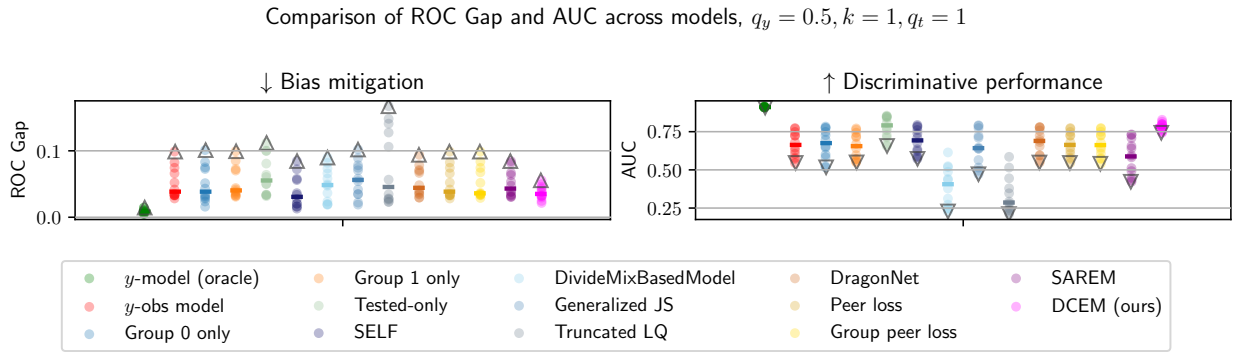


Figure A.25: ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1, q_t = 1$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

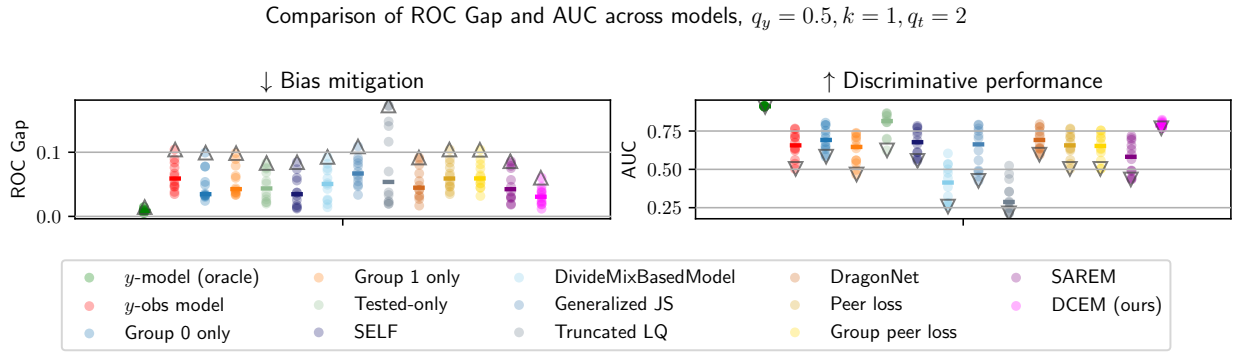


Figure A.26: ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 1, q_t = 2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

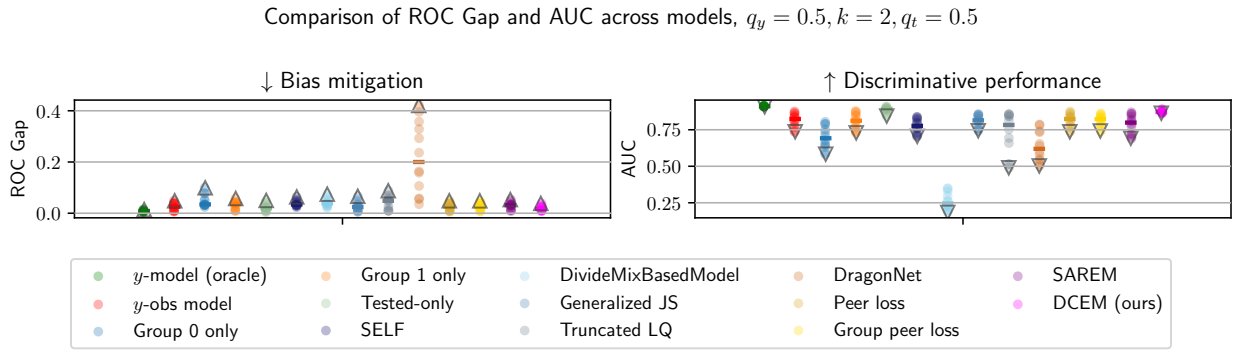


Figure A.27: ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 2, q_t = 1/2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

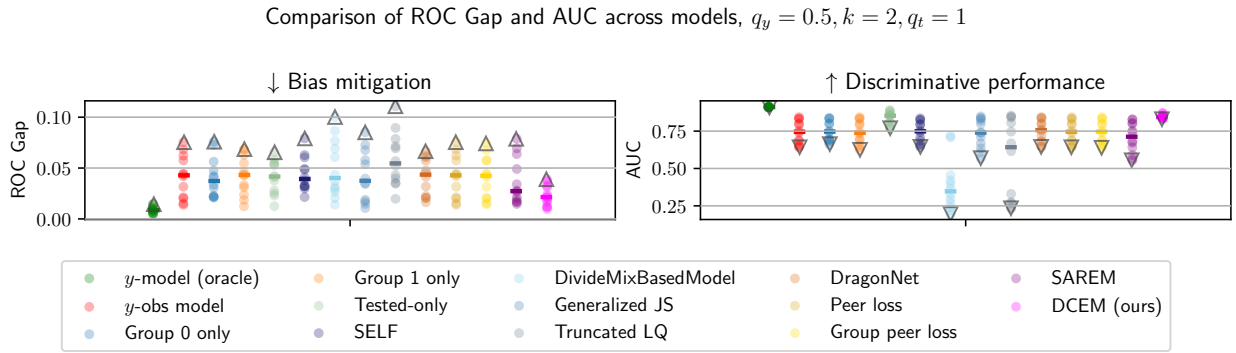


Figure A.28: ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 2, q_t = 1$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

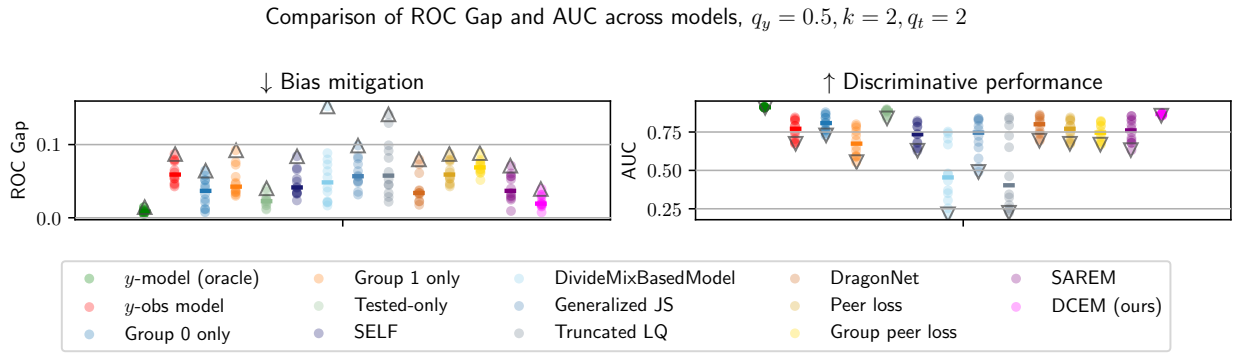


Figure A.29: ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 2, q_t = 2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

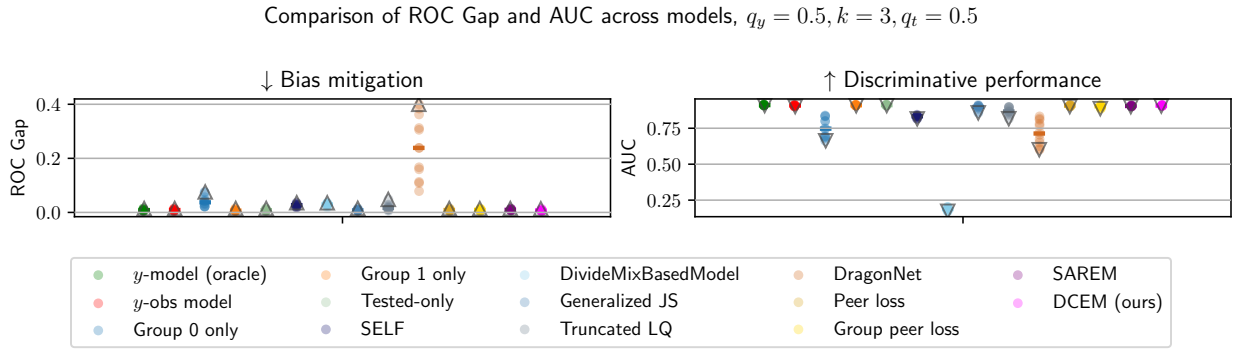


Figure A.30: ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 3, q_t = 1/2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

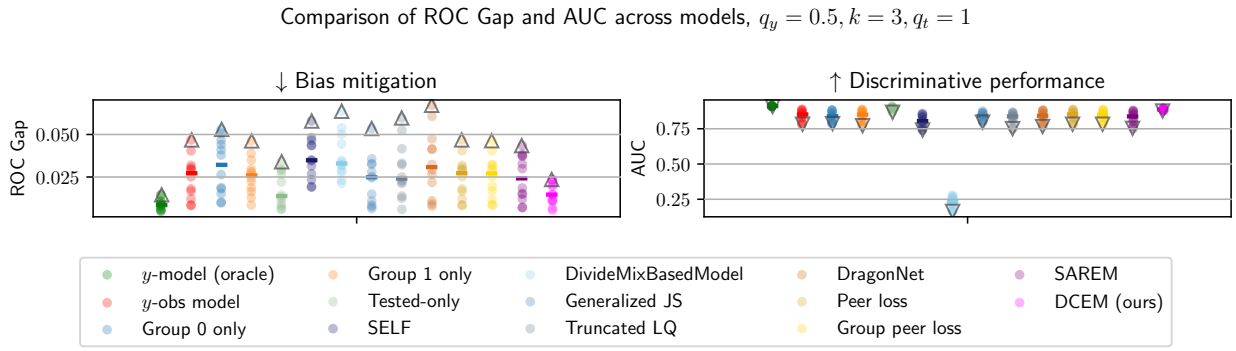


Figure A.31: ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 3, q_t = 1$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

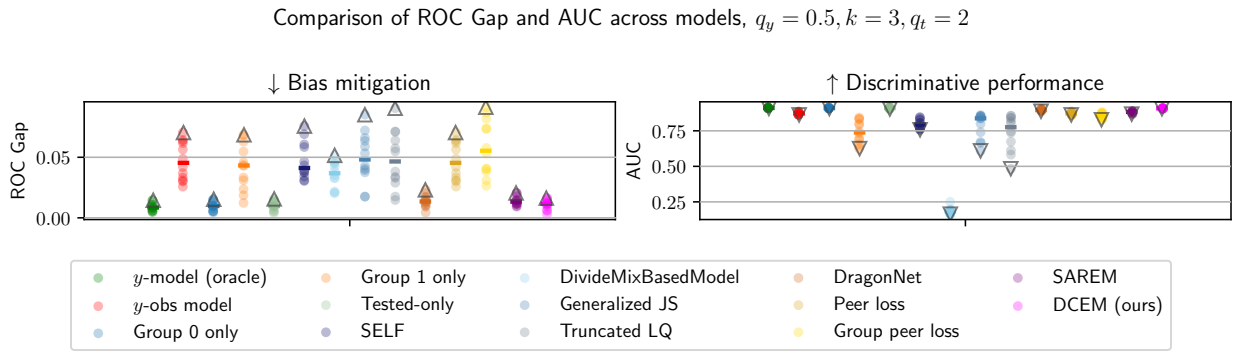


Figure A.32: ROC gap (left) and AUC (right) of baselines on simulated data at $q_y = 1/2, k = 3, q_t = 2$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

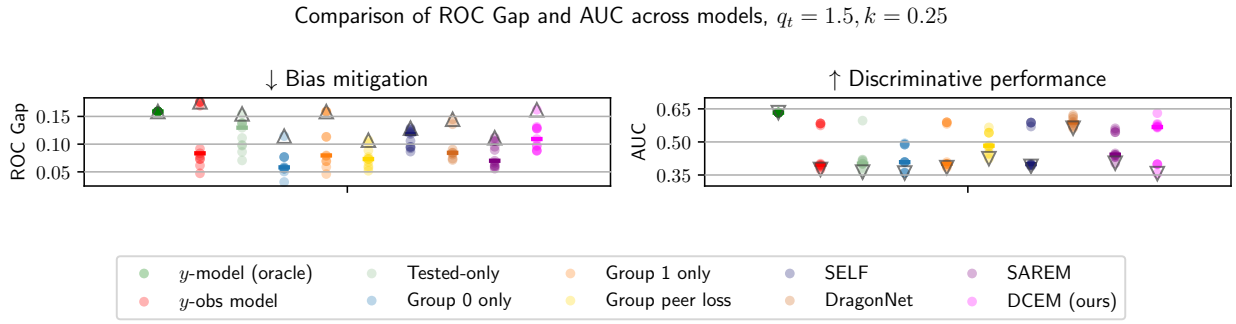


Figure A.33: ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 1/4, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

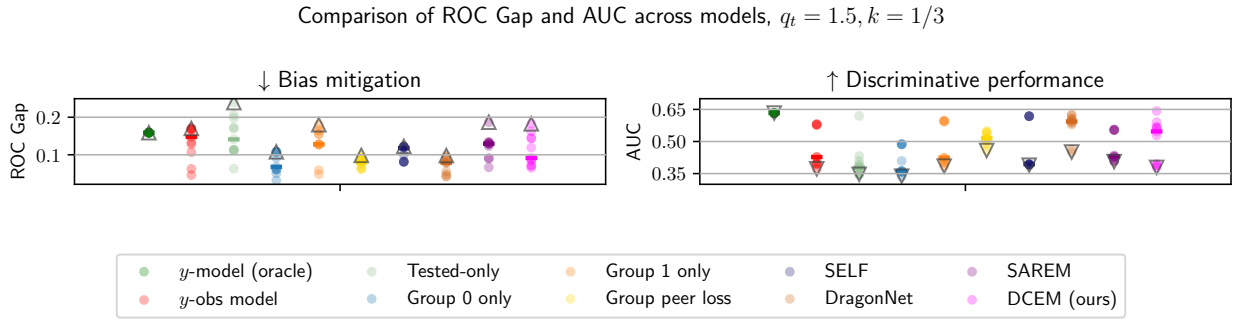


Figure A.34: ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 1/3, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

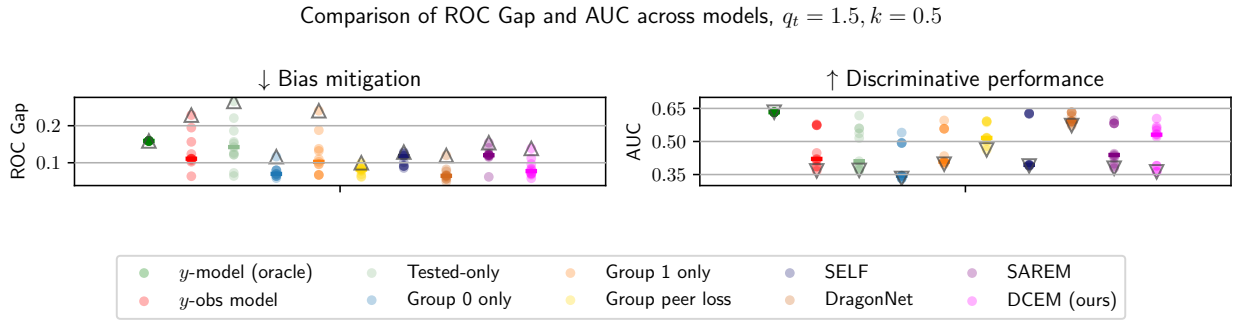


Figure A.35: ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 1/2, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

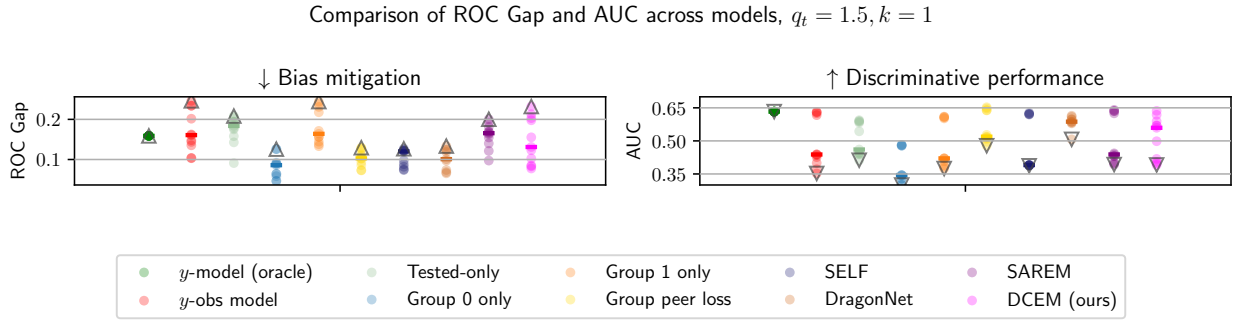


Figure A.36: ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 1, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

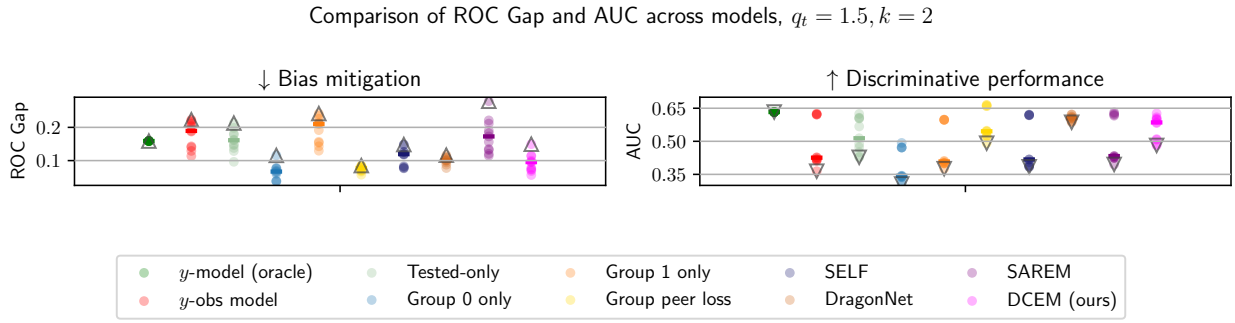


Figure A.37: ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 2, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

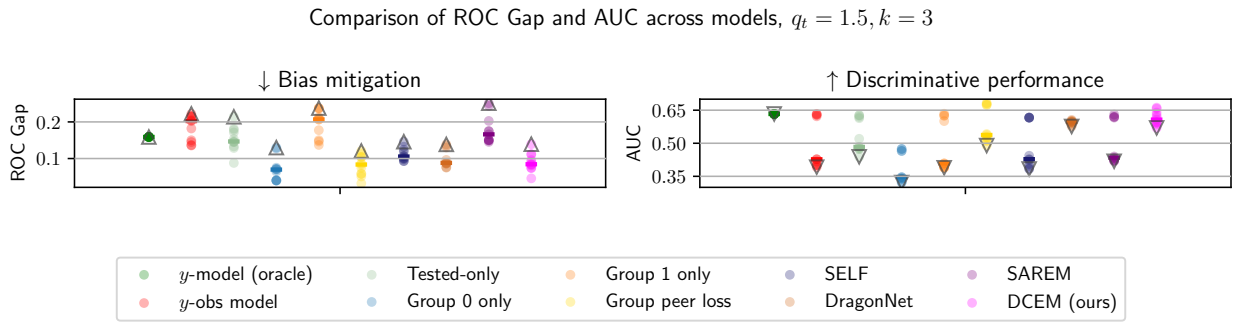


Figure A.38: ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 3, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

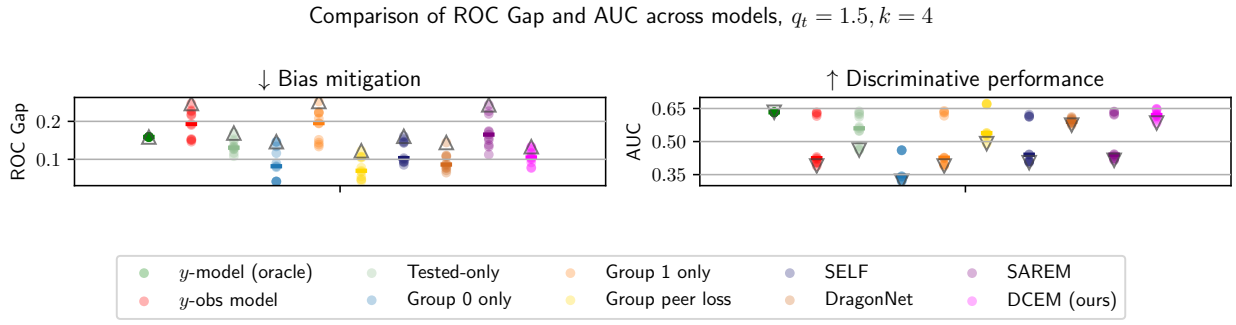


Figure A.39: ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 4, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

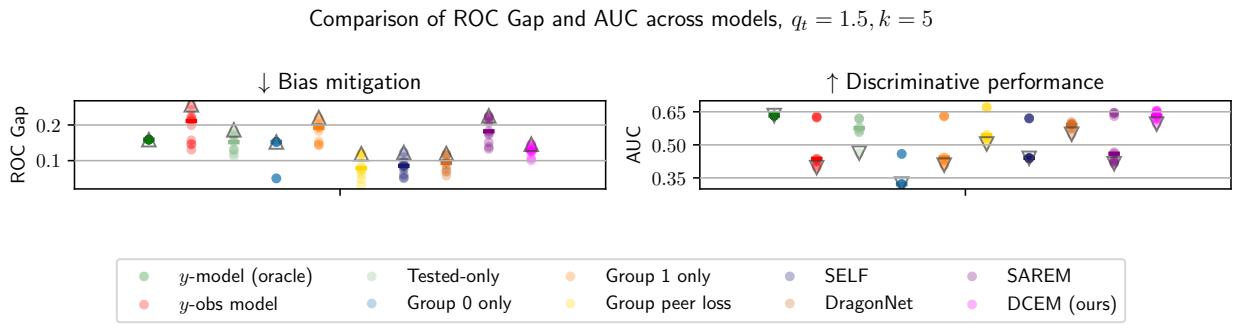


Figure A.40: ROC gap (left) and AUC (right) of baselines on sepsis classification at $k = 5, q_t = 1.5$. "-": median, "△": worst-case ROC gap, "▽": worst-case AUC.

APPENDIX B

Appendix for A Course Correction in Steerability Evaluation: Goal-conditioned Model Behavior at Inference Time

B.1 Supplementary material for A Course Correction in Steerability Evaluation: Revealing Miscalibration and Side Effects in LLMs

B.1.1 Steerability probe implementation details

B.1.1.1 Dataset preprocessing

Each dataset is pre-processed as follows. All datasets are processed via the HuggingFace `datasets` library, for which we provide dataset identifiers here. We report the number of texts from each dataset in our set of seed texts (*i.e.*, the set of texts from which the steerability probe subsamples source texts) after pre-processing.

- CNN/DailyMail (`ccdv/cnn_dailymail`, $N = 2,996$, License: MIT [249]): The CNN/DailyMail dataset is a collection of over 300,000 total online news articles from CNN from April 2007 to April 2015 and DailyMail from June 2010 to April 2015. We use the `validation` split of version 3.0.0, and extract source texts from a random subsample of 3,000 entries in the `article` column.
- BookSum (`kmfoda/booksum`, $N = 2,903$, License: BSD-3 [157]): The Booksum dataset contains public domain short stories, plays, and novels from Project Gutenberg, originally split by chapter. We use the `validation` split, and extract source texts from a random subsample of 30 entries in the `chapter` column. Since BookSum contains multiple summaries of book chapters, we de-duplicate the `chapter` column by filtering for summaries

where the `source` is equal to `"sparknotes"` prior to sampling. Each chapter is greedily chunked into paragraphs, where paragraphs are added to a "chunk" until the chunk exceeds 30 sentences in length, as measured via `nltk.sent_tokenize`.

- RedditTIFU (`ctr4si/reddit_tifu`, $N = 2,116$, License: unknown [146]): RedditTIFU is a collection of approximately 120,000 social media posts from Reddit, drawn from the "subreddit" (*i.e.*, a sub-forum) `r/tifu`. The `r/tifu` subreddit focuses on users recounting embarrassing personal experiences. We use the `train` split of the `short` version, and extract source texts from a random subsample of 3,000 entries in the `stories` documents. Due to the lack of punctuation, we detect ends of paragraphs via the delimiter `"\n\n"` and ensure such paragraphs end with periods to prevent artificial inflation of scores associated with sentence length (e.g., Flesch-Kincaid score).

- SummScreenFD (`tau/scrolls`, $N = 288$, License: unknown [251, 46]): The SummScreenFD dataset is drawn from paired TV show transcripts and human-written recaps. We select all source texts in the `validation` split in the column `output`, using version `summ_screen_fd`. We use the version of SummScreenFD included as a subset of the SCROLLS benchmark.

All seed data are filtered to be between 50 and 2048 words, inclusive as measured via NLTK's tokenization function. The upper bound of 2048 words is chosen to cap LLM generation costs. The lower bound of 50 is chosen since the measure of textual lexical diversity is only considered valid on texts of length 50 and above. After pre-processing, we have 8,303 seed texts from which steerability probes are constructed. N is reported post-filtering.

B.1.1.2 Steerability probe implementation details

Computing goal-space mappings. We map all 8,303 seed texts to goal space (see Section 4.3.1) without normalization. We measure each goal dimension as follows:

- **Reading difficulty (Flesch-Kincaid reading level)**: An approximation of reading grade level in the U.S. education system, given by

$$0.39\frac{\text{\# of words}}{\text{\# of sentences}} + 11.8 \cdot \frac{\text{\# of syllables}}{\text{\# of words}} - 15.59 \tag{B.1}$$

  which we compute via the `textstat` package.

- **Text length**: A measure of verbosity based on a function of the word count as computed via `nltk.word_tokenizer`.

- **Textual diversity (Measure of textual lexical diversity, MTLD)**: MTLD keeps a running counter of type-token ratio (TTR; # of unique tokens / # of total tokens), and defining "chunk boundaries" whenever the TTR drops beneath a pre-defined threshold (generally 0.72). The MTLD is the average length of the resultant "chunks." MTLD is calculated via the `taaled` package with pre-processing via `pylats`, a SpaCy-based pipeline for normalizing case, correcting misspellings, and part-of-speech tagging.

- **Formality (Heylighten-Dewaele F-score)**: Based on the observation that formal language tends to be less context-dependent (diectic) than informal language, and that certain parts of speech are associated with diectic vs. non-diectic language, the F-score is given by

$$\frac{(\%\text{noun} + \%\text{adj.} + \%\text{adp.} + \%\text{art.}) - (\%\text{pron.} + \%\text{verb} + \%\text{adv.} + \%\text{intj.}) + 100}{2}$$

  Part-of-speech tagging was done via `spacy` using the `en_core_web_sm` model. Since `spacy` does not have an "article" category (instead tagging determiners), we tag "a", "an", and "the" as articles manually. [1]

While we do not claim these formulae to be authoritative measures of each dimension, these dimensions were chosen as well-established, rules-based metrics for textual analysis, aiding interpretation of the results. Note that for all goal dimensions, higher numerical values indicate higher levels of the textual aspect of interest.

For validation, we prompted Llama3.1-8B to evaluate all pairs of original texts and their rewrites under the steerability probe shown in Figure 4.2 using the default prompting strategy. The LLM-as-judge prompt required the model to select whether the original text and rewrite was higher/lower on each aspect of text. We evaluated Kendall's τ between the LLM predictions and the ground-truth answers (*i.e.*, the sign of the difference between the source text and the rewrite for each goal dimension), showing at least $\approx 67.0\%$ agreement in all dimensions (Table B.1).

Goal dimension normalization. Since each dimension may be measured on a different scale, to improve comparability, we re-scale all goal dimensions to the range [0, 1], with the idea that values corresponding to zero (one) represent extremely low (high) values of each aspect.

Formally, let $\alpha_{q,i}$ be the qth quantile of goal i with respect to the seed data, and let $\tilde{z}_i$ and z_i be the raw and normalized goal-space mappings for goal i. We linearly rescale the middle 95% of each

[1] Abbreviations: adj. = adjective, adp. = adposition, art. = article, pron. = pronoun, adv. = adverb, intj. = interjection. Spacy tags adpositions, a generalization of prepositions, while the original Heylighen-Dewaele formula lists prepositions. The discrepancy likely has marginal effects: most English adpositions are prepositions with exceptions including the postpositions "ago" (e.g., three weeks ago) and "hence" (*e.g.* two days hence).

Goal dimension	Kendall's τ	Pairwise agreement
Reading difficulty	0.4644	73.2%
Textual diversity	0.3395	67.0%
Textual length	0.6814	84.1%
Formality	0.4668	73.3%

Table B.1: Kendall's τ of an LLM-based evaluation of each text dimension compared to the goal-space mapping function used. Evaluation performed on rewrites by Llama3.1-8B (direct + neg. prompt). We also provide the pairwise agreement rate ($(\tau + 1)/2$).

Goal dimension	Min.	Max.
Reading difficulty	2.8	12.9
Formality	40.4	69.1
Textual diversity	44.8	128.5
Text length	78	1,509

Table B.2: Values for each goal-dimension corresponding to $[0, 1]$ in normalized goal-space.

goal dimension to cover [0, 1] and clip values accordingly:

$$z_i = \text{clip}\left(\frac{\tilde{z}_i - \alpha_{0.025,1}}{\alpha_{0.975,i} - \alpha_{0.025,i}}, 0, 1\right). \tag{B.2}$$

Thus, goal-space is $\mathcal{Z} \equiv [0, 1]^4$. The ranges that correspond to each goal dimension in our study are reported in Table B.2.

Generating instructions. For each source text, we generate goal vectors in $\mathcal{Z}$ by adding a random offset δ_i uniformly sampled from [-0.7, -0.1] or [0.1, 0.7] to three randomly-selected goal dimensions z_i. To ensure that $z_i + \delta_i \in [0, 1]$, if applicable, we clip the minimum or maximum value of δ_i and sample uniformly from the resultant range. For example, if $z_i = 0.8$, we would sample uniformly from [-0.7, -0.1] $\cup$ [0.1, 0.2]. This reflects the assumption that requests must be *feasible*: requests to make extreme texts even more extreme are not meaningful, since a very informal message cannot be made even more informal.

Generating sampling weights. To generate sampling weights, we use probabilistic-classifier based density ratio estimation [26]. Formally, for each seed text, we draw a vector from a uniform distribution $\mathcal{U}(0, 1)^4$ for all source texts. Let $C = 1$ be the class generating the distribution of goal-space mappings z on the seed data, and let $C = 0$ be the class generating a random uniform distribution of goal-space mappings. Fitting a logistic regression to predict $P(C = 1 \mid z)$, the

sampling weight for each example with goal-space mapping z is given by $\frac{P(z|C=0)}{P(z|C=1)}$, which, via Bayes' rule, is equal to $\frac{1-P(C=1|z)}{P(C=1|z)}$ since $P(C=0) = P(C=1)$ by construction. We repurpose these sampling weights for RL.

Steerability probe settings. Steerability probes are sampled according to the sampling weights, with goal-space mappings and instruction vectors created according to the above processes. The steerability probe used for our main results features 64 source texts, with 32 goal vectors per source text ($N = 2,048$) in a 4D goal-space (Section 4.4.1 and 4.4.2). For RL fine-tuning, we sample 384 source texts and 8 goal vectors per source text ($N = 3,072$). Evaluation of models post-RL (and pre-RL models/best-of-N comparisons) is conducted on a probe with 64 source texts never seen during RL fine-tuning, with 16 goal vectors each ($N = 1,024$).

B.1.2 Text-rewriting task implementation details

B.1.2.1 Text rewriting task setup

Our main steerability probe consists of requests to ask models to rewrite texts; *i.e.*, to "move" a source text with goal-space mapping $\mathbf{z}_0$ to some ideal point $\mathbf{z}^*$. For this section, let δ_i be the requested change in an arbitrary goal dimension.

Response post-processing. All responses are post-processed via regex-based searches to filter extraneous text that precede the rewrite (e.g., `<think>` tokens from DeepSeek-family models, or phrases such as "Sure, here is your rewritten text..."). For chain-of-thought prompting only, the block under "## Rewritten text" is explicitly extracted via regex. The following paragraphs show examples of all prompting strategies evaluated.

Rewrite-filtering. A trivial way to game the vanilla steerability probe is to output text unrelated to the original. For example, in a response to produce a much harder to read text, the model could simply produce a completely unrelated text with a high Flesch-Kincaid score. Models may also refuse to write certain texts, hallucinate that the source text is not provided, or truncate their responses.

To prevent such cases from polluting our evaluation, we passed post-processed outputs into an LLM-as-judge setup, asking the model to evaluate whether the rewrite vs. original are variations of the same text as judged by the prompt in Appendix B.1.2.2. We randomize whether the rewrite or original appears first. The LLM-as-judge response is parsed via regex to extract a "Yes/No" answer and a rationale. In the case of a parse failure, the answer is recorded as "None."

All "No/None" decisions and a sample of 16 random "Yes" decisions are flagged for review by the authors. The human is provided with an interactive command-line dialog showing the original text, the rewritten text, the LLM decision ("Yes/No"), and the LLM's extracted rationale. The human is given the final say to approve or overrule the LLM's rationale. While evaluating groundedness is inherently subjective, the human review is intended to target false positive decisions by the LLM judge, rather than to judge the correctness of the response with respect to the rewriting prompt.

Re-prompting. For fair comparison across prompting strategies and models, we never re-prompt the model when it returns a valid text response. In other words, we only re-prompt on API failures (4XX/5XX HTTP responses). This approach gives all models the same number of attempts to produce a rewritten text that aligns with the generated user goal.

B.1.2.2 Prompt samples

Direct prompt. This prompt is a simple, template-based prompt where the model is asked to increase or decrease various aspects of text, with optional modifiers "slightly" for requests where $|\delta_i| < 0.2$, and "much" for requests where $|\delta_i| > 0.5$. An example prompt is shown here.

Prompt

```
Please rewrite the following, but make it longer, use less
diverse language, and more formal.  Respond with only the
rewritten text and do not explain your response.
```

Note that the "slots" for each aspect of text are randomly shuffled (*i.e.*, text length, diversity, and formality do not necessarily appear in the order given above).

Negative prompting. A negative prompt explicitly instructs the model not to change any other aspects of the text, even if it would be otherwise undesirable. All negative prompt messages are injected immediately before the source text. An example is provided here with the negative prompt underlined:

Prompt

```
Please rewrite the following, but make it longer,
use less diverse language, and more formal.
You MUST not change anything else about the other parts of the
text, even if it makes the rewritten text sound unnatural or
otherwise awkward. Respond with only the rewritten text and do
not explain your response.
```

Negative-prompt variations are only well-defined for prompt strategies that explicitly name aspects of text to change (*i.e.*, direct, direct + instructions, and chain of thought), since a directive to

leave "other" aspects unmodified is only meaningful if the prompt mentions specific attributes of text to change. We omit examples of the other prompting strategies with negative prompts since their construction is identical.

Chain-of-thought. Here, the model is asked to explain its edits concretely prior to outputting the rewritten text.

Prompt

```
Please rewrite the following, but make it more formal, use less
diverse language, and longer.  Respond with only the rewritten
text and do not explain your response.  Before outputting the
rewritten text, propose and discuss a few concrete edits you
might apply to this specific text using the following format
and replacing the placeholders in []:


## Edits


[your proposed edits]


## Rewritten text


[your rewritten text]
```

Instruction-only. In lieu of naming specific goal dimensions, model is provided with specific instructions anchored to the source text to satisfy the users' goal. The instructions are extracted via an LLM from chain-of-thought responses.[2]

Prompt

```
Please rewrite the following.  Respond with only the rewritten
text and do not explain your response.  Some ways that you can
do so might include:
```

[2]Responses are extracted from chain-of-thought rewrites by Llama-3.1 (8B) without negative prompting. Empirically, proposed edits did not qualitatively differ whether or not a negative prompt was used. Instructions had varying specificity, with some referencing parts of the text, and others containing high-level strategies.

```
- Combine shorter sentences into longer ones using more complex
sentence structures.
- Replace vague or imprecise words with more specific
alternatives.
- Reduce the use of contractions.
```

Direct + instruction. This prompt combines the direct prompt and instruction-only prompt.

Prompt

```
Please rewrite the following, but make it use less diverse
language, longer, and more formal.  Respond with only the
rewritten text and do not explain your response.  Some ways
that you can do so might include:
- Combine shorter sentences into longer ones using more complex
sentence structures.
- Replace vague or imprecise words with more specific
alternatives.
- Reduce the use of contractions.
```

Underspecified. The underspecified prompt contains a vague normative phrase as an instruction. One example follows:

Prompt

```
Please rewrite the following, but make it higher-quality.
Respond with only the rewritten text and do not explain your
response.
```

Note that no direction is provided beyond "higher-quality." In addition to "make it higher-quality," we draw variations of such vague normative phrases randomly from a preset list.

B.1.2.3 LLM-as-judge details

We use Llama-3.1 (8B) as a model to extract relevant information from various texts. We provide prompt examples used for such sub-tasks.

Evaluating goal dimension validity. Here, we show our prompt for evaluating the agreement between the LLM and each goal-space mapping function:

Prompt

```
You are an impartial judge comparing two versions of a
particular text.


**Version A:** $version_a
**Version B:** $version_b


Please evaluate the two versions by answering the following
questions.


**Questions:**
* Which version is harder to read?
* Which version uses more diverse language?
* Which version is more verbose?
* Which version is more formal?


For all questions, you may answer "A", "B", or "Tie," and
provide a brief rationale (1-2 sentences).  Return only a
valid JSON object in the following format, with no additional
commentary.

```
{
"higher_reading_difficulty": {"answer": [A|B|Tie], "rationale":
[your reasoning]},
"higher_textual_diversity": {"answer": [A|B|Tie], "rationale":
[your reasoning]},
"higher_text_length": {"answer": [A|B|Tie], "rationale":
[your reasoning]},
"higher_formality": {"answer": [A|B|Tie], "rationale": [your
reasoning]},
}
```
```

```
```

This prompt is tested on rewrites produced by Llama3.1-8B on the direct + negative prompt. Note that we randomize the order in which the rewritten vs. original text appears.

Evaluating groundedness. Groundedness proceeds in two stages. The first stage is an LLM-as-judge evaluation, using the template below:

Prompt

```
You are an impartial judge comparing two texts.  Your task
is to determine whether these two texts could possibly be
describing the same event or story, or otherwise be variations
of the same text.  The two texts may vary drastically in tone,
formality, verbosity, length, or other aspects of style and
wording, and can be drawn from a variety of sources, including
but not limited to news articles, creative writing, social
media, and others.  However, the texts should very broadly
discuss the same topics and/or events.


**Version A:**
$version_a


**Version B:**
$version_b


Answer yes or no and provide a brief rationale (1-2 sentences).
Return only a valid JSON object in the following format, with
no additional commentary.

```
{
"answer": [Yes|No],
"rationale": [your reasoning],
}
```
```

The ordering of "Version A" and "Version B" is randomized. All "no" decisions and a random subsample of 16 "yes" decisions are manually-reviewed by a human familiar with the experimental setup of the text-rewriting task. The human is given a yes/no option to reject (*i.e.*, flip) the judge response, or approve the judge response. We report counts of valid responses for all steerability probes in Tables B.6 through B.7. Note that we randomize the order in which the rewrite vs. original text appears.

Instruction extraction from chain-of-thought Due to inconsistencies in the format of edits outputted by the chain-of-thought prompt, we leverage an LLM as an extraction subroutine.

Prompt

```
A writer was asked to rewrite texts in a variety of ways.
Before answering, they were asked to list some changes that
they intended to make to the text.  Given their responses,
could you help me extract the proposed edits as a numbered
list?  Respond with only the numbered list and do not explain
your response.


**Proposed edits:**
$edits
```

B.1.3 LLM inference & fine-tuning details

B.1.3.1 Inference details

Models are either hosted via vLLM locally using the OpenAI API-compatible server, or accessed via the OpenAI API. Table B.3 lists the exact model versions used for our experiments. The "model endpoint name" column is equivalent to the model key for the OpenAI API, or the HuggingFace model name for vLLM-hosted models. For HuggingFace models, we provide the first seven characters of commit hash. Local models were hosted on 1-4 A6000 GPUs via tensor parallelism.

B.1.3.2 RL objective design details

Here, we derive for various components of our RL approach. We do not claim these as new results, but as an aid to understanding our methods.

Main objective. Assume all goal dimensions are independent, and that all $\mathbf{z}^*$ are reachable for all $\mathbf{z}_0$. The first assumption implies that miscalibration and orthogonality never trade off, so

Model name	Model endpoint name	Revision
GPT-3.5 (turbo)	`gpt-3.5-turbo-0125`	N/A
GPT-4 (turbo)	`gpt-4-turbo-2024-04-09`	N/A
GPT-4o	`gpt-4o-2024-08-06`	N/A
GPT-4.1	`gpt-4.1-2025-04-14`	N/A
o1-mini	`o1-mini-2024-09-12`	N/A
o3-mini	`o3-mini-2025-01-31`	N/A
Llama3-8B	`meta-llama/Meta-Llama-3-8B-Instruct`	5f0b02c
Llama3-70B	`meta-llama/Meta-Llama-3-70B-Instruct`	28bd9fa
Llama3.1-8B	`meta-llama/Llama-3.1-8B-Instruct`	0e9e39f
Llama3.1-70B	`meta-llama/Llama-3.1-70B-Instruct`	1605565
Llama3.3-70B	`meta-llama/Llama-3.3-70B-Instruct`	6f6073b
Deepseek-8B	`deepseek-ai/DeepSeek-R1-Distill-Llama-8B`	ebf7e8d
Deepseek-70B	`deepseek-ai/DeepSeek-R1-Distill-Llama-8B`	008f3f3
Qwen-4B	`Qwen3/Qwen-4B`	82d62bb
Qwen-32B	`Qwen3/Qwen-32B`	30b8421
Qwen-30B-A3B	`Qwen3/Qwen-30B-A3B`	4c44647

Table B.3: Model endpoint and version information for all models evaluated. For HuggingFace models, the first seven digits of the model commit hash on the HuggingFace model hub is provided.

optimizing ℓ = steering error is well-principled. The second assumption ensures that $P(\mathbf{z}_0, \mathbf{z}^*) > 0$, so steerability is achievable. Then, recall that we optimize a sample-weighted objective:

$$\max . \mathop{\mathbb{E}}_{(\mathbf{z}_0,\mathbf{z}^*)\sim\mathcal{D}} \mathop{\mathbb{E}}_{\hat{\mathbf{z}}\sim f(\cdot|\mathbf{z}_0,\mathbf{z}^*)} [\hat{w}(\mathbf{z}_0, \mathbf{z}^*) \cdot \|\mathbf{z}^* - \hat{\mathbf{z}}\|_2^2] \tag{B.3}$$

To estimate $\hat{w}(\cdot)$, we make a simplifying assumption that $\mathbf{z}^* \perp\!\!\!\perp \mathbf{z}_0$. Since we can sample user goal vectors $\mathbf{z}^*$ on-demand, we simply do so uniformly. Then we need only ensure a uniform sample over source texts with respect to goal-space. Thus, we write

$$w(\mathbf{z}_0, \mathbf{z}^*) = \frac{f(\mathbf{z}_0, \mathbf{z}^* \mid \mathcal{U})}{f(\mathbf{z}_0, \mathbf{z}^* \mid \mathcal{D})} = \frac{f(\mathbf{z}_0 \mid \mathcal{U}) f(\mathbf{z}^* \mid \mathcal{U})}{f(\mathbf{z}_0 \mid \mathcal{D}) f(\mathbf{z}^* \mid \mathcal{U})} = \frac{f(\mathbf{z}_0 \mid \mathcal{U})}{f(\mathbf{z}_0 \mid \mathcal{D})} \tag{B.4}$$

where $\mathcal{U}$ is a uniform distribution and f is the respective probability density function. The first equality follows by definition. The second equality follows from our assumption that $\mathbf{z}_0$ and $\mathbf{z}^*$ are independent, and we have substituted $f(\mathbf{z}^* \mid \mathcal{D})$ with $f(\mathbf{z}^* \mid \mathcal{U})$ in the denominator since, given a known goal-space $\mathcal{Z}$, we can generate $\mathbf{z}^*$ arbitrarily. The final equality is a simple cancellation, from Eq. B.4 can be written as $w(\mathbf{z}_0)$, which we estimate via classifier-based density ratio estimation [26].

Margin-aware leave-one-out policy optimization (MA-LOOP). Our RL algorithm is a variant of leave-one-out proximal policy optimization (LOOP) [45]. For reward r, LOOP optimizes:

$$\mathcal{J}_{\text{LOOP}}(\pi; \mathcal{D}, r) = \mathbb{E}_{\mathcal{D}} \left[\frac{1}{\sum_{i=1}^{|\mathcal{G}|} |y_i|} \sum_{i=1}^{|\mathcal{G}|} \left(\sum_{t=1}^{|y_i|} \frac{\pi(y_t \mid x_i, y_{<t})}{\pi_{\text{ref}}(y_t \mid x_i, y_{<t})} \hat{A}_i - \beta D_{KL}(\pi || \pi_{\text{ref}}) \right) \right] \tag{B.5}$$

$$\hat{A}_i \triangleq \frac{|\mathcal{G}|}{|\mathcal{G}-1|} \left(r(\mathbf{z}_i^*, \hat{\mathbf{z}}_i) - \frac{1}{|\mathcal{G}|} \sum_{j=1}^{|\mathcal{G}|} r(\mathbf{z}_j^*, \hat{\mathbf{z}}_j) \right). \tag{B.6}$$

We drop the PPO clipping operation, assuming one PPO epoch. Thus, Eq. B.5 is equivalent to trust region policy optimization [247] with a leave-one-out advantage estimate (Eq. B.6) [152]. We then normalize loss per-token to mitigate response length bias [301] and rejection-sample $\mathcal{G}'$ by retaining the top- and bottom-$K/2$ responses with respect to r. To enrich reward signal, we augment Eq. B.5 with an identity preference optimization-based regularizer [10] to scale the difference between top- and bottom-$K/2$ responses with the reward difference (see Appendix B.1.3.2), yielding:

$$\mathcal{J}_{\text{MA-LOOP}}(\pi; \mathcal{D}, \mathcal{Z}, r) = \mathbb{E}_{\mathcal{D}} \left[\tilde{w}(\mathbf{z}_0) \left(\mathcal{J}_{\text{LOOP}}(\pi; \cdot) + \lambda_\tau \sum_{i=1}^{|\mathcal{G}'|} \sum_{j=1}^{K} \sum_{k=1}^{K} \Delta_{i,j,k}^2 \right) \right] \tag{B.7}$$

$$\Delta_{i,j,k} = \log \frac{\pi(y_j \mid x_i)/\pi_{\text{ref}}(y_j \mid x_i)}{\pi(y_k \mid x_i)/\pi_{\text{ref}}(y_k \mid x_i)} - \frac{r(\mathbf{z}_j^*, \hat{\mathbf{z}}_j) - r(\mathbf{z}_k^*, \hat{\mathbf{z}}_k)}{\tau}, \tag{B.8}$$

where τ and λ_τ are hyperparameters. Note the sample weight $\tilde{w}(\mathbf{z}_0)$ in Eq. B.7. The first term of Eq. B.8 is the difference in the "implicit reward" between top-$K/2$ vs. bottom-$K/2$ responses under the preference model induced by the LLM [228], while the second term is the reward difference scaled by τ. Intuitively, the difference in the "true" reward should be proportional to the difference in the LLM-induced "implicit reward." We proceed with optimizing the policy π.

Identity preference optimization-based regularization. Here, we derive the pairwise-preference based regularizer in Eq. B.8. To further enrich the signal from our completions, we aim to apply pairwise-preference loss such as the direct preference optimization (DPO) objective [228]. This is given by

$$\mathcal{L}_{\text{DPO}}(\pi; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y,y') \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi(y \mid x)}{\pi_{\text{ref}}(y \mid x)} - \beta \log \frac{\pi(y' \mid x)}{\pi_{\text{ref}}(y' \mid x)} \right) \right] \tag{B.9}$$

for some KL-regularization parameter β and preferred (non-preferred) responses y (y'). and we can use a threshold on advantages $\hat{A}_i$ to create pseudo-preferred and non-preferred classes. The implicit reward $r(x, y)$ optimized by Eq. B.9 is proportional to $\pi(y \mid x)/\pi_{\text{ref}}(y \mid x)$. Intuitively, optimization

of Eq. B.9 encourages the model to "push" the log-likelihood of preferred responses away from that of non-preferred responses.

However, the reward of a preferred (or non-preferred) response is not necessarily constant. In our setting, we have access to an oracle reward function via goal-space mappings (*i.e.*, we can calculate $\|\mathbf{z}^* - \hat{\mathbf{z}}\|_2$). Thus, the ideal "margin" between the log-likelihood of preferred vs. non-preferred responses ought to be proportional to the ground-truth reward gap. To that end, Azar et al. introduce Ψ-preference optimization, which aims to optimize

$$\mathcal{L}(\text{IPO}) = -\mathbb{E}_{x\sim P(\cdot),\, y\sim\pi(\cdot|x),\, y'\sim\mu(\cdot|x)} \left[\Psi\left(p^*(y \succ y' \mid x)\right)\right] + \tau D_{\text{KL}}(\pi \,\|\, \pi_{\text{ref}}). \tag{B.10}$$

for some non-decreasing $\Psi : [0, 1] \to \mathbb{R}$. Let $p^*(y \succ \mu) = \mathbb{E}_{y'\sim\mu(\cdot|x)}[p^*(y \succ y' \mid x)]$ for an arbitrary prior μ, and $p^*(y \succ y' \mid x)$ is the probability that y is preferred to y'. We introduce the notation $p^*_{\mathbf{z}^*}(y \succ \mu)$ to denote $\mathop{\mathbb{E}}_{x\sim P(\cdot|\mathbf{z}^*)}[p^*(y \succ \mu)]$.[3] Assume the following two statements:

$$p^*_{\mathbf{z}^*}(y \succ \mu) > p^*_{\mathbf{z}^*}(y' \succ \mu) \iff \|\mathbf{z}^* - \hat{\mathbf{z}}\|_2 < \|\mathbf{z}^* - \hat{\mathbf{z}}'\|_2, \tag{B.11}$$

$$p^*_{\mathbf{z}^*}(y \succ \mu) = p^*_{\mathbf{z}^*}(y' \succ \mu) \iff \|\mathbf{z}^* - \hat{\mathbf{z}}\|_2 = \|\mathbf{z}^* - \hat{\mathbf{z}}'\|_2; \tag{B.12}$$

i.e., in expectation over all prompts expressing intent $\mathbf{z}^*$, the negative steering error and the probability that the response is preferred induce *identical* preferences. Note that, since $\mathbf{z}^*$ is as a function of x (*i.e.*, intents are expressed through prompts), and that $\hat{\mathbf{z}}$ is a function of y ($\hat{\mathbf{z}} \triangleq \mathbf{g}(y)$), this assumption is well-posed; *i.e.*, there exists a non-decreasing $\Psi : [0, 1] \to \mathbb{R}$ such that

$$\Psi(p^*_{\mathbf{z}^*}(y \succ \mu)) - \Psi(p^*_{\mathbf{z}^*}(y' \succ \mu)) = \|\mathbf{z}^* - \hat{\mathbf{z}}'\|_2 - \|\mathbf{z}^* - \hat{\mathbf{z}}\|_2. \tag{B.13}$$

Substitution into Eq. 12 of Azar et al. yields an objective with an identical form to our regularizer.

B.1.3.3 RL implementation details

Text pre-processing. We apply the default conversational template to all texts; *i.e.*, all prompts are represented as dictionaries:

```
{"role": "user", "content": [prompt]}
```

Exploration policy. To generate rollouts prior to policy updates, we sample 64 ($|\mathcal{G}|$) completions are sampled on-policy with temperature 1.0, min-p sampling ($p = 0.2$), and a frequency penalty of 0.1, before applying rejection sampling.

[3] Note that $P(\cdot \mid \mathbf{z}^*)$ is equivalent to the "context distribution" $x \sim \rho$ of [10].

Optimization hyperparameters. We use the following hyperparameters to train our model:

- Learning rate: 2.5×10^{-7}, with linear warmup for the first 20% of training
- Optimizer: AdamW [180]
- Batch size: 4, via gradient accumulation (total: $4 \cdot K = 64$ completions per gradient update)
- Gradient clipping norm: 1.0
- LoRA [75], rank 256, α: 512, with rank-stabilization [136] and no dropout
- Rollouts per prompt ($|\mathcal{G}|$): 64
- Rejection sample size (K): 16
- KL divergence regularization parameter (β): 0.01
- IPO-regularization strength (λ_τ): 1
- IPO-regularization scale (τ): 1
- Model context length: 4,096

Memory-efficiency optimizations. To maximize memory savings, we leveraged CUDA kernels for cut-cross entropy loss for calculating sequence log-probabilities [296] and DeepSpeed ZeRO Stage 2 [231] with optimizer offloading to CPU, as well as the default `trl` gradient checkpointing. All training was done using bfloat16.

Speed-efficiency optimizations. To improve generation speed, we employed tensor parallelism (size: 2) via vLLM during training time. Goal-space mappings were implemented as a local server with 16 workers for asynchronous, parallel computation of goal-space mappings.

Software acknowledgement. Our method is built with a customized version of the GRPO impelmentation in `trl` 0.16.0.

B.1.4 Additional results

B.1.4.1 Steerability probe results

Steering error, all models. We show results for all models evaluated on our steerability probe. Figure B.1 shows that, within model families, some improvements in steering error is visible, and miscalibration improves as well. However, for all models, side effects are severe: orthogonality remains skewed toward one.

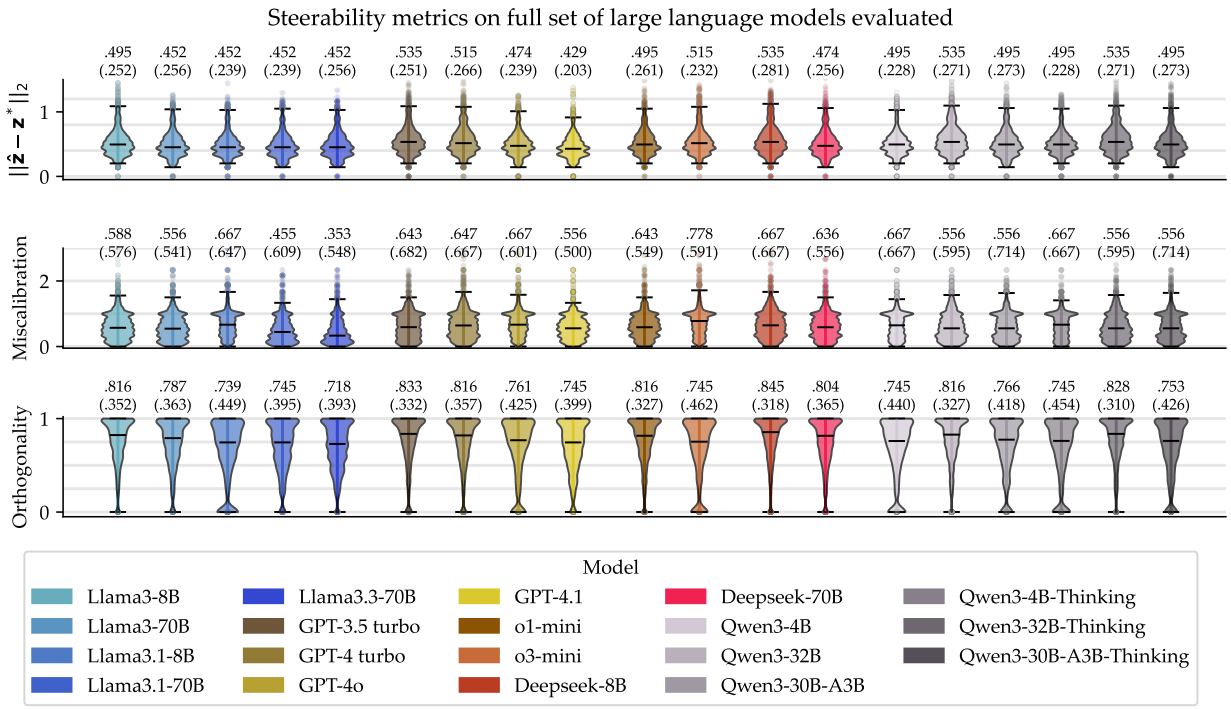


Figure B.1: Median and IQR steerability metrics for (from left to right) Llama3 (blue), GPT (yellow), o1/o3-mini (orange), Deepseek (red), and Qwen3 (gray) family models.

Correlated vs. anti-correlated requests, (reading difficulty, formality) subspace. We show a subgroup analysis of prompts requesting correlated (same direction) vs. anti-correlated (opposite direction) changes in reading difficulty and formality in Figure B.2, for the largest model we evaluated in each class (Llama3.3-70B, GPT-4.1, Deepseek-70B, and Qwen-32B).

All prompts. Here, we show results for all prompting strategies evaluated on our steerability probe. Fig. B.3 shows that prompt engineering has marginal effects on steerability metrics. Providing more detail compared to a direct prompt (e.g., via instructions, including self-generated via chain-of-thought, or specifying a 1-10 scale) as well as negative prompting yield minor improvements to steerability metrics.

The impact of reweighting. In Figure B.4, we find that sampling source texts to target a uniform goal-space (left; darker) vs. naive sampling (right; lighter) yields similar steerability metrics. The no-reweighting probe has slightly worse steerability metrics, suggesting that steerability is actually slightly more difficult on more frequently-encountered goals, though the difference is minor. Ultimately, reweighting goals to target a uniform goal-space remains principled, and ensures that differences in steerability metrics across models or other interventions are not driven solely by improvements in frequent or easy goals.

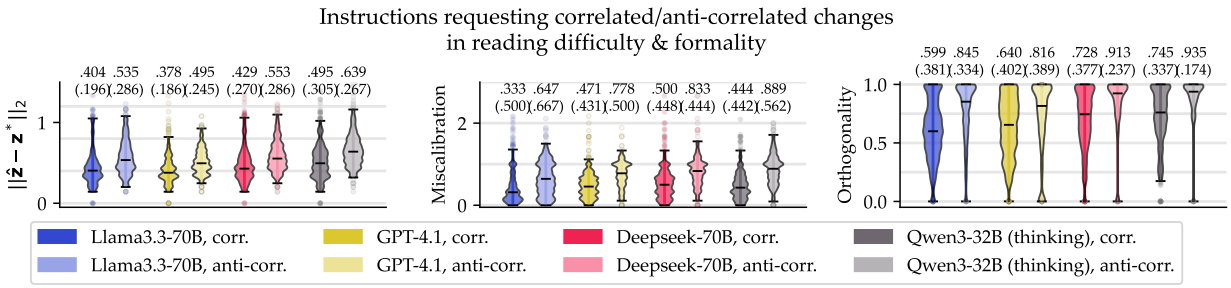


Figure B.2: Median and IQR steerability metrics for (from left to right) Llama3.3-70B (blue), GPT-4.1 (yellow), Deepseek-80B (red), and Qwen3-32B (gray), stratified by requests for correlated (darker) vs. anti-correlated (anti-correlated) changes to reading difficulty and formality. Across all model families, LLMs struggle to satisfy in anti-correlated requests more so than correlated requests.

B.1.4.2 Flow diagrams

Here, we display flow diagrams for all pairs of (requested goal, non-requested goal) for a subset of all models and prompting strategies tested. For transparency, we show both the observed goal-space movement as well as the interpolated flows. Results are organized as outlined below at the end of the Appendix.

Flow diagrams, various models. We show flow diagrams for GPT-4.1 (Figures B.9 through B.20, inclusive). Note that all such experiments use a direct + negative prompting strategy.

Pre- vs. post-RL flow diagrams. We also show flow diagrams for pre- vs. post-RL:

- Pre-RL, reading difficulty specified, formality not specified: Figure B.5
- Pre-RL, formality specified, reading difficulty not specified: Figure B.6
- Post-RL, reading difficulty specified, formality not specified: Figure B.7
- Post-RL, formality specified, reading difficulty not specified: Figure B.8

We observe that, before RL, the model exhibits no movement on many source texts, indicative of the copy-pasting behavior noted in Section 4.4.3, which artificially lowers orthogonality. After RL, the model consistently exhibits movement in many source texts with similar orthogonality to the base model. This result suggests that RL allows the model to independently discover a strategy for mitigating side effects. While side effects visually improve with respect to the base model (less vertical movement), non-trivial vertical movement is still visible post-RL, especially in instructions to change formality, but not reading difficulty (Figure B.8). Thus, room for improvement remains.

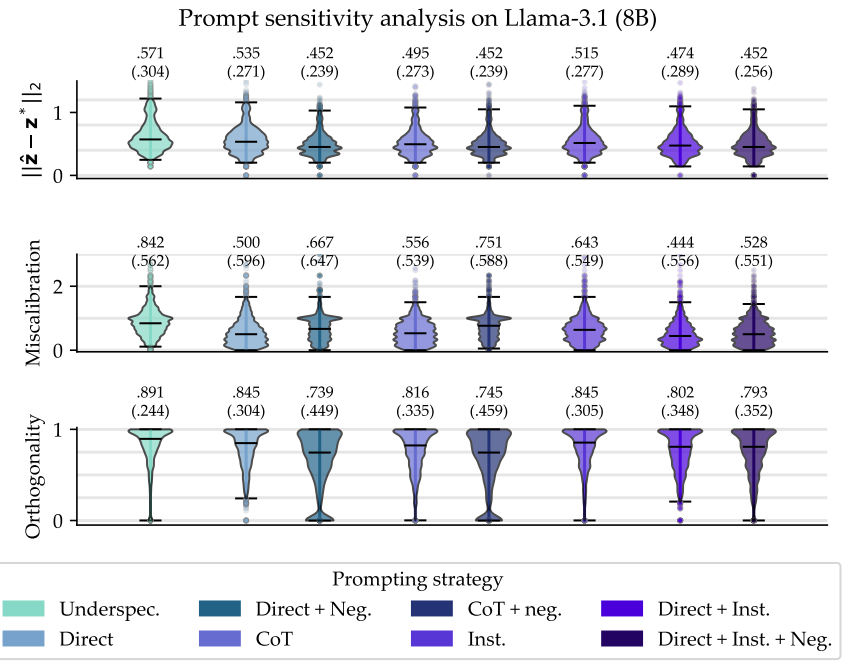


Figure B.3: Median and IQR steerability metrics for various prompting strategies, Llama3.1-8B (from left to right): underspecified, direct, direct with negative prompt, chain-of-thought, chain-of-thoughts with negative prompt, instruction-only, direct with instructions, and direct with instructions and negative prompt.

B.1.4.3 Examples of rewritten texts pre- and post-RL

Here, we show two examples of rewritten texts pre- and post-RL sampled from our steerability probe that demonstrate different rewriting "techniques" learned via RL. For ease of visualization, we truncate the texts and verify that, post-truncation, all excerpts correspond to the same part of the written text (e.g., same events are described). To aid interpretation, we report steerability metrics as well as relevant goal-dimension metrics (normalized and unnormalized Flesch-Kincaid and Heylighen-Dewaele scores). We provide commentary on all rewrites as well, though we caution that analysis is specific to the texts visualized and should not be taken as general statements for all rewrites.

Copy-pasting behavior. Table B.4 shows an example of a source text that is copy-pasted by the base model, but not so by the model post-RL. For ease of qualitative analysis, the shown example is the example with the lowest BLEU score between the source text and rewritten text post-RL, conditioned on being copy-pasted by the original model.

TL;DR: the post-RL uses adverbs and interjections to decrease formality. Though it avoids copy-pasting, the model introduces filler phrases that inflate average sentence length, causing

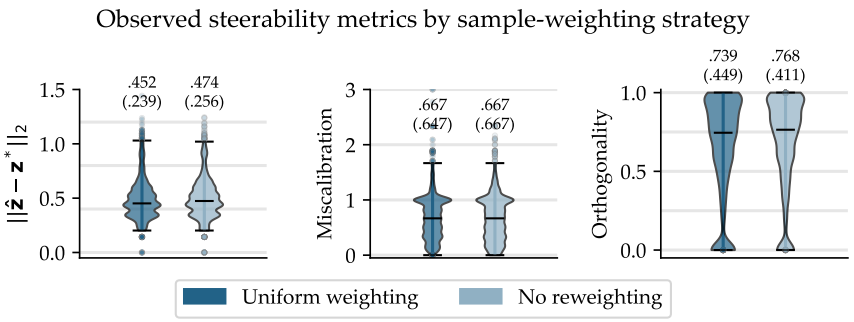


Figure B.4: Median and IQR steerability metrics for a probe sampled with (left) and without (right) uniform reweighting. Steerability metrics remain similar across probes, but reweighting remains a principled approach to ensure that differences in steerability metrics are not dominated by common/easy goals.

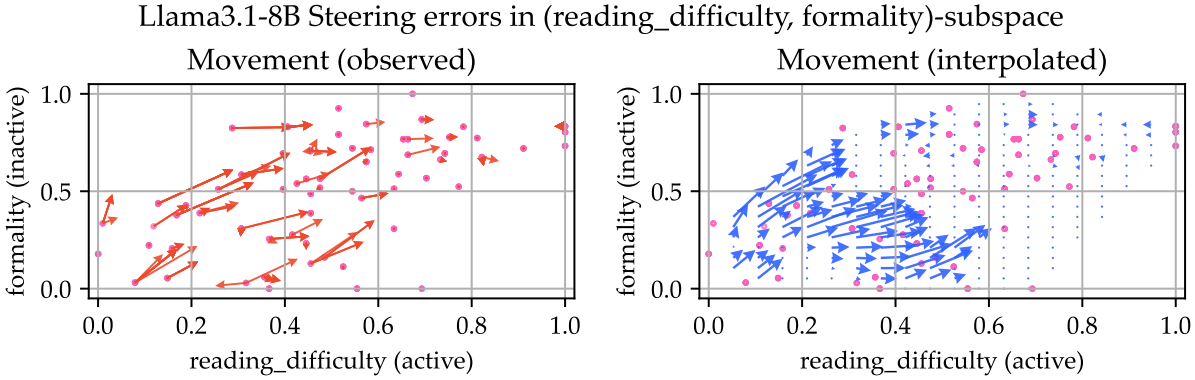


Figure B.5: Llama3.1-8B flow diagram on steerability-tuning evaluation set prior to RL on instructions where reading difficulty is specified, but formality is not.

a side effect in reading difficulty. Post-RL, the model correctly follows the instruction to make the text more informal, as shown by a drop of 17.3 Heylighen-Dewaele points. Such a value indicates that the part-of-speech distribution has shifted in favor of non-diectic text by 34.6%, as evidenced by the increased prevalence of adverbial rewrites such as "thrust into the limelight" $\rightarrow$ "totally getting roasted", and "after accusing" $\rightarrow$ "all like super angry at this guy." In particular, the proportion of adverbs and interjections increases by 9.9% (4.3% $\rightarrow$ 14.2%) and 7.3% (0.1% $\rightarrow$ 7.4%), respectively, supporting the analysis. Note that more colloquial language is not explicitly rewarded by Heylighen-Dewaele.

However, despite unlearning copy-pasting behavior, the post-RL rewrite introduces higher reading difficulty, as the Flesch-Kincaid score goes from 9.9 to 12.6. While the increase in reading difficulty may appear counterintuitive due to the increased usage of colloquialisms in the rewritten text, Flesch-Kincaid monotonically increases in the number of words per sentence, and the average

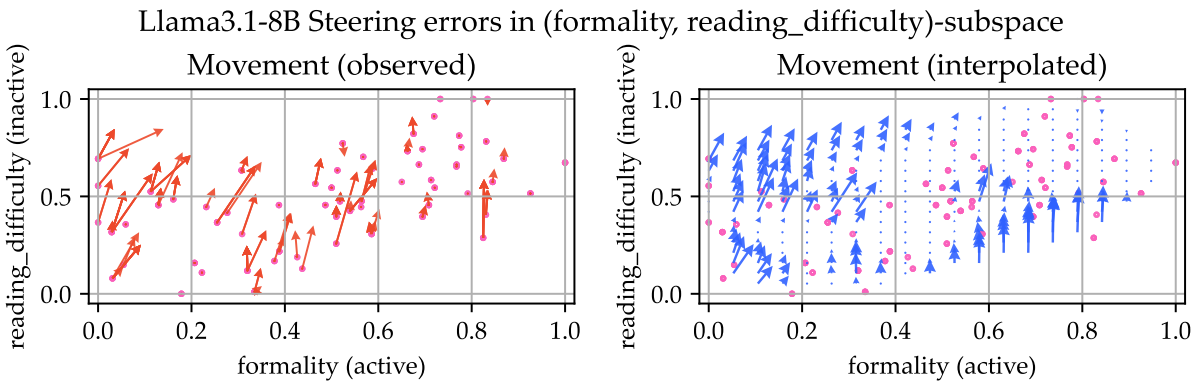


Figure B.6: Llama3.1-8B flow diagram on steerability-tuning evaluation set prior to RL on instructions where formality is specified, but reading difficulty is not.

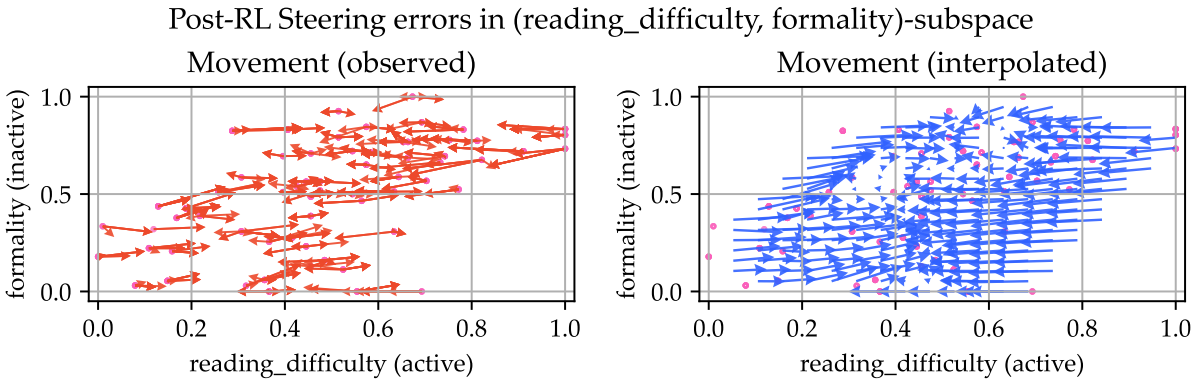


Figure B.7: Llama3.1-8B flow diagram on steerability-tuning evaluation set after RL on instructions where reading difficulty is specified, but formality is not.

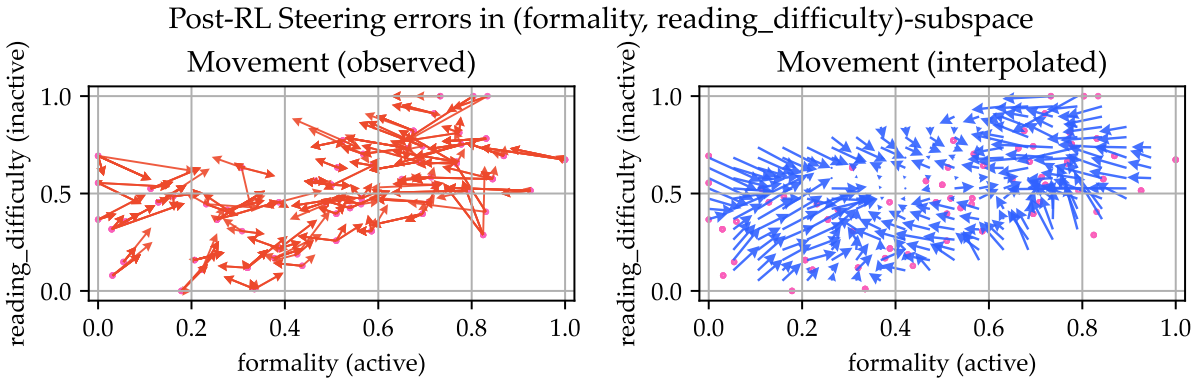


Figure B.8: Llama3.1-8B flow diagram on steerability-tuning evaluation set after RL on instructions where formality is specified, but reading difficulty is not.

syllables per word. Indeed, the number of words per sentence in the post-RL rewrite spikes from 23.0 to 33.0 words per sentence on average, accounting for an increase in 3.9 Flesch-Kincaid grade levels, while the average syllables per word decrease by 0.1, accounting for a decrease of 1.2 grade levels, yielding the observed net 2.7 increase. The increase in words per sentence is likely introduced by the additional adverbial phrases/filler words used in the more informal rewrite. While arguably less brittle than copy-pasting, such filler-phrase usage is may also be detrimental to steerability in the reading difficulty-formality subspace.

Goal disentanglement techniques used post-RL. Table B.5 shows a rewritten text exhibiting disentangled adjustment of reading difficulty and formality on an anti-correlated prompt (increase formality; decrease reading difficulty). To facilitate qualitative interpretation, we choose the prompt with the largest improvement in unnormalized orthogonality on the 2D evaluation probe.

TL;DR: Both models increase formality, but the pre-RL model uses longer words, causing a side effect in reading difficulty. The post-RL model also uses longer words, but mitigates the side effect by using shorter sentences. The instruction provided to the model requires increasing formality, but decreasing reading difficulty. While the pre-RL model increases both goal dimensions, the post-RL model is "directionally correct" in both dimensions. The pre-RL rewrite significantly increases formality, eschewing pronouns (e.g., "i [sic] know x from high school" $\rightarrow$ "The individual in question, who shall be referred to as X, is an acquaintance from high school", or "she is going back to toronto" $\rightarrow$ "X intends to return to Toronto"). The former phrase also replaces the verb "know" with an article+noun ("an acquaintance"). Such edits result in increases to the Heylighen-Dewaele score, and indeed, the proportion of pronouns decreases in the base model's rewrite by 12.2% (20.9% $\rightarrow$ 8.7%), while the proportion of nouns increases by 13.2% (8.8% $\rightarrow$ 22.0%). However, the base model also relies heavily on introducing more polysyllabic words (average words per sentence: 1.1 pre-RL to 1.7 post-RL, accounting for a increase of 11.8 * 0.6 = 7.1 grade levels), such that the Flesch-Kincaid score increases extraneously.

The post-RL model avoids such extraneous increases to the Flesch-Kincaid score. While the post-RL rewrite still increases the number of syllables per word (1.1 $\rightarrow$ 1.4), it compensates by decreasing the sentence length (original: 22.5 words/sentence; post-RL rewrite: 12.5 words/sentence), which the base model fails to do (pre-RL rewrite: 22.0 words/sentence). The model uses similar techniques to increase formality, using nouns to describe events (e.g., "well this happened" $\rightarrow$ "The events of last night"). Interestingly, instead of eschewing pronouns, the model sometimes adds more details to the rewrite, introducing adjectives and verbs that increase Heylighen-Dewaele (e.g., "she went to study to toronto after" $\rightarrow$ "She relocated to Toronto for further education after completing her *high school education*;" italics added for emphasis).

This example demonstrates that, on some source texts, a model trained via RL is able to

disentangle reading difficulty and formality. However, we emphasize that the high variance in orthogonality means that some undesired correlation in model behaviors remains.

B.1.4.4 Groundedness evaluation

In the tables below, we report the number of valid rewrites in each probe, as judged by the pipeline in Appendix B.1.2.3. Results are organized as follows:

- Table B.6: Groundedness counts for sensitivity analysis of prompting strategies on Llama3.1-8B.
- Table B.7: Groundedness counts for best-of-N response on Llama3.1-8B. Note that groundedness is only evaluated on "best" response (*i.e.*, the generation with the lowest steering error).
- Table B.8: Groundedness counts for all models evaluated (direct + negative prompt).

Note that the number of grounded responses (column: "**Count (%)**") is not necessarily equal to 2,048 (total probe size) minus the number of overruled responses. The total number of grounded responses can be lower if any "Yes" responses are also overruled during human review.

B.1.4.5 Examples of real-world user requests

Here, we show a non-exhaustive collection of conversation IDs of real-world user requests in the WildChat dataset [307] related to our choice of goal-dimensions. Conversation IDs were found using the WildVis interactive search tool [64] via a keyword based search, filtering to English-language conversations.[4] Retrieved IDs were then manually reviewed to verify that intents to modify text in a certain manner were present in the conversation.

Note that requests to produce text with a specific value along a goal-dimension without a source text are excluded (e.g., "write a story about a pool table tournament at a grade 1 reading level", ID: 1013074). However, requests that remain "active" across multiple turns are included (*e.g.* "Please rewrite all the following paragraphs that I send to a 9th grade reading level with engaging language that doesn't change the meaning," ID: 26308591fd814f779cd8511e3e449d61), since we merely aim to showcase examples where users desire specific rewrites to text without anchoring to a prompt format. Feedback given to the model expressing an intent (e.g., "write less formal [sic]" in response to a first draft of a text, ID: 2090aca2b7bb4ef8b069e1d43943b007) is also counted.

Warning: Note that linked examples include those flagged as toxic by WildVis.

[4] https://wildvisualizer.com/

Reading difficulty. The following conversation IDs contain intents to modify the reading level.

- **Keywords/keyphrases**: reading level, reading difficulty, advanced
- **IDs**: 1339815, 1368597, 1abecad7158d426282e35c0c91106206, 271c9d0b08b749bbbf67527aa98c08c7, 1262863, 1089442, f075b8061bb6449390c3368207af745b, 314190da3eff40a2838773249c9167d8, 1373275, 1368816, 1f9202e58d3a451998a230c268c292a1, 1947279, 350789

Formality. The following conversation IDs contain intents to modify text formality.

- **Keywords/keyphrases**: formal, formality, informal
- **IDs**: 09b203dc594e4ee7a0fc79dad9efa69d, e0f245efe54943aaa9889299a8599cc3, 600c946054b447de9969925ae27bc09a, cc7dfeaa931e4c3c9f21bec5c22132e6, 37a428638c134d128a67f81e1949156b, 9f71f443b9764eb094173bbc39705a70, 62a2597fc186451c968582b8a0af6a3f, 231912e90cbe4fa28afe282099602139, 3bb584091ce64b6b84906792fc75397d, 2090aca2b7bb4ef8b069e1d43943b007, 44f9d6df33084db2b9a6278ce2407140

Textual diversity. The following conversation IDs contain intents to modify text diversity.[5]

- **Keywords/keyphrases:** repetitive
- **IDs:** 2474721, 58a5f9af506e4f14a3219f352a509c6c, 20cc3c53d38d42bab096e4715eae85f5, bf6efffd89c24f38affb83b24e85e622,

Text length. The following conversation IDs contain intents to modify text length.

- **Keywords/keyphrases:** longer, shorter, concise, verbose
- **IDs:** 6e38bdb4034c4bc0a6b63645e2f49166, fcf6bb4ea8a91b552a1a38db3c32a2ee33ef72a619156, 24766f0ee8a149ef962cc1e8c41f2bf0, 74cd6c3ffb8d46db87e7b5a8c3b87a0f, 432ee46108db4f3aa6a6cb5e36e5ac9f, 4e6723ddb6784b40805f1120ce04c5ac, b52aa0765d1f40e9a276d93feb4403b3, ec055ca5800640158be7639bdb9b073d, 06f78ce84e8045faaf5270c561150db3, 8676fcded9fb4be6b1398cc2eeeec995,

[5]Other keywords for which we were unable to find relevant conversation IDs include: diversity, variety, diverse vocabulary, variety of words. Our search was not exhaustive due to the high false positive rate for "diversity" and "variety."

061284abd9dc409e8beb90dc59e88849, 3bef1c236ecb49b4ae69f72be6947ce3, d5821704834b4866808abf6d642d1b16, 2667173, 175693, 366339, 70f1f211c9f44096a635b08ac8f6e311

Other intents. The following conversation IDs contain assorted intents, which we categorize. Note that some IDs appear twice since they request changes to multiple aspects of text.

- **Keywords/keyphrases:** rewrite, make it, change the, make this, not enough, no make it
- **Aspect change:** 1793b1ad8f1748a98f7889ee2bb52962 (clarity), 73ec56e74cc244cabdda9a1a88e5e597, 2734150 (politeness), 973648 (aggressive), 1019036, 1124733 (historical language), 217907 (allegorical), 6b1578ff9de944bf9b7e8d0f45f29e69 (scientific)
- **Stylistic change:** cffa19ef42df4e49a0e8aad86947722c (dramatic), cffa19ef42df4e49a0e8aad86947722c (romantic), 68b0dd0186da45debfb300acedae5013 (modern), 2274771 (textbook-like), eb63ef22d66846f18f65b5680e5437a7 ("more human")
- **Toxicity:** f2a4cab2129b4f4a9088c74656fd61b8
- **Underspecified:** 54324834ac124a2fa04e267e057976ff ("change the format more"), 390c70b49ae0416e850cbe3c92617c08 ("more awesome"), 390c70b49ae0416e850cbe3c92617c08 ("make X better"), 75fb787148674c81a66549464ed6d724, 1144654, 1814613, 331820, 434828 ("rewrite/again"), 3e0c5523d09a47a68e261ddaa04263c9 ("improve"), 1947139, 2033487, 1925167, 1990811 ("edit")
- **best-of-N/re-prompting:** 2113066, 0628c4bfaf1541e4879f5a96a215c1fb

B.1.5 Computational details

Software. We used a custom fork of `trl` 0.16.0 (License: Apache 2.0, [285]) for model training, along with vLLM 0.8.4 for fast model rollouts (License: Apache 2.0 [159]), `deepspeed` [234] (License: Apache 2.0) and `cut-cross-entropy` [296] (License: Apple) for memory-efficient training,`accelerate` for multi-GPU training (License: Apache 2.0 [100]), `peft` for LoRA (License: Apache 2.0 [183]), and FlashAttention-2 (License: BSD-3 [57]). The PyTorch version is 2.6.0 (License: BSD-style [212]). with CUDA 12.4. During inference, LLM API calls were made via SAMMO (License: MIT [245]). Various text-processing packages were used to compute goal dimensions, namely, `nltk` (License: MIT) and `spacy` (License: MIT), `textstat` (License:

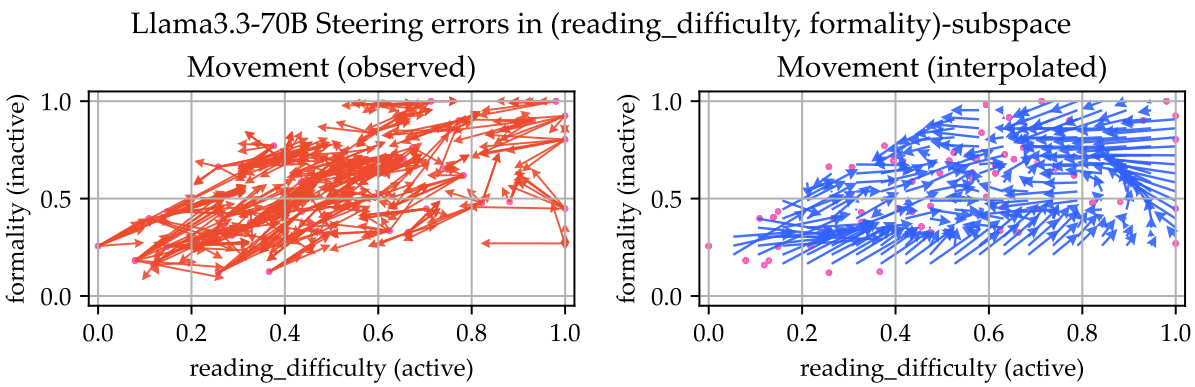


Figure B.9: Llama3.3-70B flow diagram, (reading difficulty, formality) subspace.

MIT), `taaled` (License: CC BY-NC-SA 4.0), and `pylats` (License: CC BY-NC-SA 4.0), which were hosted locally as a server via Uvicorn (License: BSD-3) and FastAPI (License: MIT) with Pydantic-based type-validation (License: MIT) during training. Fastsafetensors (License: Apache 2.0) was used to load LoRA adapters from trained models via vLLM. Scikit-learn (License: BSD-3) was used to implement classifier-based density ratio estimation for the purpose of computing sampling weights.

Hardware. All experiments were run on 4 GPUs (RTX A6000, 48GB VRAM) on an 8 GPU Ubuntu 22.04.5 machine with 256 CPUs (processor type: AMD EPYC 7763 64-Core)

Estimated compute. Main training runs took six GPU-days ($\approx$ 1.5 days $\times$ 4 GPUs). While inference took 30 GPU-minutes per model (smallest models; e.g., Llama3.1-8B) to 4 GPU-hours per model (approx. 1 hr * 4 GPUs). Best-of-N approaches took up to 8 GPU hours ($N = 128$); we approximate the total compute time for such experiments as 16 GPU hours. Experiments reported in the paper constitute a total of approximately 25 GPU-days (inference: approx. 27.5h; training: approx. 24 days). Additional preliminary training and inference experiments for debugging and exploration required approximately 50 additional GPU-days, with the vast majority of additional compute devoted to model training.

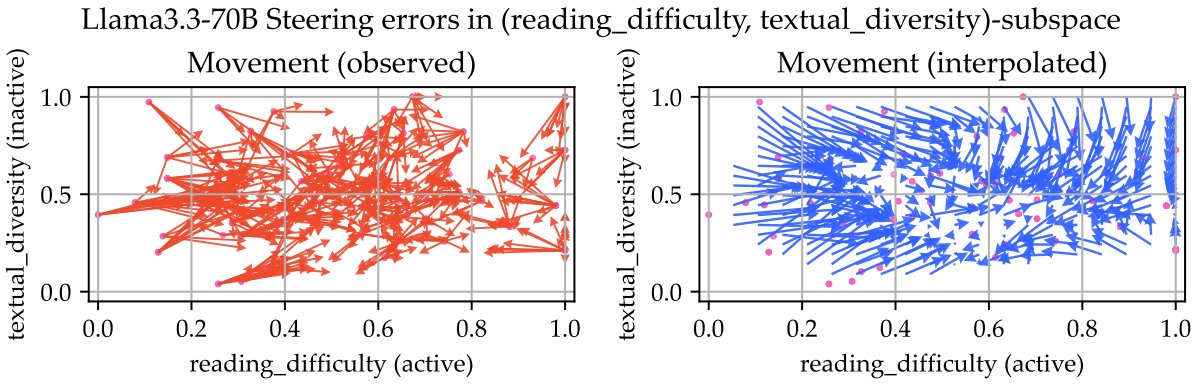


Figure B.10: Llama3.3-70B flow diagram, (reading difficulty, textual diversity) subspace.

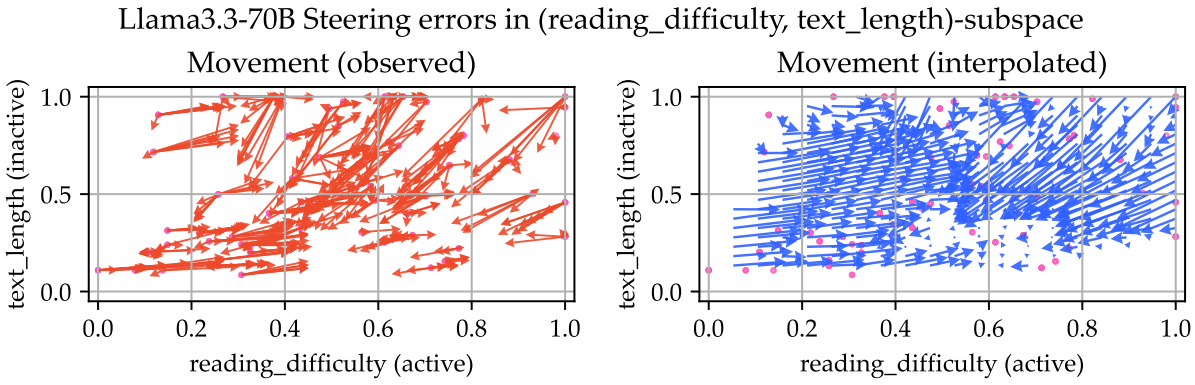


Figure B.11: Llama3.3-70B flow diagram, (reading difficulty, text length) subspace.

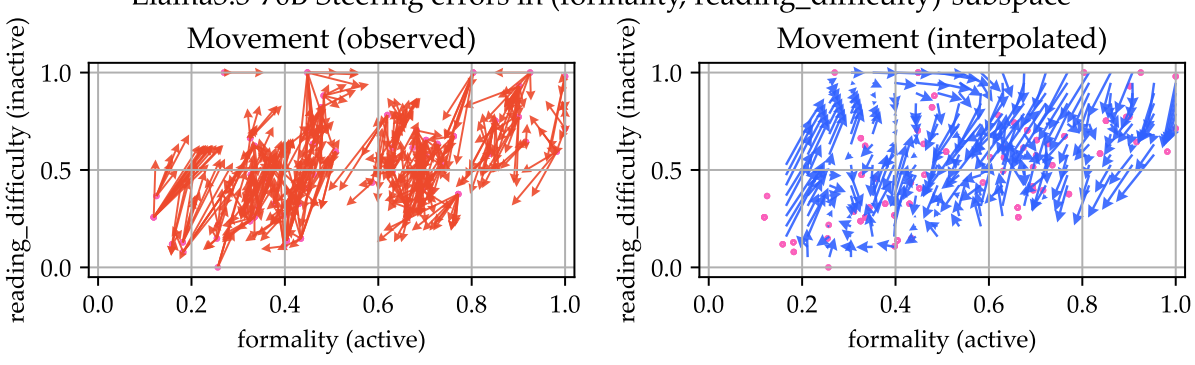


Figure B.12: Llama3.3-70B flow diagram, (formality, reading difficulty) subspace.

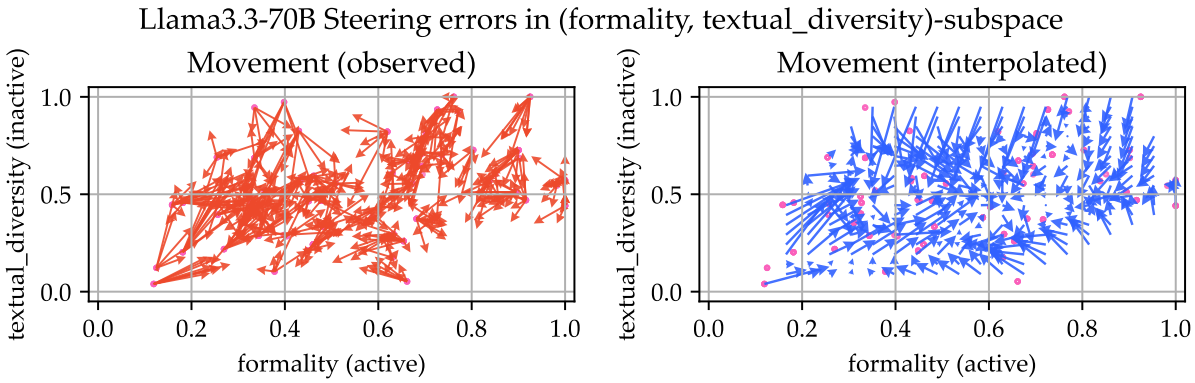


Figure B.13: Llama3.3-70B flow diagram, (formality, textual diversity) subspace.

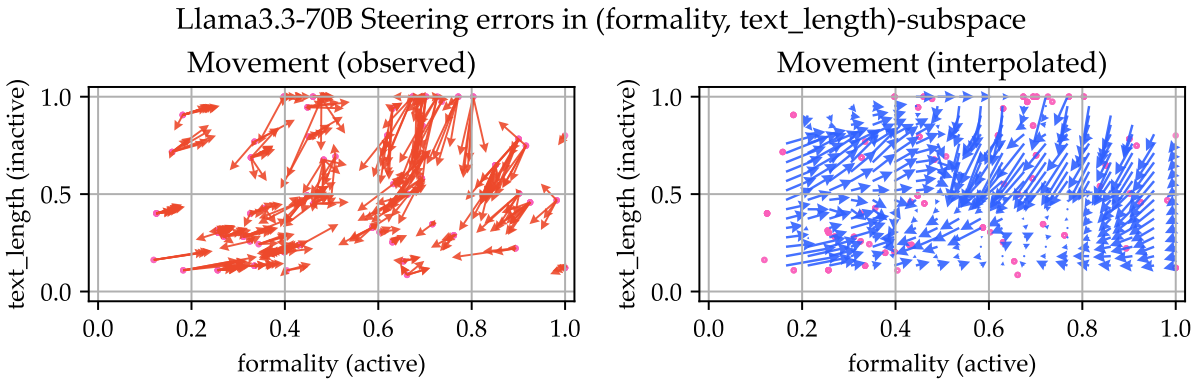


Figure B.14: Llama3.3-70B flow diagram, (formality, textual length) subspace.

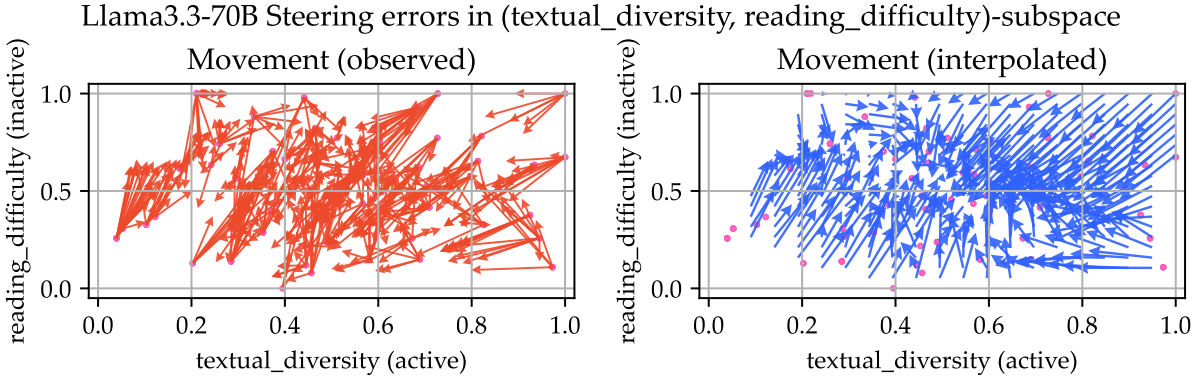


Figure B.15: Llama3.3-70B flow diagram, (textual diversity, reading difficulty) subspace.

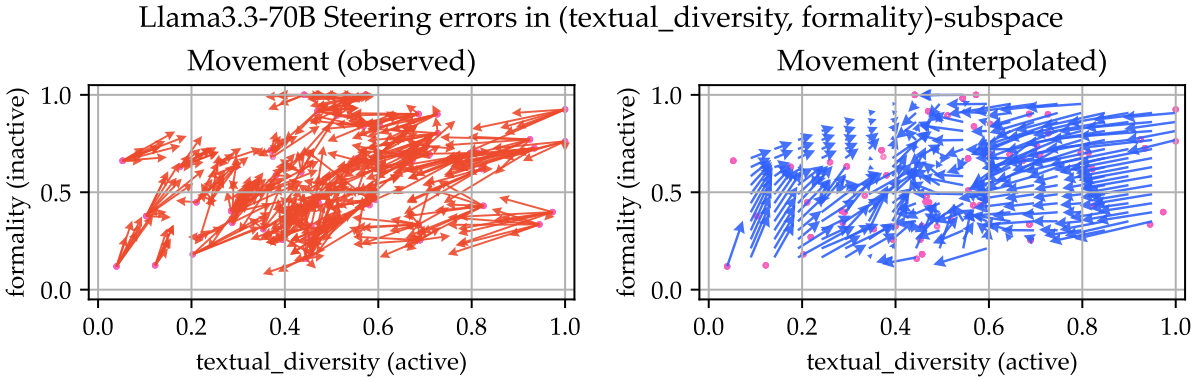


Figure B.16: Llama3.3-70B flow diagram, (textual diversity, formality) subspace.

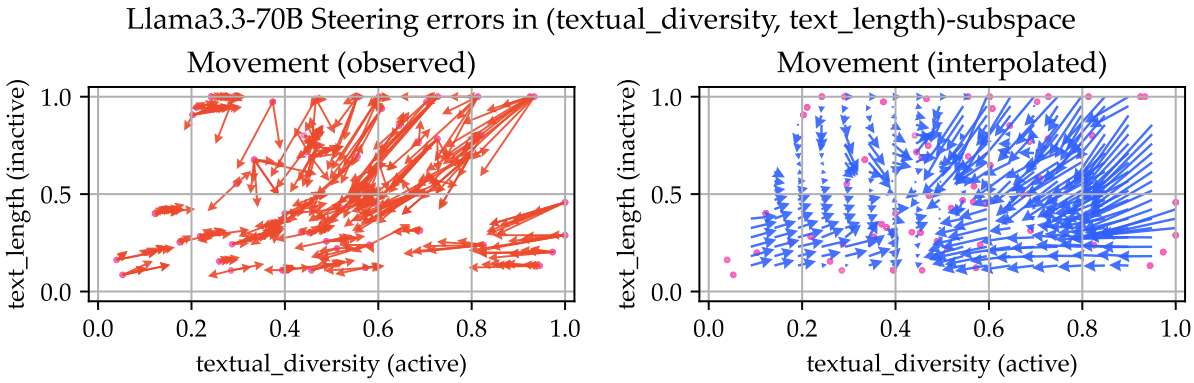


Figure B.17: Llama3.3-70B flow diagram, (textual diversity, text length) subspace.

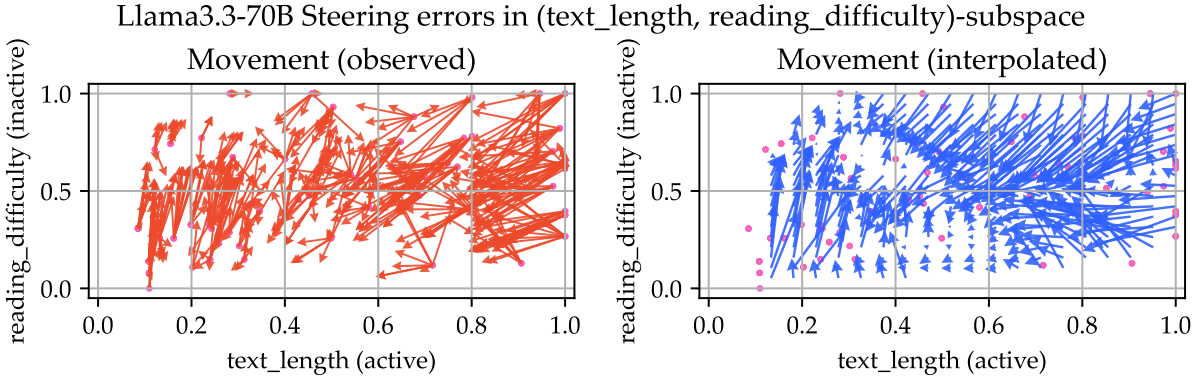


Figure B.18: Llama3.3-70B flow diagram, (text length, reading difficulty) subspace.

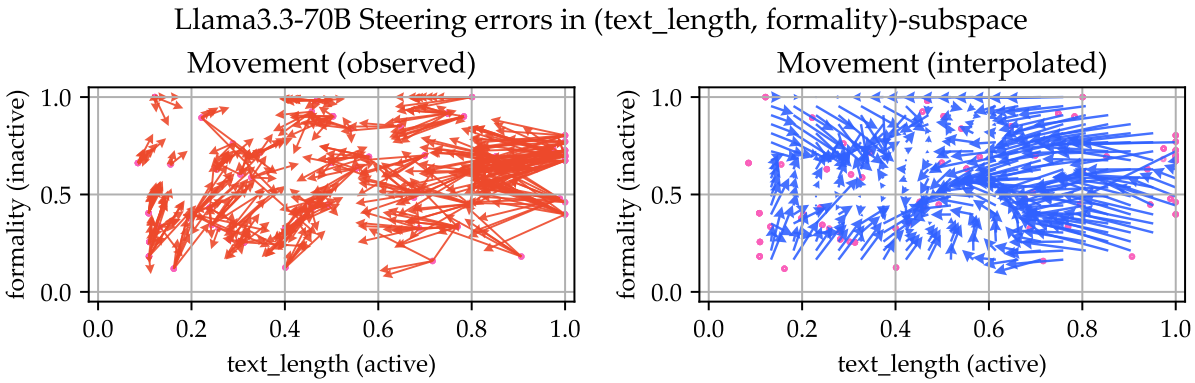


Figure B.19: Llama3.3-70B flow diagram, (text length, formality) subspace.

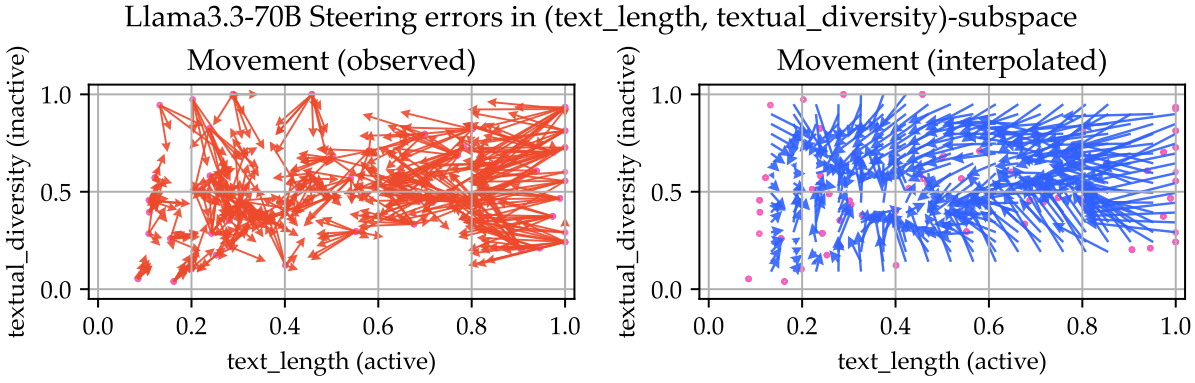


Figure B.20: Llama3.3-70B flow diagram, (text length, textual diversity) subspace.

	Original/pre-RL	Post-RL
	She's suddenly been thrust into the limelight after accusing a talent show contestant of stealing someone else's style - but according to users on Twitter, Natalia Kills's own look is far from original. The Bradford-born singer, 28, was fired from the judging panel of New Zealand's X Factor, along with her husband Willy Moon, after accusing hopeful Joe Irvine of copying her spouse on Sunday's show. However, viewers have taken to social media to point out that Natalia appears to have turned to a range of different artists including Jessie J, Lady Gaga and even Johnny Depp for fashion tips of her own. Scroll down for video . Natalia launched an angry tirade at X Factor New Zealand contestant Joe Irvine, she said that she was 'disgusted,' at the fact that she thought he looked like her husband, Willy Moon .	So like Natalia Kills is just totally getting roasted on Twitter right now, and it's all because she was all like super angry at this guy Joe Irvine on X Factor New Zealand, and she was all 'he's copying my husband' and stuff, but honestly, it's kinda like, she's the one who's been copying loads of other people's styles, you know? She's this singer from Bradford, and she's 28, and she's all like super famous now, but she got fired from X Factor with her husband Willy Moon [...]
FK Score (norm.)	9.90 (0.703)	12.6 (0.970)
HD Score (norm.)	56.7 (0.568)	39.4 (0.000)
Steering err.	0.564	**0.267**

Table B.4: Example showing copy-pasting behavior in the pre-RL base model un-learned after RL, sourced from CNN/DailyMail validation split. [...] added to indicate mid-sentence truncation. FK: Flesch-Kincaid (reading difficulty). HD: Heylighen-Dewaele (formality). Note that metrics are for the entire source text and may not match the excerpt provided. **Instruction** (immediately before source text): "Please rewrite the following, but make it much more informal. You MUST not change anything else about the other parts of the text, even if it makes the rewritten text sound unnatural or otherwise awkward. Respond with only the rewritten text and do not explain your response."

	Original	Pre-RL	Post-RL
	well this happened last night past midnight. a friend had a good bye party because she's going to italy. take into account that the night out lasted from 9pm to 8am. the girl i like was also going, we'll call her x. i know x from high school but she went to study to toronto after. she was here for the last month to visit so we've seen each other a lot because of the friends we have in common. she is going back to toronto on tuesday and she doesn't know if she'll be coming back again since she'll probably stay there to work too.	The aforementioned incident transpired last night, subsequent to the stroke of midnight. A friend had convened a farewell gathering in anticipation of her forthcoming relocation to Italy. It is pertinent to note that the nocturnal excursion commenced at 9pm and culminated at 8am. The individual in question, who shall be referred to as X, is an acquaintance from high school who had relocated to Toronto for academic purposes. However, she had recently returned to our vicinity for a month-long visit, thereby facilitating numerous encounters with the narrator due to their shared acquaintances. It is worth mentioning that X intends to return to Toronto on Tuesday, with the likelihood of her remaining in that city for professional purposes.	The events of last night occurred after midnight. A friend had a farewell party in celebration of her departure to Italy. It is essential to note that the evening lasted from 9pm to 8am. The individual I have a romantic interest in will be referred to as x. I am acquainted with x from high school. She relocated to Toronto for further education after completing her high school education. x was in the area for a month to visit and as a result, we had numerous opportunities to interact due to our shared acquaintances. x is scheduled to return to Toronto on Tuesday. It is uncertain whether she will return to the area in the future as she intends to pursue employment in Toronto.
FK Score (norm.)	6.00 (0.317)	13.2 (1.000)	5.80 (0.297)
HD Score (norm.)	41.2 (0.029)	64.0 (0.821)	53.7 (0.464)
Steering err.	–	0.623	**0.366**

Table B.5: Example showing the model demonstrating the ability to disentangle goals post-RL, sourced from Reddit TIFU. FK: Flesch-Kincaid (reading difficulty). HD: Heylighen-Dewaele (formality). Note that metrics are for the entire source text and may not match the excerpt provided. **Instruction** (immediately before source text): "Please rewrite the following, but make it more formal, and easier to read. You MUST not change anything else about the other parts of the text, even if it makes the rewritten text sound unnatural or otherwise awkward. Respond with only the rewritten text and do not explain your response."

Prompt strategy	NP?	Count (%)	# (%) LLM flagged	# (%) overruled
Direct	No	2047 (99.95%)	6 (0.29%)	5 (0.24%)
Direct	Yes	2042 (99.71%)	6 (0.29%)	2 (0.10%)
Underspecified	No	2048 (100.0%)	0 (0.00%)	0 (0.00%)
1-10 scale	No	2043 (99.76%)	13 (0.63%)	10 (0.49%)
1-10 scale	Yes	2044 (99.80%)	15 (0.74%)	11 (0.54%)
Inst.-based	No	2044 (99.80%)	5 (0.24%)	1 (0.05%)
Direct + inst.	No	2047 (99.95%)	5 (0.24%)	4 (0.20%)
Direct + inst.	Yes	2048 (100.0%)	4 (0.20%)	4 (0.20%)
CoT	No	2048 (100.0%)	2 (0.10%)	2 (0.10%)
CoT	Yes	2048 (100.0%)	2 (0.10%)	2 (0.10%)

Table B.6: Counts of grounded rewrites in main steerability probe for Llama3.1-8B under different prompting strategies, with number of rewrites flagged as un-grounded by the LLM, and number of LLM-flagged rewrites overruled. NP: negative prompt. Inst.: instruction(s).

N (rewrites per prompt	Count (%)	# (%) LLM flagged	# (%) overruled
1 (zero-temp.)	2042 (99.71%)	6 (0.29%)	2 (0.10%)
4	2048 (100.0%)	1 (0.05%)	1 (0.05%)
8	2047 (99.95%)	9 (0.44%)	8 (0.39%)
16	2048 (100.0%)	4 (0.20%)	4 (0.20%)
32	2048 (100.0%)	4 (0.20%)	4 (0.20%)
64	2046 (99.90%)	3 (0.15%)	2 (0.10%)
128	2048 (100.0%)	1 (0.05%)	1 (0.05%)

Table B.7: Counts of grounded rewrites in main steerability probe under different best-of-N strategies. Note that only the lowest steering-error response for each was evaluated.

N **(rewrites per prompt)**	**Count (%)**	**# (%) LLM flagged**	**# (%) overruled**
Llama3-8B	2048 (100.0%)	8 (0.39%)	8 (0.39%)
Llama3-70B	2048 (100.0%)	8 (0.39%)	8 (0.39%)
Llama3.1-8B	2042 (99.71%)	6 (0.29%)	2 (0.10%)
Llama3.1-70B	2047 (99.95%)	6 (0.29%)	7 (0.34%)
Llama3.3-70B	2047 (99.95%)	19 (0.93%)	18 (0.88%)
GPT-3.5 turbo	2045 (99.86%)	41 (2.00%)	40 (1.95%)
GPT-4 turbo	2048 (100.0%)	0 (0.00%)	0 (0.00%)
GPT-4o	2048 (100.0%)	0 (0.00%)	0 (0.00%)
GPT-4.1	2048 (100.0%)	0 (0.00%)	0 (0.00%)
o1-mini	2026 (98.93%)	23 (1.12%)	1 (0.05%)
o3-mini	2024 (98.83%)	23 (1.12%)	4 (0.20%)
Deepseek-8B	2047 (99.95%)	6 (0.29%)	5 (0.24%)
Deepseek-70B	2046 (99.90%)	4 (0.20%)	4 (0.20%)
Qwen3-4B (no thinking)	2046 (99.90%)	5 (0.24%)	3 (0.15%)
Qwen3-4B (+thinking)	2047 (99.95%)	4 (0.20%)	1 (0.05%)
Qwen3-32B (no thinking)	2045 (99.86%)	2 (0.10%)	4 (0.20%)
Qwen3-32B (+thinking)	2045 (99.86%)	3 (0.15%)	3 (0.15%)
Qwen3-30B-A3B (no thinking)	2047 (99.95%)	5 (0.24%)	0 (0.00%)
Qwen3-30B-A3B (+thinking)	2046 (99.90%)	7 (0.34%)	5 (0.24%)

Table B.8: Counts of grounded rewrites in main steerability probe under different models.

APPENDIX C

Appendix for Gaming and Strategic Adaptation: Facilitating Oversight over Deployment-time Interactions

C.1 Supplementary material for "Who's Gaming the System? A Causally-Motivated Approach to Detecting Strategic Adaptation"

C.1.1 Additional related works

Algorithmic anomaly/fraud detection. Our framework can be understood as an algorithmic auditing method for fraud/anomaly detection. Many approaches assume that ground-truth fraud/gaming labels are available, reducing gaming detection to supervised learning [17, 102, 282]). We do not assume access to such labels. Unsupervised approaches for anomaly/fraud detection [60, 176, 105, 96] generally assume that anomalies are outliers with respect to some distribution. Mixture-modeling approaches similarly assume that fraudulent/non-fraudulent decisions correspond to distributions learnable under restrictive parametric assumptions [76]. Instead of distributional assumptions, we make behavioral assumptions about agents following strategic classification. Existing models of agent behavior in the context of fraud detection make domain-specific assumptions about features predictive of fraud [273], agent utility (*e.g.*, constant penalties [66]), or access to audit labels [273]. We generalize past work by making looser assumptions on agent utility and does not assume access to auxiliary information (*e.g.*, audit labels), and circumvents the need for ground-truth labels by making assumptions about gaming.

C.1.2 Omitted Proofs

Here, we provide detailed proofs for all theoretical results.

C.1.2.1 Proposition 11

Proposition. *Let* $R' \triangleq \frac{d}{dd_p^*}$ *and* $c' \triangleq \frac{d}{dd_p^*}$. *For any agent* p*, given Assumptions 1- 4 and a fixed observation of* $\Delta_p(d_p^*)$*,*

$$\lambda_p \in \left[\frac{R'(\Delta_p(d_p^*))}{c'(\Delta_p(d_p^*))}, \infty\right). \tag{C.1}$$

Proof. Fix some R, c satisfying assumptions Assumptions 1- 4 and an arbitrary $\Delta_p(d_p^*)$. Note that, for all $\Delta_p(d_p^*) \neq d_p^*$, we can write

$$\lambda_p = \frac{R'(\Delta_p(d_p^*))}{c'(\Delta_p(d_p^*) - d_p^*)}. \tag{C.2}$$

Note that this quantity is monotonic in d^*, since the strict convexity of c implies that c' is strictly increasing. Hence, we can substitute an upper and lower bound on d^* to obtain our desired result. First, since $d^* \in [0, \Delta_p(d_p^*))]$, we can substitute $d^* = 0$ to reach

$$\lambda_p \geq \frac{R'(\Delta_p(d_p^*))}{c'(\Delta_p(d_p^*))}. \tag{C.3}$$

Before considering the case where $d^* = \Delta_p(d_p^*))$, we note that a direct substitution yields $c'(\Delta_p(d_p^*) - \Delta_p(d_p^*)) = c'(0) = 0$, since c is strictly convex and minimized at 0 via Assumption 3. However, we can use a limiting argument since c is differentiable and therefore continuous:

$$\lim_{d^* \to \Delta_p(d_p^*)^-} \lambda_p = \lim_{d^* \to \Delta_p(d_p^*)^-} \frac{R'(\Delta_p(d_p^*))}{c'(\Delta_p(d_p^*) - d^*)}. \tag{C.4}$$

To evaluate the limit, we treat R' as a positive constant (via Assumption 2) and note that the limit

$$\lim_{d^* \to \Delta_p(d_p^*)^-} c'(\Delta_p(d_p^*) - d^*) = 0, \tag{C.5}$$

from which we conclude $\lim_{d^* \to \Delta_p(d_p^*)^-} = \infty$. Combining with the lower bound on λ_p yields the desired bounds.

□

A causal DAG-based perspective on non-identifiability of λ_p. Consider the causal graph shown in Figure C.1 (top). Since d^*, or the true diagnosis rate, is unobserved, non-identifiability of the causal effect of p on d is immediate. By extension, λ_p is similarly non-identifiable. However, under causal graph of strategic adaptation (Figure C.1, bottom), we assume access to confounders x, which may be fully observed (as shown in the main paper in Figure 5.2, right), or partially observed.

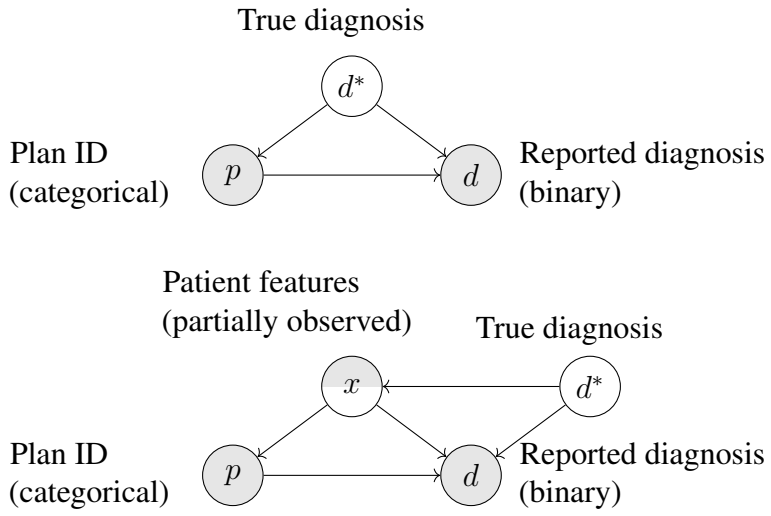


Figure C.1: **Top:** Causal graph for gaming detection with agent indicator p, ground truth diagnosis d^*, and agent decision d. **Bottom:** Causal graph for gaming detection with added partially observed confounders $\mathbf{x}$.

If x is only partially observed, conditional exchangeability (Assumption 6) does not hold, though bounds on the causal effect of p on d may still be achievable through sensitivity analyses that bound outcome probabilities (e.g., [185]) or propensity scores (e.g., [137]), though the resulting bounds on treatment effect estimates may not be tight. Such unobserved confounding is a potential limitation of all observational studies.

In the case of fully observed x, the causal effect is identifiable, but that alone may not imply identifiability of λ_p itself as we can only ensure a monotonic relationship between λ_p and the diagnosis rate under one particular plan p. Intuitively, even when we observe an unbiased estimate of the "treatment effect" comparing two plans p and p', without knowledge of d^* as an anchor, λ_p remains unidentifiable. In theory, x is a sufficient proxy for d^*, and if the relationship between d^* and x is invertible and known, recovering d^* and λ_p is straightforward if we know the utility function. However, without observations of d^* itself it is unclear whether the mapping from d^* to x is identifiable, or whether an invertible mapping even exists.

However, given our goal of identifying worst offenders in terms of λ_p, a ranking is sufficient. Thus, in addition to including x as covariates in estimates of treatment effects, we leverage assumptions from the utility-maximization formulation of gaming borrowed from strategic classification. These assumptions provide additional information not encoded in the causal graph. Namely, we have that $R'(\Delta_p(d_p^*))/c'(\Delta_p(d_p^*) - d_p^*)$ is monotonic in d_p^* and $d_p^* \in [0, \Delta_p(d_p^*)]$, provided that R and c satisfy Assumptions 1- 4. Such assumptions allow us to rank by gaming deterrence parameter λ_p.

C.1.2.2 Theorem 12

Theorem. *Define* $\Delta_{p'}(d_p^*)$ *as*

$$\Delta_{p'}(d_p^*) \triangleq \underset{\bar{\mathbf{d}} \in [0,1]}{\arg\max}\ R(\bar{d}) - \lambda_{p'} c(\bar{\mathbf{d}} - d_p^*) \tag{C.6}$$

Then, given Assumptions 1- 4, $\Delta_p(d_p^*) < \Delta_{p'}(d_p^*)$ *if and only if* $\lambda_p > \lambda_{p'}$.

Proof. For all portions of the proof, let $R' \triangleq \frac{d}{dd_p^*}$ and $c \triangleq \frac{d}{dd_p^*}$.

($\implies$) We have the simultaneous first-order conditions

$$R'(\Delta_p(d_p^*)) - \lambda_p c'(\Delta_p(d_p^*) - d^*) = R'(\Delta_{p'}(d_p^*)) - \lambda_{p'} c'(\Delta_{p'}(d_p^*) - d^*) \tag{C.7}$$

$$\iff \underbrace{R'(\Delta_p(d_p^*)) - R'(\Delta_{p'}(d_p^*))}_{:=C_0 \geq 0} - \lambda_p c'(\Delta_p(d_p^*) - d^*) = -\lambda_{p'} c'(\Delta_{p'}(d_p^*) - d^*) \tag{C.8}$$

$$\iff \lambda_p c'(\Delta_p(d_p^*) - d^*) - C = \lambda_{p'} c'(\Delta_{p'}(d_p^*) - d^*) \tag{C.9}$$

$$\iff \lambda_p = \lambda_{p'} \cdot \underbrace{\frac{c'(\Delta_{p'}(d_p^*) - d^*) + C}{c'(\Delta_p(d_p^*) - d^*)}}_{C_1 > 1} > \lambda_{p'}, \tag{C.10}$$

and hence $\lambda_p > \lambda_{p'}$ as desired. Note that $C_0 \geq 0$ as per Assumption 2, and $C_1 > 1$ as per our assumption that $\Delta_p(d_p^*) < \Delta_{p'}(d_p^*)$, from which we can conclude $c'(\Delta_{p'}(d_p^*) - d^*) > c'(\Delta_p(d_p^*) - d^*)$, because the strict convexity of c implies that c' strictly increases (Assumption 3).

($\impliedby$) Pick any $\lambda_{p'} < \lambda_p$ and fix some d^*. By contradiction; suppose $\Delta_p(d_p^*) \geq \Delta_{p'}(d_p^*)$ as well. Then:

$$R'(\Delta_p(d_p^*)) - \lambda_p c'(\Delta_p(d_p^*) - d^*) = R'(\Delta_{p'}(d_p^*)) - \lambda_{p'} c'(\Delta_{p'}(d_p^*) - d^*) \tag{C.11}$$

$$\iff R'(\Delta_p(d_p^*)) - R'(\Delta_{p'}(d_p^*)) = \lambda_p c'(\Delta_p(d_p^*) - d^*) - \lambda_{p'} c'(\Delta_{p'}(d_p^*) - d^*). \tag{C.12}$$

Now, note that $R'(\Delta_p(d_p^*)) - R'(\Delta_{p'}(d_p^*)) \leq 0$, since $\Delta_p(d_p^*) \geq \Delta_{p'}(d_p^*)$ by assumption and R' is non-increasing in its argument. Furthermore, we can write

$$\lambda_p c'(\Delta_p(d_p^*) - d^*) - \lambda_{p'} c'(\Delta_{p'}(d_p^*) - d^*) > \lambda_{p'}(c'(\Delta_p(d_p^*) - d^*) - c'(\Delta_{p'}(d_p^*) - d^*)) > 0, \tag{C.13}$$

where the first inequality is due to $\lambda_{p'} < \lambda_p$ and factoring, and the second inequality is since $c'(\Delta_p(d_p^*) - d^*) - c'(\Delta_{p'}(d_p^*) - d^*) > 0$ (via Assumption 3) and $\lambda_{(\cdot)} > 0$ (by definition). Returning to Eq. C.12, we can write

$$0 \geq R'(\Delta_p(d_p^*)) - R'(\Delta_{p'}(d_p^*)) = \lambda_{p'} c'(\Delta_{p'}(d_p^*) - d^*) > 0 \tag{C.14}$$

which is a contradiction ($0 \not> 0$). Thus, $\lambda_{p'} < \lambda_p$ implies $\Delta_p(d_p^*) < \Delta_{p'}(d_p^*)$ as desired. □

Potential extensions to upcoding detection. With stronger assumptions, a weaker form of upcoding detection, which is stronger than ranking, is possible:

Assumption 9 (Known cost and reward derivatives)**.** *R' and c' can be evaluated at arbitrary points.*

Assumption 10 (Uncertainty in d_p^* is bounded)**.** *Lower and upper bounds on d_p^* are possible to obtain.*

Then, define the following:

Definition 32 (ε-gaming)**.** *For $\varepsilon > 0$, an agent p is ε-gaming if $\Delta_p(d_p^*) - d_p^* > \varepsilon$.*

In other words, we can define a threshold based on some ε tolerance in deviations from d_p^* for detecting gaming. Now, recall that

$$\lambda_p = \frac{R'(\Delta_p(d_p^*))}{c'(\Delta_p(d_p^*) - d_p^*)}, \tag{C.15}$$

which is *non-increasing* in $\Delta_p(d_p^*)$ due to the concavity of R and strict convexity of c. Thus, substituting, we could define an agent-specific threshold

$$\lambda^*(p) = \frac{R'(\varepsilon + d_p^*)}{c'(\varepsilon)}, \tag{C.16}$$

and if some procedure for estimating λ_p yields an estimate $\hat{\lambda}_p < \lambda^*(p)$, we conclude that the agent is ε-gaming, while $\hat{\lambda}_p > \lambda^*(p)$ rules out gaming.

$$\lambda^*(p) = \frac{R'(\varepsilon + d_p^*)}{c'(\varepsilon)}, \tag{C.17}$$

However, d_p^* is still unknown. Suppose that, we can bound $d_p^* \in [\underline{d}_p, \overline{d}_p]$ with probability $1 - \delta$: (*e.g.*, via a parametric binomial proportion confidence interval about some estimate of d_p^*):

$$\lambda^*(p) \in \left[\frac{R'(\varepsilon + \overline{d_p^*})}{c'(\varepsilon)}, \frac{R'(\varepsilon + \underline{d_p^*})}{c'(\varepsilon)} \right], \tag{C.18}$$

and abbreviate these bounds to $\lambda^*(p) \in [\underline{\lambda}^*(p;\varepsilon), \overline{\lambda}^*(p;\varepsilon)]$. We can use the same bounds on d_p^* to

produce a range of estimates for λ_p:

$$\hat{\lambda}_p \in \left[\frac{R'(\Delta_p(d_p^*))}{c'(\Delta_p(d_p^*) - \overline{d}_p^*)}, \frac{R'(\Delta_p(d_p^*))}{c'(\Delta_p(d_p^*) - \underline{d}_p^*)} \right], \tag{C.19}$$

and abbreviate these bounds to $\hat{\lambda}_p \in [\underline{\hat{\lambda}}_p, \overline{\hat{\lambda}}_p]$. Then, we can compare the intervals for $\lambda^*(p;\varepsilon)$ and $\hat{\lambda}_p$ as follows:

- If $\overline{\hat{\lambda}}_p < \underline{\lambda}^*(p;\varepsilon)$, then agent p is ε-gaming with probability $1-\delta$.
- If $\underline{\hat{\lambda}}_p > \overline{\lambda}^*(p;\varepsilon)$, then we can rule out that agent p is ε-gaming with probability $1-\delta$.
- Otherwise, $\hat{\lambda}_p \in [\underline{\hat{\lambda}}_p, \overline{\hat{\lambda}}_p] \cap [\underline{\lambda}^*(p;\varepsilon), \overline{\lambda}^*(p;\varepsilon)]$ is non-empty, and we cannot definitively rule out or prove the existence of ε-gaming.

However, this algorithm may be of purely technical interest: it is doubtful that the requisite assumptions are satisfied in our motivating setting (upcoding detection), especially Assumption 9, and it is similarly unclear whether such assumptions apply in other settings. In addition, if agents game similarly (*i.e.*, have similar values of λ_p), it may be difficult to rule out/prove ε-gaming for the vast majority of agents.

C.1.2.3 Corollary 13

Corollary. *Define $\tau(p,p')$ as above. Then, given Assumptions 1- 8, $\tau(p,p') > 0$ if and only if $\lambda_p < \lambda_{p'}$.*

Proof. It is sufficient to show that $\tau(p,p') > 0$ if and only if $\Delta_p(d_p^*) < \Delta_{p'}(d_p^*)$, from which Theorem 12 yields the desired result. We can do so by showing (without loss of generality) that $\mathbb{E}[\mathbb{E}[d_i \mid p, x_i]]$ is an unbiased estimate of $\Delta_p(d_p^*)$ (and the case for $\Delta_{p'}(d_p^*)$ proceeds symmetrically). Since $\Delta_p(d_p^*)$ is equivalent to $\mathbb{P}[d_i = 1 \mid p]$ (Eq. 5.2), the result is immediate:

$$\mathbb{E}[\mathbb{E}[d_i \mid p, x_i]] = \mathbb{E}[d_i \mid p] = \mathbb{P}[d_i = 1 \mid p] = \Delta_p(d_p^*). \tag{C.20}$$

□

C.1.2.4 Proposition 14

Proposition. *Let $\tau(\cdot)$ be the oracle treatment effect function, and $\hat{\tau}$ be some sample estimate of τ. Given Assumptions 1- 8, for any $\varepsilon > 0$, if $\sup |\hat{\tau}(p,p') - \tau(p,p')| \leq \varepsilon$, then, for all p, p' such that $\inf_{p,p'} |\tau(p,p')| > \varepsilon$, $\hat{\tau}(p,p) > 0$ if and only if $\lambda_p < \lambda_{p'}$.*

Proof. Choose $\tau, \hat{\tau}$, and some arbitrary ε as specified in the theorem statement. It suffices to show that for all p, p' such that $\inf_{p,p'} |\tau(p, p')| > \varepsilon$, it holds that $\hat{\tau}(p, p') > 0$ if and only if $\tau(p, p') > 0$.

($\implies$) By contradiction; suppose that $\hat{\tau}(p, p') > 0$ but $\tau(p, p') < 0$. Then, by assumption, $\tau(p, p') < -\varepsilon$. But $\sup\ |\hat{\tau}(p, p') - \tau(p, p')| \leq \varepsilon$, so $\hat{\tau}(p, p') < 0$, yielding a contradiction. Thus, $\hat{\tau}(p, p') > 0 \implies \tau(p, p') > 0$. The reverse direction ($\impliedby$) proceeds identically. Thus, the proposition is true. □

Connections to robustness to non-rational actors. We assume throughout that agents behave rationally; *i.e.*, always perfectly maximize the utility function given by Eq. 5.2. However, the rational actor assumption may not always hold; *e.g.*, if agents do not have the resources to carry out the resource maximizing action, or incorrectly estimate their costs. The former is particularly salient for our motivating problem of Medicare upcoding, in which representatives of certain plans may shedule a home visit with a healthcare professional to generate diagnosis codes [292]—a potentially resource-intensive process. Our ranking formulation can afford some robustness to violations of the rational actor assumption:

Remark 33 (Robustness to bounded rationality violations)**.** *Let $\Delta_p(d_p^*)$ be the utility-maximizing action, and let $\tilde{\Delta}_p(d_p^*)$ be an agent's observed action, such that $|\Delta_p(d_p^*) - \tilde{\Delta}_p(d_p^*)| < \varepsilon_1/2$ for some $\varepsilon_1 > 0$. Let $\tilde{\tau}(p, p)$ be some sample estimate of τ, fitted via $\tilde{\Delta}(\cdot)$. Then:*

$$|\hat{\tau}(p, p') - \tilde{\tau}(p, p')| = |(\Delta_p(d_p^*) - \tilde{\Delta}_p(d_p^*)) - (\Delta_{p'}(d_p^*) - \tilde{\Delta}_{p'}(d_p^*))| \leq \varepsilon_1. \tag{C.21}$$

Then, if $\sup\ |\hat{\tau}(p, p') - \tau(p, p')| \leq \varepsilon_2$ such that $\varepsilon \triangleq \varepsilon_1 + \varepsilon_2$, we can apply Proposition 14 directly to conclude that $\hat{\tau}(p, p') > 0$ if and only if $\lambda_p < \lambda_{p'}$.

Intuitively, violations of the rational actor assumption are another source of noise in the estimation of τ. If the noise due to rationality violations plus noise due to standard estimation error are low, then no rankings are flipped, as desired.

C.1.2.5 Identifiability result

For completeness, we show the derivation of the standard causal effect identifiability result for our problem setting (*i.e.*, as in [124]). We note that this is a direct application of a known result in the literature.

Proposition. *Given Assumptions 6- 8, $\mathbb{E}[d_i(p) \mid x_i] = \mathbb{E}[d_i \mid x_i, p]$ and $\mathbb{E}[d_i(p') \mid x_i] = \mathbb{E}[d_i \mid x_i, p']$.*

Proof. We show that $\mathbb{E}[d_i(p) \mid x_i] = \mathbb{E}[d_i \mid x_i, p]$; the case for $\mathbb{E}[d_i(p') \mid x_i]$ proceeds identically. We can write

$$\mathbb{E}[d_i(p) \mid x_i] = \mathbb{E}[d_i(p) \mid x_i, p] = \mathbb{E}[d_i \mid x_i, p] \tag{C.22}$$

where the first equality is an application of conditional exchangeability (Assumption 6), and the second is an application of consistency (Assumption 7). Positivity (Assumption 8) ensures that the conditional expectations are well-defined. □

C.1.2.6 What is the expected AUSC?

For completeness, we also analyze the AUSC metric and provide a closed-form expression for its expected value under a random ranking. First, consider a population of $K \in \mathbb{N}$ agents, where we are interested in top-k sensitivity for some $k \in \{1, \ldots, K\}$. Suppose that we conduct m random audits, for some $m \in \{1, \ldots, K\}$.

The number of ground truth top-k agents that are in the set of m audited agents can be modeled as a hypergeometric random variable $n_m \sim Hyp(k, K, m)$, which is a variable describing the number of "successes" observed by a random draw from a population of K objects with k "success states" in m draws *without replacement*. By standard properties of the hypergeometric distribution, we know that $\mathbb{E}[n_k] = km/K$. Thus, by linearity of expectation, the average (across audit intensities m) number of agents identified by random auditing is given by

$$\frac{1}{K}\sum_{m=1}^{K}\mathbb{E}[n_m] = \frac{k}{K^2}\cdot\frac{K(K-1)}{2} = \frac{k(K-1)}{2K}. \tag{C.23}$$

We divide the result by k to obtain a *proportion* of the top-k identified (as used in the definition of top-k sensitivity), which yields

$$\frac{1}{2}\cdot\frac{K-1}{K}. \tag{C.24}$$

For a finite number of agents, this is bounded above by 0.5, but approaches 0.5 as $K \to \infty$ (a population of infinite agents).

C.1.3 Data Processing

C.1.3.1 Fully synthetic data

Overview. Our general fully synthetic data-generation pipeline is as follows. First, for each agent, we draw some value of $\boldsymbol{\mu}_p \in \mathbb{R}^2$. We draw individual observations $\mathbf{x}^{(i)} \in \mathbb{R}^2$ for each agent from a Gaussian with mean $\boldsymbol{\mu}_p$ and some fixed variance σ^2. Given the $\mathbf{x}^{(i)}$, we simulate a "ground-truth"

decision $d^{*(i)}$. Note that $d^{*(i)}$ represents the ground truth decision rate. Then, for each agent p, we simulate a "gamed" (*i.e.*, strategically perturbed) version of $d^{(i)}$ via a version of the agent utility function in Eq. 5.2. We now formally describe the data-generation process.

Generating a "population" for each agent. To generate a "population" upon which each agent makes decisions, we randomly draw a mean vector specifying a Gaussian distribution for each agent. Given P agents, the mean depends on λ_p as follows:

$$\boldsymbol{\mu}_p = R_\mu \cdot \left(\frac{\tilde{\lambda}_p - \min_{p'} \tilde{\lambda}_p}{\max_{p'} \tilde{\lambda}_p - \min_{p'} \tilde{\lambda}_p} \right) + b; \quad \tilde{\lambda}_p = \log(\lambda_p),\ \bar{\lambda} = \frac{1}{P} \sum_{i=1}^{P} \tilde{\lambda}_p, \tag{C.25}$$

where $a > 0$ is a parameter controlling the range of agent-specific means, and $b \in \mathbb{R}$ is some constnat offset. In other words, we log-transform the λ_p values, then apply min-max scaling and a constant shift. This design allows us to control the level of confounding in the synthetic data by changing R_μ. We use $\boldsymbol{\mu}_p$ to generate populations for each agent as described below.

Concretely, we consider $\lambda_{(\cdot)} \in [0.001, 0.003, 0.005, 0.007, 0.009, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04, 0.045, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3]$ (20 agents), and set $b = -1$. In practice, this means that $\boldsymbol{\mu}_p \in [-1, R_\mu - 1]$. We manually chose $\lambda_{(\cdot)}$ values to set the difficulty of ranking agents by such that a payout-only approach would succeed under low confounding, but fail under high levels of confounding.

Generating ground-truth and gamed decisions. The covariates $\mathbf{x}^{(i)}$ and ground-truth decisions $d^{*(i)}$ are drawn as per

$$\begin{aligned} \mathbf{x}^{(i)} &\sim \mathcal{N}(\boldsymbol{\mu}_{p^{(i)}}, \sigma^2 \mathbf{I}_{2\times 2}) \\ d^{*(i)} &\sim Ber(\alpha^{*(i)}); \quad \alpha^{*(i)} = \sigma(\mathbf{w}^\top \mathbf{x}^{(i)} + b_d), \end{aligned}$$

where $\mu_{(\cdot)}$ is generated for each agent $p^{(i)}$ as described previously, $\mathbf{w}$ is a randomly generated positive vector, and $b_d \in \mathbb{R}$ is some offset.

The $p^{(i)}$s are assigned deterministically; *i.e.*, we sequentially draw a fixed number of $\mathbf{x}^{(i)}$ from each agent-specific distribution and concatenate the results. For realism, inspired by the health insurance setting, and the fact that diagnosis rates for most conditions are relatively low, we set b_d to $\text{logit}(0.05) - \hat{\mathbb{E}}[\mathbf{w}^\top \mathbf{x}]$ such that $\alpha^{*(i)}$ is relatively low, where $\hat{\mathbb{E}}[\cdot]$ denotes the sample mean.

Lastly, we simulate gamed agent decisions $d^{(i)}$ by solving a per-agent utility maximization

problem:

$$d^{(i)} \sim Ber(\alpha_p^{(i)}); \quad \alpha_p^{(i)} = \underset{\tilde{d}_p \in [0,1]}{\arg\max}\ \log(\tilde{d}_p) - \lambda_p(\tilde{d}_p - d^{*(i)})^2.$$

Implementation details. All randomness is seeded *once* at the start of the entire data-generation process.

C.1.3.2 Medicare cohort

We select a 0.2% pseudo-random sample of all U.S. Medicare beneficiaries using the last characters of encrypted beneficiary IDs for inclusion in the cohort. As covariates, we choose age, racial category, biological sex, and diagnosis code categories (defined using the 2018 version of the "Hierarchical Condition Category" [HCC] schema released by the Center for Medicare Services). This yields a total of 97 features.

We exclude enrollees who did not generate any claims (no healthcare utilization; *i.e.*, zero cost as recorded by Medicare), those not located in the 50 U.S. states and the District of Columbia, as well as *dual-eligible* beneficiaries. Dual-eligibility refers to individuals simultaneously eligible for U.S. Medicare and U.S. Medicaid. While eligibility for Medicare is primarily based on age, eligibility for Medicaid is primarily based on disability status. Dual-eligible beneficiaries are excluded since the U.S. government uses a different payout model for dual-eligible vs. non-dual-eligible enrollees, potentially violating Assumption 1 (shared rewards), since agents may not be reacting to the same payout model for all enrollees.

Licensing. Our cohort is drawn from a 20% sample of all U.S. Medicare beneficiaries provided to the authors under a data usage agreement with the Center for Medicare & Medicaid Services.

State-level healthcare statistics. We use a mix of raw and engineered features from the CMS (Center for Medicare & Medicaid Services) Provider of Service file (license: Public Use File) [2][1] and the National Neighborhood Data Archive (NANDA; CC-BY 4.0) [144]. In NANDA, data is already aggregated at the state level (including the District of Columbia). We keep statistics pertaining to per-capita or per-sq. mi. healthcare provider density (50 features), filtering to providers with non-zero sales. From the CMS Provider of Service file, data is reported at the provider level. First, we filter out providers ineligible for Medicare participation, and providers that are no longer active. We then manually code ownership information following the provided data dictionaries (for-profit vs. non-profit vs. publicly owned). Next, we compute at the state level the prevalence

[1]Full information on Public Use File licensing as defined by CMS is here: https://www.cms.gov/research-statistics-data-and-systems/files-for-order/nonidentifiabledatafiles

of for-profit, non-profit, and publicly-owned providers of each type (as defined by CMS) by state, as well as the ratio of for-profit to non-profit providers. Additionally, we extract the average per-hospital bed count, physician count, medical school affiliation rate, and compliance rate as certified by CMS, for a total of 54 features. All aggregations for the Provider of Service file are unweighted (*i.e.*, all hospitals/other healthcare facilities contribute equally). This yields a total of 104 state-level health indicators.

We report the statistics derived from the provider of service file (Table C.1) and NANDA (Table C.2), with summary statistics across states. Summary statistics are computed excluding infinite and NaN values (*i.e.*, ratios of 0/0 or 1/0). When computing correlations, we exclude states with NaN values for that feature (*i.e.*, ratio of 0/0). Proportions of publicly-owned, non-profit, and for-profit providers do not sum to 1, because providers reporting "unknown" or "other" ownership are excluded. For further information on the feature definitions, consult the data dictionaries for the provider of service[2] and NANDA[3] files.

Table C.1: Features chosen for analysis derived from the provider of service file, with mean and standard deviation across states. S.d.: standard deviation. OP PT/SP: outpatient physical therapy and speech pathology. ICF/ID: intermediate care facility for individuals with intellectual disability. *: invalid (1/0 or 0/0) ratios dropped.

Feature name	**Mean±s.d.**
bed count	95.125±27.826
% affiliated with a medical school	3.385±0.350
# of personnel	87.182±47.477
% non-compliant	0.115±0.074
% of providers that are non-profit hospitals	0.055±0.030
% of providers that are for-profit hospitals	0.024±0.015
% of providers that are publicly-owned hospitals	0.028±0.027
% of providers that are non-profit skilled nursing facilities	0.069±0.046
% of providers that are for-profit skilled nursing facilities	0.171±0.068
% of providers that are publicly-owned skilled nursing facilities	0.020±0.021
% of providers that are non-profit home health agencies	0.033±0.018
% of providers that are for-profit home health agencies	0.104±0.075
% of providers that are publicly-owned home health agencies	0.008±0.010
% of providers that are non-profit OP PT/SP providers	0.005±0.010

[2] https://www.nber.org/research/data/provider-services-files
[3] https://www.openicpsr.org/openicpsr/project/120907/version/V3/view

Feature name	Mean±s.d.
% of providers that are for-profit OP PT/SP providers	0.025±0.018
% of providers that are publicly-owned OP PT/SP providers	0.001±0.001
% of providers that are non-profit end-stage renal disease facilities	0.014±0.012
% of providers that are for-profit end-stage renal disease facilities	0.096±0.042
% of providers that are publicly-owned end-stage renal disease facilities	0.001±0.002
% of providers that are non-profit ICF/IDs	0.041±0.052
% of providers that are for-profit ICF/IDs	0.025±0.050
% of providers that are publicly-owned ICF/IDs	0.005±0.007
% of providers that are non-profit rural health clinics	0.044±0.044
% of providers that are for-profit rural health clinics	0.024±0.028
% of providers that are publicly-owned rural health clinics	0.014±0.020
% of providers that are non-profit ambulatory surgery centers	0.004±0.005
% of providers that are for-profit ambulatory surgery centers	0.084±0.061
% of providers that are publicly-owned ambulatory surgery centers	0.002±0.004
% of providers that are non-profit hospice care facilities	0.020±0.012
% of providers that are for-profit hospice care facilities	0.039±0.033
% of providers that are publicly-owned hospice care facilities	0.002±0.004
% of providers that are non-profit	0.285±0.139
% of providers that are for-profit	0.592±0.154
% of providers that are publicly-owned	0.081±0.066
% of urbanized population	0.679±0.228
% of providers that are ambulatory surgery centers	0.091±0.064
% of providers that are end-stage renal disease facilities	0.111±0.040
% of providers that are home health agencies	0.145±0.069
% of providers that are hospitals	0.118±0.038
% of providers that are hospice care facilities	0.070±0.035
% of providers that are ICF/IDs	0.073±0.088
% of providers that are OP PT/SP providers	0.007±0.009
% of providers that are rural health clinics	0.082±0.067
% of providers that are skilled nursing facilities	0.261±0.073
ratio of for-profit to non-profit providers	2.838±2.024
ratio of for-profit to non-profit hospitals	0.730±0.955
ratio of for-profit to non-profit skilled nursing facilities	3.633±2.387
ratio of for-profit to non-profit home health agencies	5.135±6.020

Feature name	Mean±s.d.
ratio of for-profit to non-profit OP PT/SP providers	10.149*±10.108
ratio of for-profit to non-profit end-stage renal disease facilities	15.905*±18.546
ratio of for-profit to non-profit ICF/IDs	1.074*±1.969
ratio of for-profit to non-profit rural health clinics	0.936±1.380
ratio of for-profit to non-profit ambulatory surgery centers	32.826*±30.949
ratio of for-profit to non-profit hospice care facilities	3.989±6.797

Table C.2: Features chosen for analysis derived from NANDA, with mean and standard deviation across states. All chosen summary statistics are for providers with > 0 U.S. dollars in sales. S.d.: standard deviation. Sq. mi.: square mile. P/O/ST: physical, occupational, and speech therapists. IDD: intellectual and developmental disabilities. MHSA: mental health and substance abuse.

Feature name	Mean±s.d.
ambulatory health care specialists per 1000 people	4.429±2.745
ambulatory health care specialists per sq. mi.	15.480±16.928
physicians per 1000 people	1.753±1.196
physicians per sq. mi.	6.477±7.593
physicians (excluding mental health) per 1000 people	1.656±1.090
physicians (excluding mental health) per sq. mi.	6.137±7.195
mental health physicians per 1000	0.096±0.112
mental health physicians per sq. mi.	0.340±0.412
dentists per 1000 people	0.639±0.242
dentists per sq. mi.	2.553±2.945
all other health practitioners (besides the above) per 1000 people	1.483±1.039
all other health practitioners (besides the above) per sq. mi.	4.949±5.264
chiropractors per 1000 people	0.282±0.257
chiropractors per sq. mi.	0.840±0.999
optometrists per 1000 people	0.100±0.026
optometrists per sq. mi.	0.307±0.240
non-physician mental health practitioners per 1000	0.176±0.390
non-physician mental health practitioners per square	0.687±0.929
P/O/ST and audiologists per 1000 people	0.133±0.054
P/O/ST and audiologists per sq. mi.	0.474±0.657
other health practitioners (besides the above) per 1000 people	0.792±0.610

Feature name	Mean±s.d.
other health practitioners (besides the above) per sq. mi.	2.640±2.742
outpatient care centers per 1000 people	0.181±0.133
outpatient care centers per sq. mi.	0.556±0.516
diagnostic labs per 1000 people	0.075±0.114
diagnostic labs per sq. mi.	0.229±0.230
home health services per 1000 people	0.173±0.112
home health services per sq. mi.	0.517±0.441
other ambulatory care services (besides the above) per 1000 people	0.125±0.327
other ambulatory care services (besides the above) per sq. mi.	0.198±0.176
all nursing and residential care facilities per 1000 people	0.334±0.324
all nursing and residential care facilities per sq. mi.	0.836±0.650
nursing care facilities per 1000 people	0.224±0.330
nursing care facilities per sq. mi.	0.473±0.308
residential IDD care facilities per 1000 people	0.026±0.014
residential IDD care facilities per sq. mi.	0.102±0.174
inpatient facilities for care of individuals with IDDs per 1000 people	0.012±0.010
inpatient facilities for care of individuals with IDDs per sq. mi.	0.039±0.068
inpatient facilities providing MHSA care per 1000 people	0.014±0.007
inpatient facilities providing MHSA care per sq. mi.	0.063±0.109
continuing care and assisted living facilities per 1000 people	0.023±0.011
continuing care and assisted living per sq. mi.	0.070±0.062
other residential care facilities (besides the above) per 1000 people	0.061±0.031
other residential care facilities (besides the above) per sq. mi.	0.191±0.179
pharmacies and drug stores per 1000 people	0.314±0.669
pharmacies and drug stores per sq. mi.	0.853±1.336
optical goods stores per 1000 people	0.076±0.025
optical goods stores per sq. mi.	0.282±0.360
misc. other health and personal care stores per 1000 people	0.054±0.014
misc. other health and personal care stores per sq. mi.	0.149±0.090

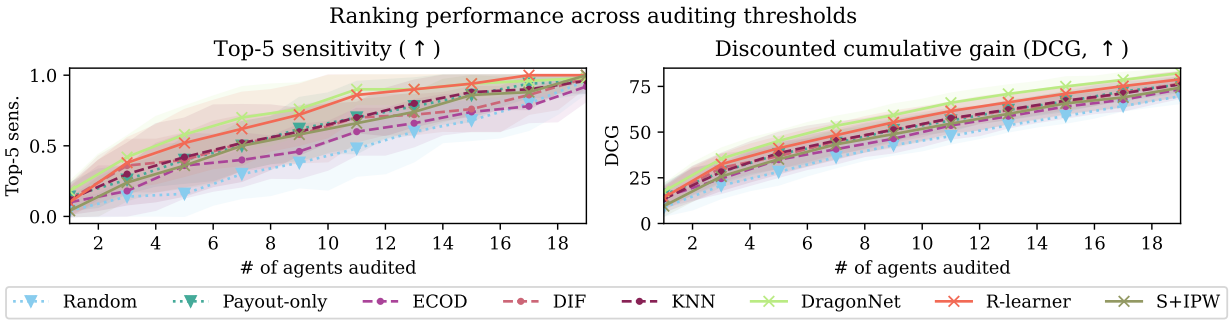


Figure C.2: Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.0, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator.

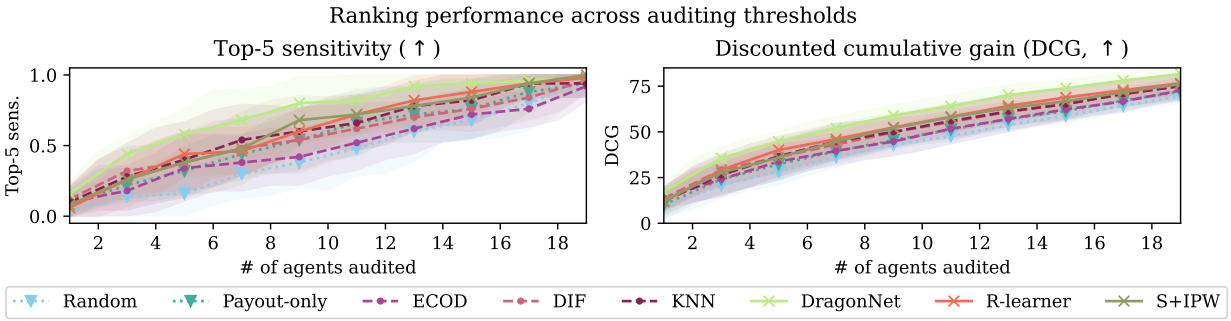


Figure C.3: Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.1, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator.

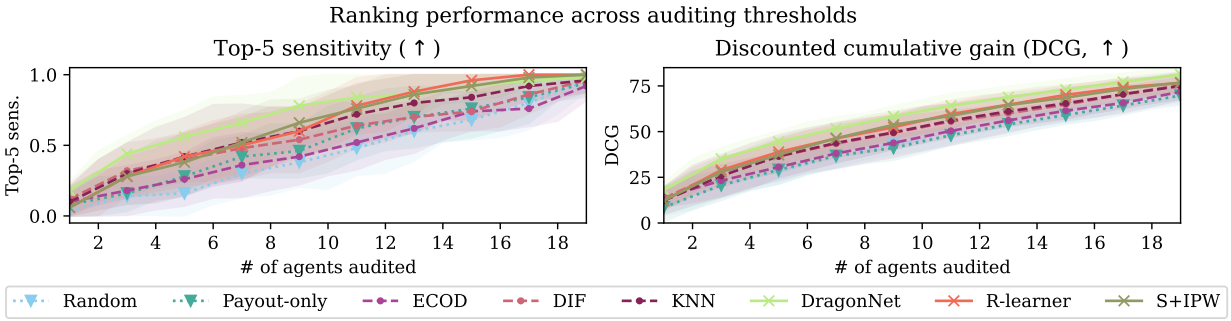


Figure C.4: Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.2, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator.

C.1.4 Supplementary results

C.1.4.1 Causal vs. non-causal approaches, all audit thresholds & all levels of confounding

We show plots with the top-5 sensitivity and DCG at all levels of confounding evaluated, as measured by the mean range R_μ (*i.e.*, $R_\mu \triangleq \max\ \mu_p - \min\ \mu_p$ across agents). We choose $R_\mu \in \{0, 0.1, \ldots, 1.0\}$. For convenience, we provide an index of results:

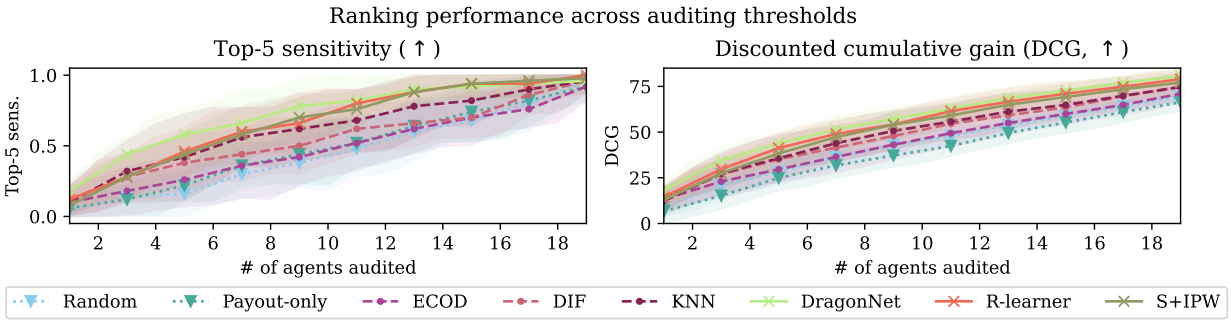


Figure C.5: Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.3, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator.

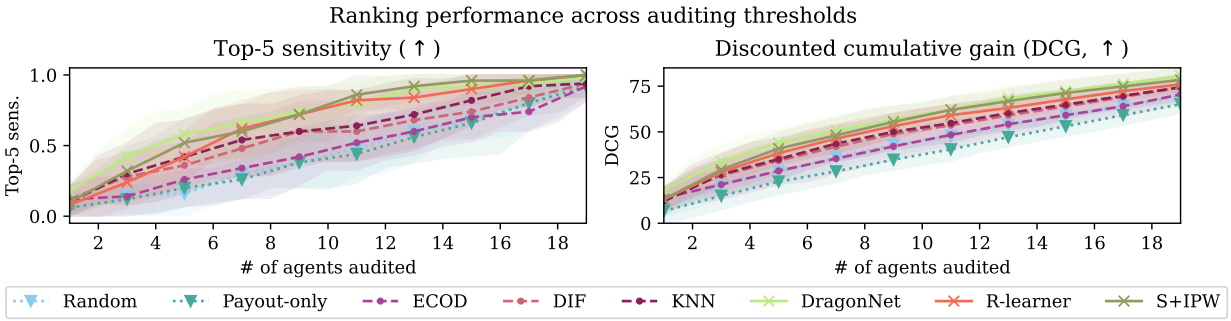


Figure C.6: Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.4, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator.

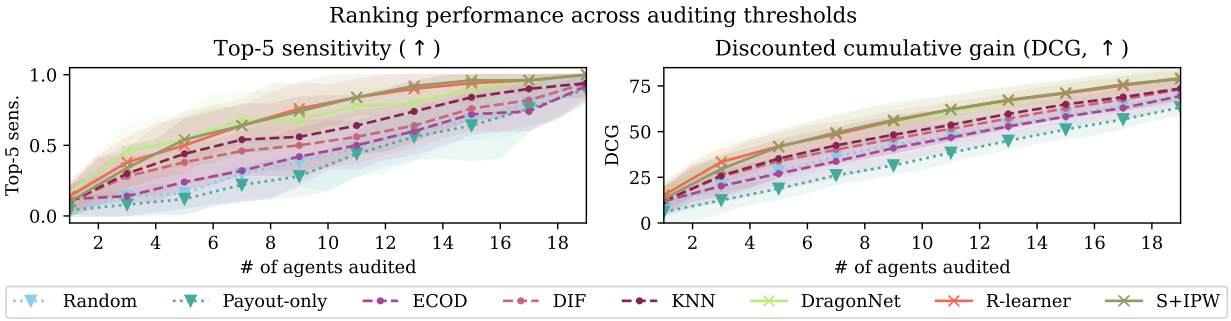


Figure C.7: Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.5, with $\pm\sigma$ error. $\triangledown$: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator.

- $R_\mu = 0.0$: Figure C.2
- $R_\mu = 0.1$: Figure C.3
- $R_\mu = 0.2$: Figure C.4
- $R_\mu = 0.3$: Figure C.5

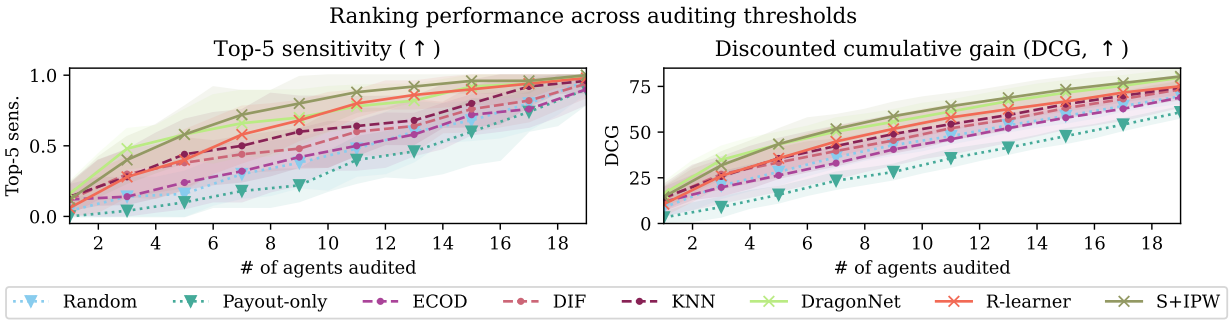


Figure C.8: Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.6, with $\pm\sigma$ error. ∇: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator.

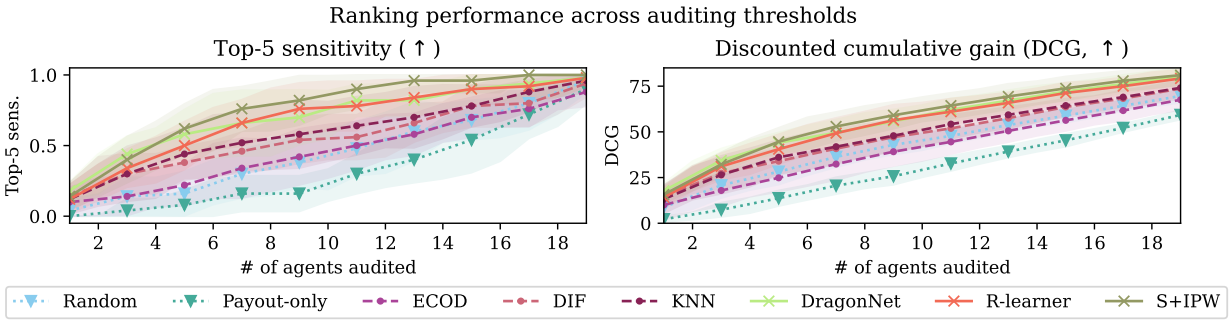


Figure C.9: Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.7, with $\pm\sigma$ error. ∇: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator.

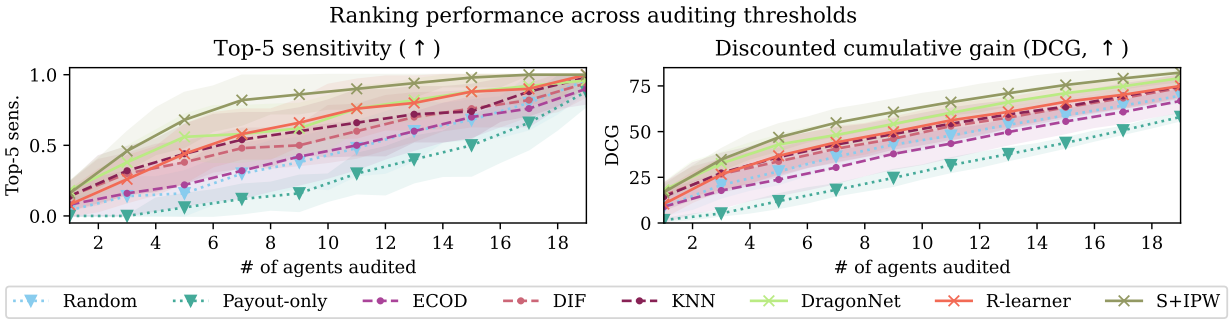


Figure C.10: Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.8, with $\pm\sigma$ error. ∇: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator.

- $R_\mu = 0.4$: Figure C.6
- $R_\mu = 0.5$: Figure C.7
- $R_\mu = 0.6$: Figure C.8
- $R_\mu = 0.7$: Figure C.9

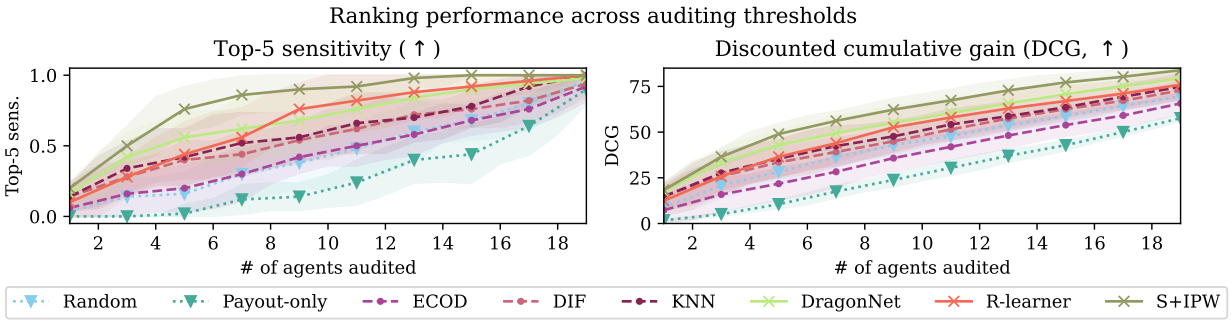


Figure C.11: Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 0.9, with $\pm\sigma$ error. ∇: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator.

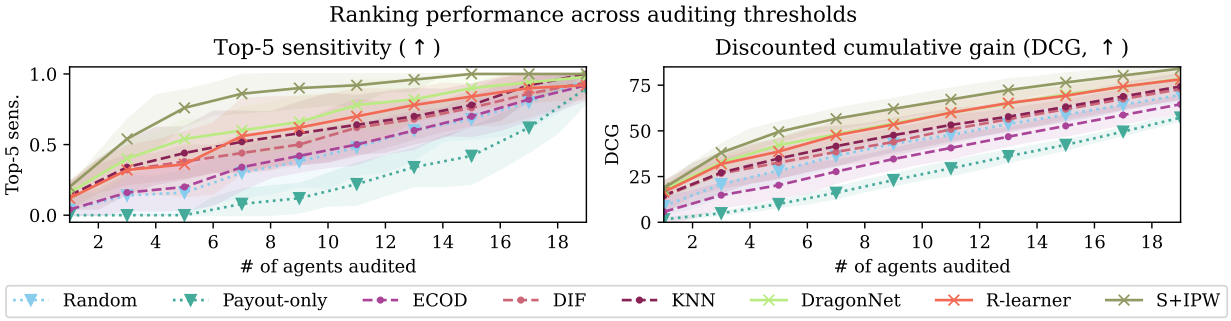


Figure C.12: Mean top-5 sensitivity (left) and DCG (right) across # of agents audited at mean range 1.0, with $\pm\sigma$ error. ∇: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator.

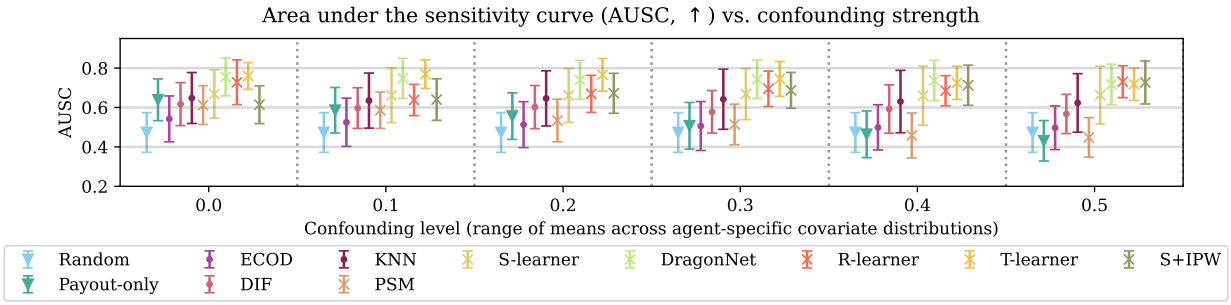


Figure C.13: Area under the sensitivity curve (AUSC) for all methods tested across levels of confounding (mean range $R_\mu \leq 0.5$). ∇: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator.

- $R_\mu = 0.8$: Figure C.10
- $R_\mu = 0.9$: Figure C.11
- $R_\mu = 1.0$: Figure C.12

We summarize the main trends for non-causal approaches and defer discussion of causal methods to the sensitivity analysis of all causal effect estimators. For convenience, we also plot the AUSC

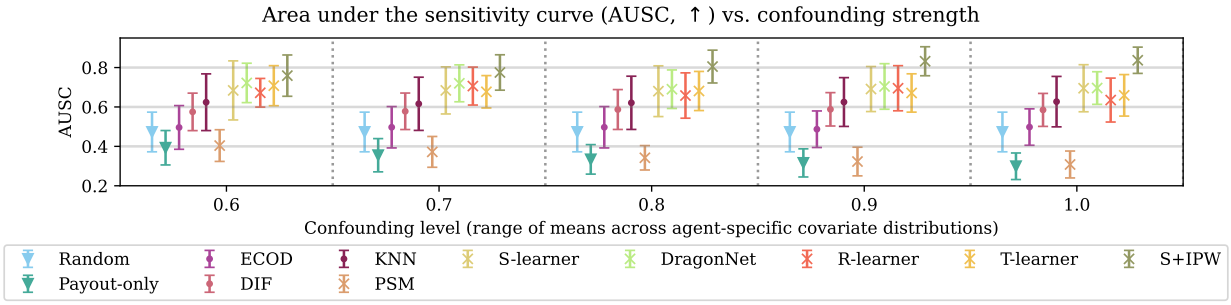


Figure C.14: Area under the sensitivity curve (AUSC) for all methods tested across levels of confounding (mean range $R_\mu > 0.5$). ∇: naïve baseline. $\circ$: anomaly detection method. $\times$: causal effect estimator.

for *all* approaches across levels of confounding in Figures C.13 ($R_\mu \leq 0.5$) and C.14 ($R_\mu > 0.5$). **For readability, in contrast to our other figures, causal effect estimators are marked with "$\circ$" instead of "$\times$".**

The payout-only approach can degrade to worse-than-random ranking due to confounding. The random auditing method performs similarly across all levels of confounding, as expected. However, as confounding increases, the payout-only ranking degrades toward random, then worse than random. The latter can occur if confounding is so strong that the relationship between gaming and observed diagnosis rates flips; *e.g.*, if (in the health insurance setting) very dishonest plans tend to serve relatively healthy populations compared to more gaming-averse plans.

In the synthetic dataset, anomaly detection methods use the variance of the observed decision ($\mathbb{V}[d_i]$) as a gaming signature. Most anomaly detection methods perform near-random, or slightly better than random. While this is due to the properties of the synthetic dataset, the results highlight potentially interesting characteristics of anomaly detection methods for gaming detection. Recall that the anomaly detection methods take covariates and agent decisions (x_i, d_i) as input. $\mathbb{V}[x_i]$ is identical across agents by design, and all agents see the same number of observations. However, $\mathbb{V}[d_i]$ may vary across agents. Thus, KNN uses $\mathbb{V}[d_i]$ as a signal of gaming, which has utility under low confounding, but is less useful as confounding increases.

To see this, recall that d_i is a binary decision, and let $p_d = P(d_i = 1)$ for some agent. $\mathbb{V}[d_i]$ is proportional to $p_d \cdot (1 - p_d)$, and is concave in p_d with maximizer $p_d = 0.5$. Since $P(d_i)$ is generally low in simulation (*i.e.*, $\ll 0.5$), agents with higher observed $P(d_i)$ rates will generally have higher $\mathbb{V}[d_i]$ as well, yielding a higher anomaly score. Under no confounding, these are precisely the agents that are gaming more. Indeed, KNN performs slightly better than random, but its advantage over random performance diminishes slightly with confounding as the utility of $\mathbb{V}[d_i]$ as a signature

for gaming. However, even at low confounding, KNN and related anomaly detection methods are inherently unable to detect gaming in non-outlier points, which occur in denser regions of covariate space. In these regions, causal methods enjoy an advantage over anomaly detection approaches due to improved overlap (Assumption 8). Thus, KNN does not exceed the ranking performance of the best causal methods. Note that this argument assumes that $\mathbb{V}[x_i]$ is similar across agents, which holds in the synthetic dataset.

More advanced anomaly detection methods can also achieve slightly better than random performance (*e.g.*, DIF), but since these approaches transform the covariate space in highly non-linear ways (*e.g.*, via random projections as in DIF, or via feature-wise CDFs as in ECOD), they may destroy outlier information useful for gaming detection. Ultimately, anomaly detection methods have inherently limited utility for gaming detection, since some gamed decisions may appear distributionally close to other gamed decisions.

C.1.4.2 Sensitivity analysis of causal effect estimators

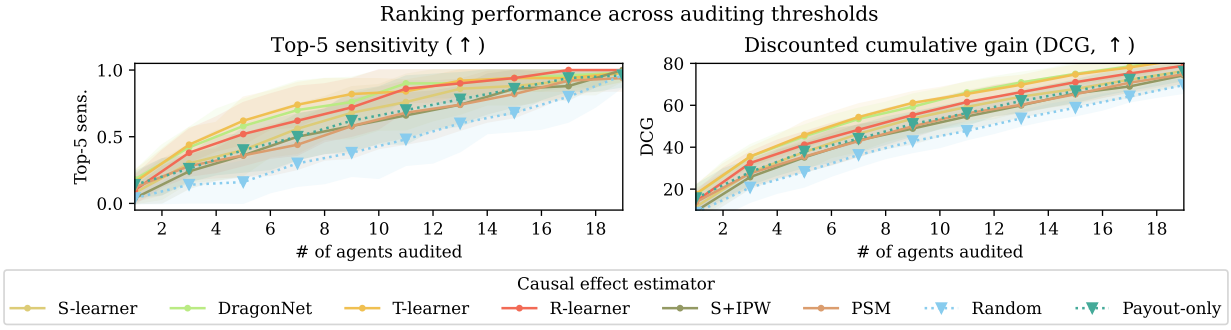


Figure C.15: Sensitivity analysis of all causal methods tested, mean range 0.0. ▽: naïve baseline. ○: causal effect estimator.

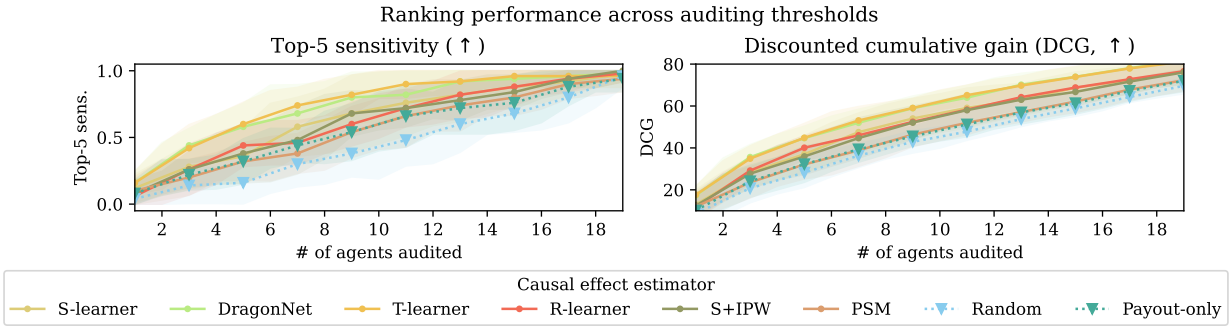


Figure C.16: Sensitivity analysis of all causal methods tested, mean range 0.1. ▽: naïve baseline. ○: causal effect estimator.

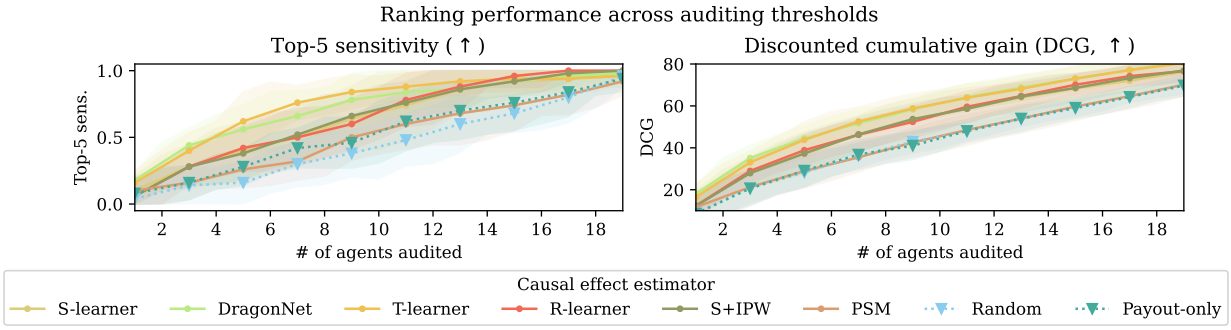


Figure C.17: Sensitivity analysis of all causal methods tested, mean range 0.2. ▽: naïve baseline. ○: causal effect estimator.

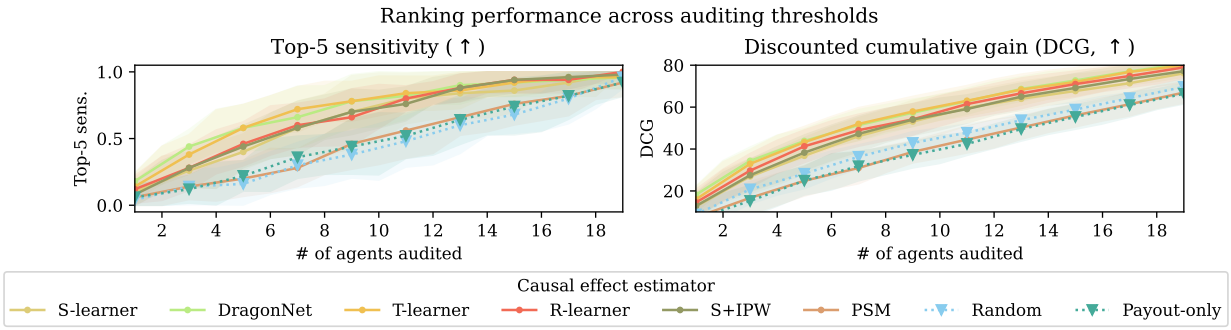


Figure C.18: Sensitivity analysis of all causal methods tested, mean range 0.3. ▽: naïve baseline. ○: causal effect estimator.

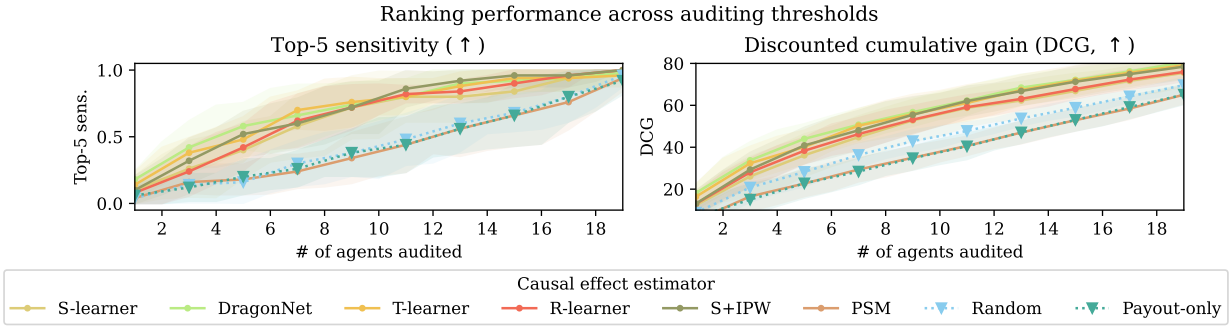


Figure C.19: Sensitivity analysis of all causal methods tested, mean range 0.4. ▽: naïve baseline. ○: causal effect estimator.

We show plots with the top-5 sensitivity and DCG at all levels of confounding evaluated (as measured by the mean range $R_\mu \in \{0, 0.1, \ldots, 1.0\}$) for causal effect estimators only, plus random and payout-only methods for comparison. An index of figures follows:

- Figure C.15: $R_\mu = 0.0$

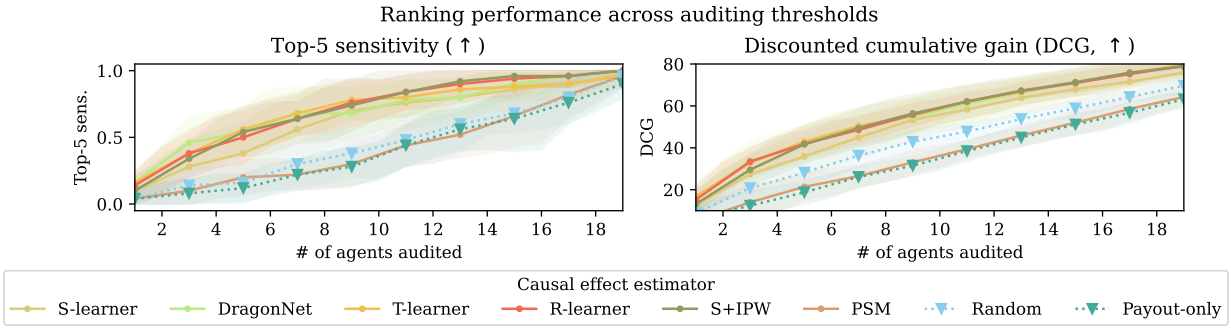


Figure C.20: Sensitivity analysis of all causal methods tested, mean range 0.5. ▽: naïve baseline. ○: causal effect estimator.

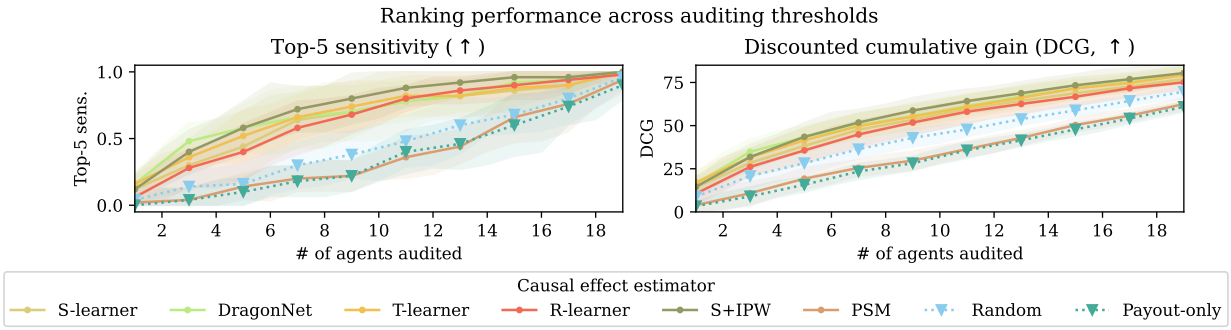


Figure C.21: Sensitivity analysis of all causal methods tested, mean range 0.6. ▽: naïve baseline. ○: causal effect estimator.

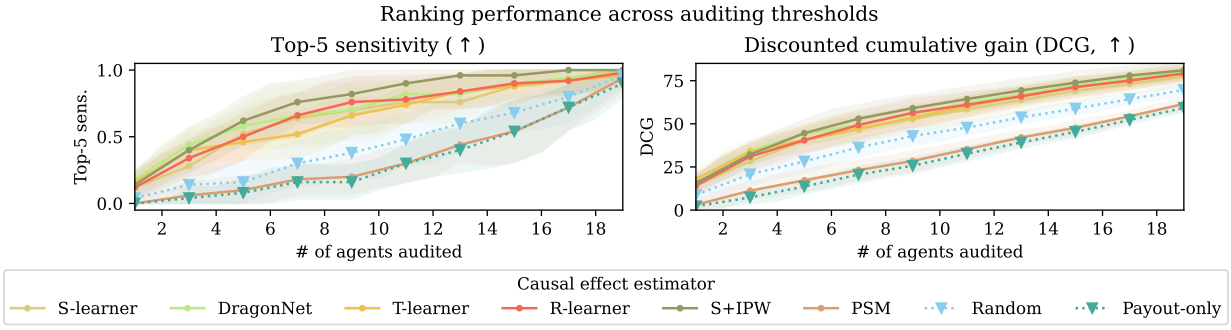


Figure C.22: Sensitivity analysis of all causal methods tested, mean range 0.7. ▽: naïve baseline. ○: causal effect estimator.

- Figure C.16: $R_\mu = 0.1$
- Figure C.17: $R_\mu = 0.2$
- Figure C.18: $R_\mu = 0.3$

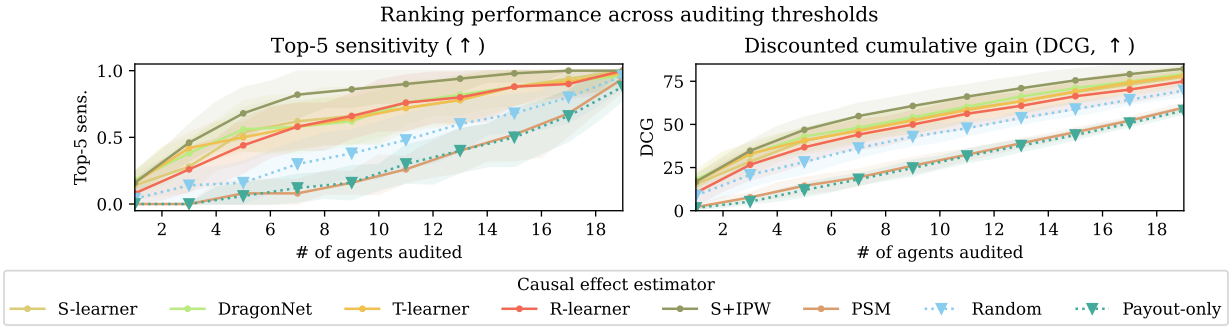


Figure C.23: Sensitivity analysis of all causal methods tested, mean range 0.8. ▽: naïve baseline. ○: causal effect estimator.

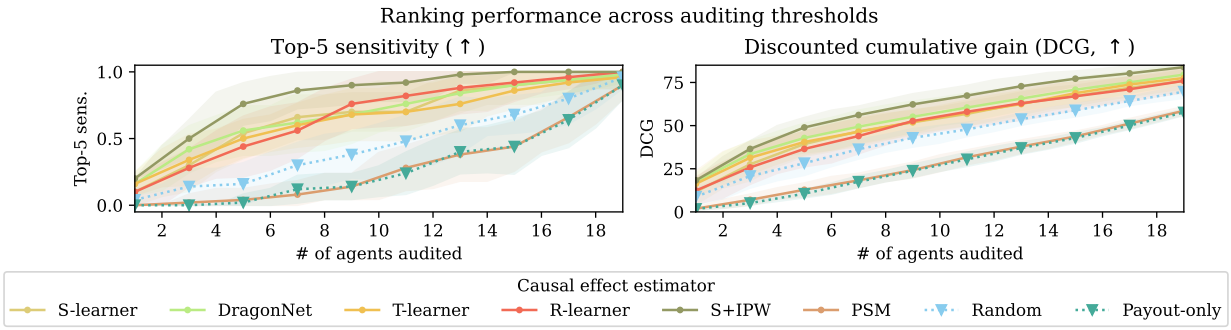


Figure C.24: Sensitivity analysis of all causal methods tested, mean range 0.9. ▽: naïve baseline. ○: causal effect estimator.

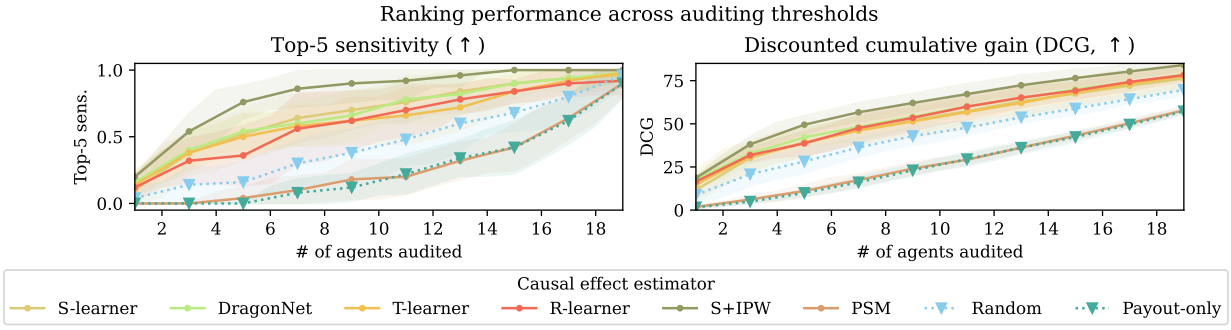


Figure C.25: Sensitivity analysis of all causal methods tested, mean range 1.0. ▽: naïve baseline. ○: causal effect estimator.

- Figure C.19: $R_\mu = 0.4$
- Figure C.20: $R_\mu = 0.5$
- Figure C.21: $R_\mu = 0.6$

- Figure C.22: $R_\mu = 0.7$
- Figure C.23: $R_\mu = 0.8$
- Figure C.24: $R_\mu = 0.9$
- Figure C.25: $R_\mu = 1.0$

Note that, even absent confounding, causal approaches outperform the payout-only model. This is because causal approaches explicitly incorporate the covariates into modeling, while the payout-only model directly uses the marginal outcome distribution. Incorporating covariates into regression models for treatment effect estimation can decrease estimator variance [124] (but is not guaranteed to do so), consistent with the empirical results.

Propensity score matching (PSM) also performs poorly across all levels of confounding. Since the matching approaches conduct matching pairwise across observations seen by agent pairs, the causal effect estimates are computed in a subset of similar observations across one pair of agents, but not the subset of similar observations across *all* distributions agent observations. This suggests that controlling for confounding simultaneously across all levels of treatment is potentially important for applying causal effect estimators to gaming detection. Ultimately, matching approaches may not scale to large numbers of treatments, such as those expected in multi-agent strategic adaptation. Non-optimal matching approaches (*e.g.*, greedy matching without replacement) are a potential workaround, but we leave the adaptation of matching methods to large numbers of treatments to future work.

We note that, at low levels of confounding, the difference between the S-learner and T-learner may be dataset-dependent: empirically, S-learners often regularize causal effect estimates towards zero, while T-learners thrive when causal effects are non-zero and heterogeneous [158, 51], as in our synthetic dataset. Thus, per-agent modeling, as done by the T-learner and DragonNet, can better capture the complex treatment effects in our dataset.

Furthermore, the R-learner and S+IPW both perform poorly at *low* levels of confounding, but improve at high levels of confounding. Since both the R-learner and S+IPW fit a nuisance propensity score estimator, this suggests that difficulties in propensity score estimation at low levels of confounding could potentially explain the observed trends.

The underperformance of the R-learner may be surprising given its doubly-robust properties, but the high-dimensional generalization [135] requires restrictive conditions for convergence. Formally, the R-learner fits four models m, g, h, e of the form

$$d \sim m(x) + g(x)^\top (h(p) - e(x)), \tag{C.26}$$

where m is fit independently, and g, h, e are fitted using alternating optimization. The final treatment effect estimate of swapping from agent p to p' is given by $g(x)^\top h(p) - g(x)^\top h(p')$, but the oracle representation of $h(\cdot)$ is unknown and must be fitted. Convergence of the nuisance parameter estimate of $e(x)$ to the oracle value of $h(p)$ is necessary for convergence of the overall treatment effect estimate.

C.1.5 Training

C.1.5.1 Model architectures

All approaches that fit a model are built based on a fully-connected neural network with two hidden layers and 300 neurons per layer plus ReLU activations. The output of the neural network either has size two with a softmax non-linearity (for classification; *i.e.*, predicting agent decisions), or no activation and a pre-specified output size (*i.e.*, for generating feature maps in the high-dimensional R-learner).

We describe approach-specific modifications to the architectures as follows:

S+IPW. The estimated weights are used as a sample weight when training the S-learner. At inference time, weights are also computed for test examples based on the propensity model fitted on the training set to take an inverse propensity-weighted average of the S-learner estimates.

DragonNet. We changed the propensity prediction head from a binary classification (as in the original paper [256]) to a multi-class classification head, since there are multiple treatments in our setting. Furthermore, we introduce a new outcome modeling head per agent. Targeted regularization is omitted due to the multi-treatment setup.

R-Learner. Feature maps for all nuisance parameters have dimensionality 10. The generalized R-learner uses alternating optimization to fit some nuisance parameters, where two models representing a product decomposition of the response function are updated K times for every update of a "propensity feature" model. We set $K = 10$; we defer to [135], page 6 for more details about the training procedure for the generalized R-learner, from which we designed our implementation.

C.1.5.2 Anomaly detection hyperparameters

For KNN, we keep the neighborhood size at 5, the default value. DIF uses neural network-based random projections to compute an anomaly score. Thus, we use the same architecture for DIF as

used for causal approaches, with an ensemble of 50 representations. Each representation is used as input into an isolation forest of size 6 [176].[4] ECOD does not take hyperparameters [174].

C.1.5.3 Dataset splits

We use a seeded random development-test split (7:3) for all datasets. The development split is reserved for all model fitting, while all causal effect estimates are reported on the test split. The development split is further randomly split into a training and a validation set. All model selection techniques (*e.g.*, early stopping) are performed with respect to evaluation metrics on the validation set.

C.1.5.4 Training hyperparameters

Fully synthetic data We use the following hyperparameters for training all models:

- Optimizer: SGD with learning rate 10^{-2} and weight decay 10^{-3}.
- Learning rate schedule: We reduce the learning rate by a factor of 0.1 after 5 epochs of non-improvement with respect to the validation loss.
- Training length: A *maximum* of 1000 epochs, with early stopping (patience: 10 epochs) based on validation loss.

Medicare FFS

- Optimizer: SGD with learning rate 10^{-2} and weight decay 10^{-3}.
- Learning rate schedule: We reduce the learning rate by a factor of 0.1 after 5 epochs of non-improvement with respect to the validation loss.
- Training length: A *maximum* of 1000 epochs, with early stopping (patience: 10 epochs) based on validation loss.

[4] An isolation forest is a separate anomaly detection method, in which the anomaly score is related to the number of random "splits" with respect to a single randomly-selected covariate (*i.e.*, a threshold of the form $x_i \geq r$) needed to isolate a point. This method assumes that outliers are "easier" to isolate, and require fewer splits. Multiple random-splitting routines (isolation trees) are ensembled to form an isolation forest.

C.1.6 Software and Hardware

C.1.6.1 Software

All code was written in Python 3.10.4 (license: PSF). All non-causal anomaly detection approaches were implemented using PyOD (license: BSD 2-clause) [308]. All neural networks were implemented in PyTorch 2.2.0 (license: Custom "BSD-style"[5]) [211], using Skorch 0.15.0 (license: BSD 3-clause) [275] as a wrapper. Metrics were computed using both Scikit-Learn 1.3.2 (license: BSD 3-clause) [218] and Scipy 1.11.4 (license: BSD 3-clause) [283]. For the fully synthetic data generation process, CVXPY 1.4.2 (license: Apache 2.0) [65] was used to solve each agent's utility maximization problem, and used in tandem with SCIP 9.0 (`pyscipopt` 5.0.0; license: Apache 2.0) for the matching approaches (formulated as mixed-integer programs) [31]. Numpy 1.22.3 (license: BSD-style) [107][6] and Pandas 2.0.3 (license: BSD 3-clause) [274] were used for data manipulation. Matplotlib 3.8.2 (empirical results; license: PSF-style)[7] and Adobe Illustrator 2023 (overview figures; license: commercial, "Named User Licensing"[8]) were used for figure generation. For the Medicare cohorts, we generated HCC (Hierarchical Condition Categories; used by the Center for Medicare Services) codes from raw diagnosis codes reported in claims data via HCCPy 0.1.9 (license: Apache 2.0)[9].

Other dependencies include `tqdm` 4.66.2 for rendering progress bars (license: MPL 2.0 and MIT), `gitpython` 3.1.43 for bookkeeping (license: BSD 3-clause), `pandarallel` 1.6.5 for parallel data processing (license: BSD 3-clause), and `ruamel` 0.18.6 (license: MIT) for configuration file management. All software excepting Adobe Illustrator is open-source and free for use.

C.1.6.2 Hardware

All experiments were run on either one Titan V or V100 GPU using 12.9GB of RAM as managed via a Slurm job submission system. Computing nodes had two 2.10GHz Intel Broadwell (Xeon E5-2620V4) processors each (16 cores total). Execution time was limited to six hours per run, but all training runs (one model type on 10 datasets) lasted under one hour due to the relatively small size of the architectures and datasets under consideration.

[5] See https://github.com/pytorch/pytorch/blob/main/LICENSE.
[6] See https://github.com/numpy/numpy/blob/main/LICENSE.txt.
[7] See https://matplotlib.org/stable/project/license.html.
[8] See https://helpx.adobe.com/enterprise/using/licensing.html.
[9] https://github.com/yubin-park/hccpy

C.1.7 Code

The fully-synthetic datasets, experimental code, and implementations of all approaches under evaluation can be found at https://github.com/MLD3/gaming_detection. The authors do not have permission to release any of the Medicare data, but the relevant data processing code will be included.

C.2 Supplementary material for "Streamlining Oversight in Medicare Advantage: A New Approach to Measuring Coding Intensity via Contract Rankings "

C.2.1 Methods - Exclusion Criteria/Filtering

Since exact operationalizations of definitions may be subject to discretion, code used to filter our switchers' cohort will be made publicly available.

C.2.2 Methods - Ranking Computation

We provide formal details on the ranking approach here: diagnosis rate estimation, adjustment, and rank aggregation.

To estimate diagnosis rates, we leverage the beta-binomial model to shrink estimates towards a per-market, per-stratum mean. Doing so shrinks extreme observed diagnosis rates stemming from low enrollment in specific strata. Without loss of generality, consider a single arbitrary diagnosis. We assume that diagnosis counts and rates $y_{c,m,s}$ and $p_{c,m,s}$, respectively, for a contract c in market m and stratum s of size $n_{c,m,s}$, are generated as follows:

$$y_{c,m,s} \sim \text{Binom}\left(n_{c,m,s},\ p_{c,m,s}\right), \quad p_{c,m,s} \sim \text{Beta}\left(\alpha_{m,s},\ \beta_{m,s}\right). \tag{C.27}$$

We estimate the per-stratum, per-market Beta parameters $\alpha_{m,s}$ and $\beta_{m,s}$ using an empirical Bayes approach. Formally, we optimize

$$\arg\max_{\alpha_s,\beta_s} \sum_{j=1}^{J} \log P\left(y_{s,j},\ n_{s,j} | \alpha_s, \beta_s\right)$$

$$\text{s.t.} \quad \frac{\alpha_s}{\alpha_s + \beta_s} = \frac{\sum_{j=1}^{J} y_{s,j}}{\sum_{j=1}^{J} n_{s,j}} \quad \text{and} \quad 2 \leq \alpha_s + \beta_s \leq \frac{\alpha_s \beta_s}{\varepsilon\left(\alpha_s + \beta_s\right)^2} - 1; \forall s.$$

that is, we optimize the marginal log-likelihood of our data subject to the constraint that our prior mean matches the pooled per-market, per-stratum diagnosis rate, the constraint that $\alpha_s + \beta_s$ be at least 2 to avoid an U-shaped prior, and the constraint that $\alpha_s + \beta_s$ is bounded above as a variance lower bound to ensure identifiability if only one contract is present in a stratum. In practice, it is easier to leverage the reparameterization $\mu_s = \frac{\alpha_s}{(\alpha_s+\beta_s)}, \phi_s = \alpha_s + \beta_s$ and optimize

$$\arg\max_{\phi_s} \sum_{j=1}^{J} \log B\left(y_{s,j} + \mu_s\phi_s,\ n_{s,j} - y_{s,j} + (1-\mu_s)\phi_s\right) - \log B\left(\mu_s\phi_s,\ (1-\mu_s)\phi_s\right)$$

$$\text{s.t. } 2 \leq \phi_s \leq \frac{\mu_s(1-\mu_s)}{\varepsilon} - 1; \forall s.$$

where $B(\cdot)$ is the Beta function. We take posterior mean of $y_{c,m,s}$ as our final diagnosis rate estimate, which is

$$\mathbb{E}\left[p_{c,m,s} \mid \alpha_{m,s},\ \beta_{m,s}\right] = \frac{y_{c,m,s} + \alpha_{m,s}}{n_{c,m,s} + \alpha_{m,s} + \beta_{m,s}}$$

To combine our per-stratum estimates into an adjusted per-market diagnosis rate $\widehat{p}_{c,m}$, we leverage post-stratification and target a nationally representative population with per-stratum weights w_s:

$$\widehat{p}_{c,m} = \sum_{s=1}^{S} w_s p_{c,m,s}$$

where $p_{c,m,s}$ is the posterior mean obtained in the last step, and w_s is the proportion of beneficiaries in the national switchers' cohort in the decile of interest.

After adjustment, we are left with independent, per-market "lists" of contract-level diagnosis rates. To aggregate these lists into a national ranking, we optimize a weighted version of the Bradley-Terry maximum likelihood objective using comparisons of adjusted diagnosis rates from MA contracts in the same market as input. We use boldface θ to represent the latent "strength" parameter for each contract. The standard Bradley-Terry objective is

$$\arg\max_{\boldsymbol{\theta}} \sum_{i<j} N_{ij}\ \log\frac{\theta_i}{\theta_i+\theta_j} + N_{ji}\ \log\frac{\theta_j}{\theta_i+\theta_j}$$

where i and j are contract indices, and N_{ij} is the number of weighted "matchups" in which Contract i has a higher adjusted diagnosis rate than Contract j, and N_{ji} is defined similarly:

$$N_{ij} \triangleq \sum_{m=1}^{M} \sum_{d=1}^{|HCC|} \omega_{m,i,j}^{(d)} \cdot \mathbf{1}\left[\widehat{p}_{i,m}^{(d)} > \widehat{p}_{j,m}^{(d)}\right]$$

where $\omega_{m,i,j}^{(d)}$ is a matchup weight that depends on both contracts, the market, and the diagnosis.

To mitigate the impact of small contracts in very few markets on the ranking, we add a term $\alpha \cdot \mathbb{E}\left[\omega_{m,i,j}^{(d)}\right]$\$ to the objective, equivalent to adding α pseudo-wins and pseudo-losses of "average" matchup weight to each comparison. Doing so regularizes win rates towards 50% with strength inversely proportional to the weight of "wins" in the observed data. This yields our final objective:

$$\arg\max_{\theta} \sum_{i<j} N_{ij} \ \log \frac{\theta_i}{\theta_i + \theta_j} + N_{ji} \ \log \frac{\theta_j}{\theta_i + \theta_j} + \alpha \, \mathbb{E}\left[\omega_{m,i,j}^{(d)}\right]$$

We take $\alpha = 5$ in our experiments. We define the weights to be

$$\omega_{m,i,j}^{(d)} = \left(\widehat{p}_{i,m}^{(d)} - \widehat{p}_{j,m}^{(d)}\right) \cdot HCC^{(d)} \cdot \frac{n_m}{\sum_{m=1}^{M} n_m}$$

where $HCC^{(d)}$ is the CMS-HCC V22 model weight for condition category d in the non-dual aged cohort, and n_m is the number of beneficiaries in our cohort in market m. Intuitively, we weight matchups by the "margin of victory," the CMS-HCC weight, and the overall market size. Doing so means that "wins" by a large margin, in high-leverage HCCs, in larger markets, have a greater impact on the ranking than smaller wins, wins in low-reimbursement HCCs, or small markets. The optimization proceeds by using rank centrality with a PageRank-style damping factor as an initialization to Hunter's MM, a standard approach for optimizing the Bradley-Terry model [208, 198, 122].

C.2.3 Methods - Simulation Study

To investigate the efficacy of our ranking method, we conduct a simulation study. At a high level, we simulate the process of auditing contracts following different rankings. We generate rankings with varying levels of accuracy as measured by the c-index across a variety of distributions describing simulated per-contract adjusted diagnosis misreporting rates. The c-index is a measure of

Formally, we draw per-contract diagnosis rates from a Beta distribution. We use a mean-variance parameterization of the Beta distribution $Beta\,(\mu, \sigma^2)$ where $\mu = \frac{\alpha}{\alpha+\beta}$ and $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ in terms of the standard α, β-parameterization. We sweep through $\mu \in \{0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6,\ 0.7,\ 0.8,\ 0.9\}$ and

$\sigma^2 \in \{0.020,\ 0.025,\ 0.030,\ 0.035,\ 0.040,\ 0.045,\ 0.050,\ 0.055\}$ and draw 427 per-contract diagnosis misreporting rates, matching the number of contracts in the switchers' cohort. These values stand in for the true adjusted misreporting rate for an arbitrary HCC for each contract.

To evaluate the efficiency of our ranking, we compute *audit recovery curves* on simulated noisy rankings with varying levels of accuracy. The audit recovery curve is defined as the cumulative proportion of misreported diagnoses audited as a function of the number of contracts audited when following a ranking of contracts. As a ranking accuracy metric, we use the c-index, which

corresponds to the percentage of contract pairs ranked in the same direction between the two rankings. In our setting, where all contracts are pairwise comparable, the c-index is equivalent to Kendall's tau + 1 divided by two between the predicted and true ranking.

We simulate 1000 rankings with c-index in [0.50, 0.55], [0.55, 0.60], ..., [0.95, 1] by adding mean-zero isotropic Gaussian noise to the sampled diagnosis misreporting rates from the Beta distribution described above. Note that we do not clip the outputs since we are merely interested in obtaining a noisy ranking. We adaptively set the noise as follows to ensure that ranking accuracy values span the entirety of the c-index bin.

More formally, for each c-index bin with bounds $[c_l,\ c_h]$, we begin with an input vector of diagnosis misreporting rates $\mathbf{v}$ and variance σ_t for some timestep t. The vector $\mathbf{v}$ stands in for a "perfect" ranking. Then, we generate a noisy ranking $\widetilde{\mathbf{v}} = \mathbf{v} + \varepsilon$, where $\varepsilon \ \sim \ N(0,\ \sigma_t^2 \cdot \mathbf{I})$. We compute the c-index of $\widetilde{\mathbf{v}}$ with respect to $\mathbf{v}$, which we call $C\left(\mathbf{v}, \widetilde{\mathbf{v}}\right)$. If $C\left(\mathbf{v}, \widetilde{\mathbf{v}}\right) < c_l$, our noisy ranking has a c-index that lies below the target bin, so we *reject* the sample $\widetilde{\mathbf{v}}$ and decrease the noise level: $\sigma_{t+1} := \sigma_t/(1+\alpha_t)$, where α_t is an adaptive step size parameter. Likewise, if $C\left(\mathbf{v}, \widetilde{\mathbf{v}}\right) > c_h$, our noisy ranking has a c-index above the target bin, so we similarly *reject* $\widetilde{\mathbf{v}}$ and increase the noise level $\sigma_{t+1} := \sigma_t * (1+\alpha_t)$.

If $c_l < C\left(\mathbf{v}, \widetilde{\mathbf{v}}\right) < c_h$, our noisy ranking has a c-index that lies within the target bin. Thus, we *accept* $\widetilde{\mathbf{v}}$ and adjust $\alpha_t = \alpha_0 e^{-5\,\overline{C}}$ where $\overline{C} = \frac{c_l + c_h}{2}$ is the midpoint of the c-index bin being targeted. Then, we compute the sample quartiles Q_1 and Q_3 and compute absolute distances Δ_l and Δ_h to c_l and c_h, respectively. If $\Delta_l > 1.5\Delta_h$, then the c-indices in the current bin skew high and we update $\sigma_{t+1} := \sigma_t * (1+\alpha_t)$. Otherwise, if $\Delta_h > 1.5\Delta_l$, the c-indices in the current bin skew low and we update $\sigma_{t+1} := \sigma_t/(1+\alpha_t)$. If neither holds, σ_t is left unchanged.

Table C.3: Eligible HCCs (prevalence >0.5%) in switchers' cohort, sorted in descending order of prevalence.

HCC (Version 22)	**Description**	**Count (% prevalence)**
HCC18	Diabetes with Chronic Complications	22,622 (17.61%)
HCC108	Vascular Disease	20,977 (16.33%)
HCC96	Specified Heart Arrhythmias	18,143 (14.12%)
HCC111	Chronic Obstructive Pulmonary Disease	15,277 (11.89%)
HCC85	Congestive Heart Failure	14,965 (11.65%)
HCC58	Major Depressive, Bipolar, and Paranoid Disorders	12,235 (9.52%)

Continued on next page

HCC (Version 22)	Description	Count (% prevalence)
HCC22	Morbid Obesity	10,286 (8.01%)
HCC19	Diabetes without Complication	10,071 (7.84%)
HCC40	Rheumatoid Arthritis and Inflammatory Connective Tissue Disease	8,443 (6.57%)
HCC12	Breast, Prostate, and Other Cancers and Tumors	8,394 (6.53%)
HCC48	Coagulation Defects and Other Specified Hematological Disorders	7,121 (5.54%)
HCC23	Other Significant Endocrine and Metabolic Disorders	4,790 (3.73%)
HCC135	Acute Renal Failure	4,718 (3.67%)
HCC88	Angina Pectoris	4,355 (3.39%)
HCC84	Cardio-Respiratory Failure and Shock	3,458 (2.69%)
HCC55	Drug/Alcohol Dependence	2,931 (2.28%)
HCC100	Ischemic or Unspecified Stroke	2,911 (2.27%)
HCC11	Colorectal, Bladder, and Other Cancers	2,335 (1.82%)
HCC2	Septicemia, Sepsis, Systemic Inflammatory Response Syndrome/Shock	2,292 (1.78%)
HCC161	Chronic Ulcer of Skin, Except Pressure	2,200 (1.71%)
HCC79	Seizure Disorders and Convulsions	2,138 (1.66%)
HCC107	Vascular Disease with Complications	2,137 (1.66%)
HCC124	Exudative Macular Degeneration	2,055 (1.60%)
HCC86	Acute Myocardial Infarction	1,921 (1.50%)
HCC10	Lymphoma and Other Cancers	1,836 (1.43%)
HCC47	Disorders of Immunity	1,727 (1.34%)
HCC75	Myasthenia Gravis/Myoneural Disorders and Guillain-Barre Syndrome/Inflammatory and Toxic Neuropathy	1,709 (1.33%)
HCC78	Parkinson's and Huntington's Diseases	1,678 (1.31%)
HCC176	Complications of Specified Implanted Device or Graft	1,552 (1.21%)
HCC21	Protein-Calorie Malnutrition	1,538 (1.20%)
HCC103	Hemiplegia/Hemiparesis	1,411 (1.10%)

Continued on next page

HCC (Version 22)	Description	Count (% prevalence)
HCC35	Inflammatory Bowel Disease	1,402 (1.09%)
HCC8	Metastatic Cancer and Acute Leukemia	1,360 (1.06%)
HCC33	Intestinal Obstruction/Perforation	1,343 (1.05%)
HCC112	Fibrosis of Lung and Other Chronic Lung Disorders	1,309 (1.02%)
HCC87	Unstable Angina and Other Acute Ischemic Heart Disease	1,305 (1.02%)
HCC115	Pneumococcal Pneumonia, Empyema, Lung Abscess	1,222 (0.95%)
HCC9	Lung and Other Severe Cancers	1,197 (0.93%)
HCC169	Vertebral Fractures without Spinal Cord Injury	1,143 (0.89%)
HCC170	Fracture/Dislocation	1,018 (0.79%)
HCC137	Chronic Kidney Disease, Severe (Stage 4)	1,001 (0.78%)
HCC188	Artificial Openings for Feeding or Elimination	899 (0.70%)
HCC122	Proliferative Diabetic Retinopathy and Vitreous Hemorrhage	874 (0.68%)
HCC39	Bone/Joint/Muscle Infection/Necrosis	825 (0.64%)
HCC189	Amputation Status, Lower Limb/Amputation Complications	655 (0.51%)

Table C.4: HCC to treatment (ATC drug categories or RBCS procedure categories) mappings used for treatment gap analysis.

HCC(s)	ATCs / procedures
Diabetes with complications (HCC 18)	ATC A10
Major Depressive, Bipolar, and Paranoid Disorders (HCC58)	ATC N06A; RBCS EB
Specified Heart Arrhythmias, Vascular Disease with Complications, Vascular Disease (HCC96, 107, 108)	ATC B01A, C01B, C07A, C04A; RBCS PC
Chronic Obstructive Pulmonary Disease (HCC111)	ATC R03; RBCS TF, TP
Parkinson's and Huntington's Diseases (HCC78)	ATC N04

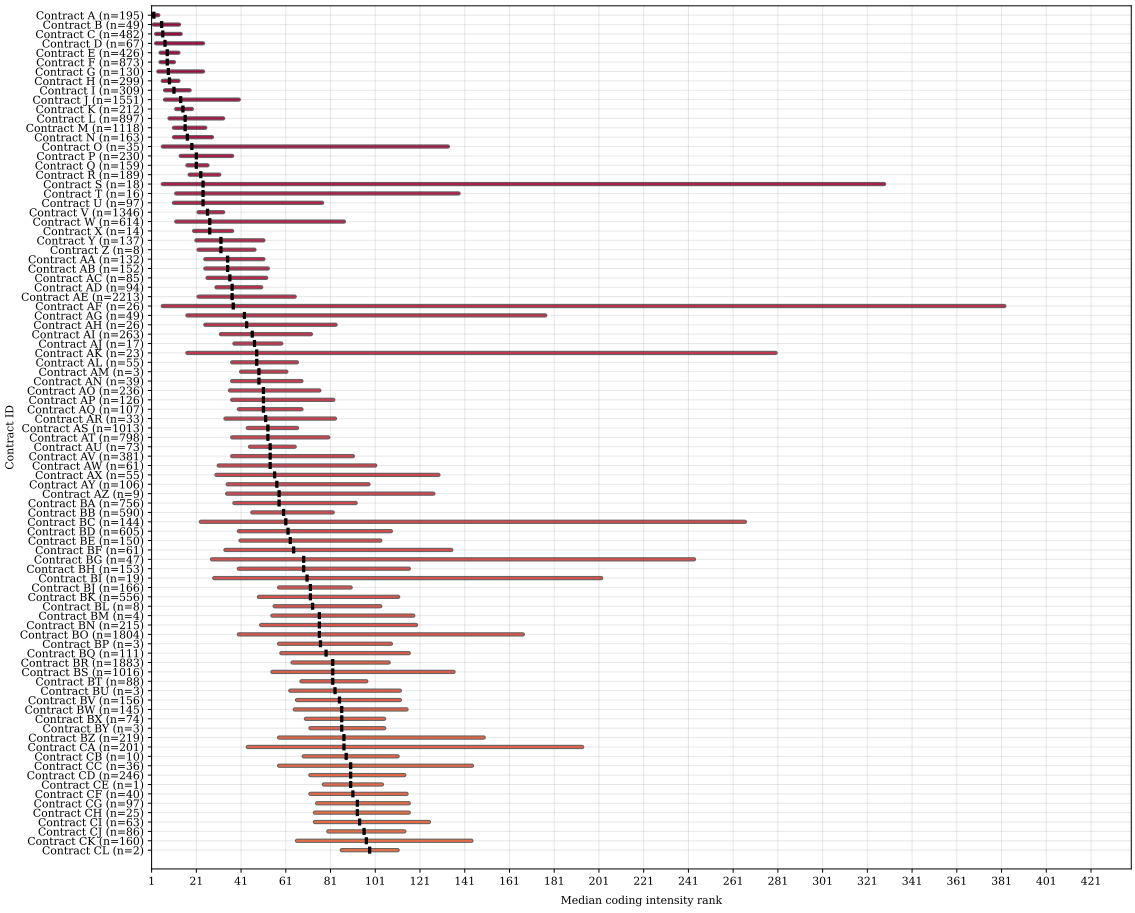


Figure C.26: Contracts ranking 1st to 90th (inclusive) in switchers' cohort, (2018 TM to 2019 MA).

Table C.5: Correlation matrix (Kendall's τ) between different rankings with empirical 95% confidence intervals across bootstraps.

	Ours	**No market split**	**No BT**	**No adjustment**
Ours	1	0.196 (0.157, 0.239)	0.249 (0.215, 0.281)	0.191 (0.158, 0.223)
No market split		1	0.409 (0.375, 0.443)	0.276 (0.248, 0.303)
No BT			1	0.468 (0.442, 0.494)
No adjustment				1

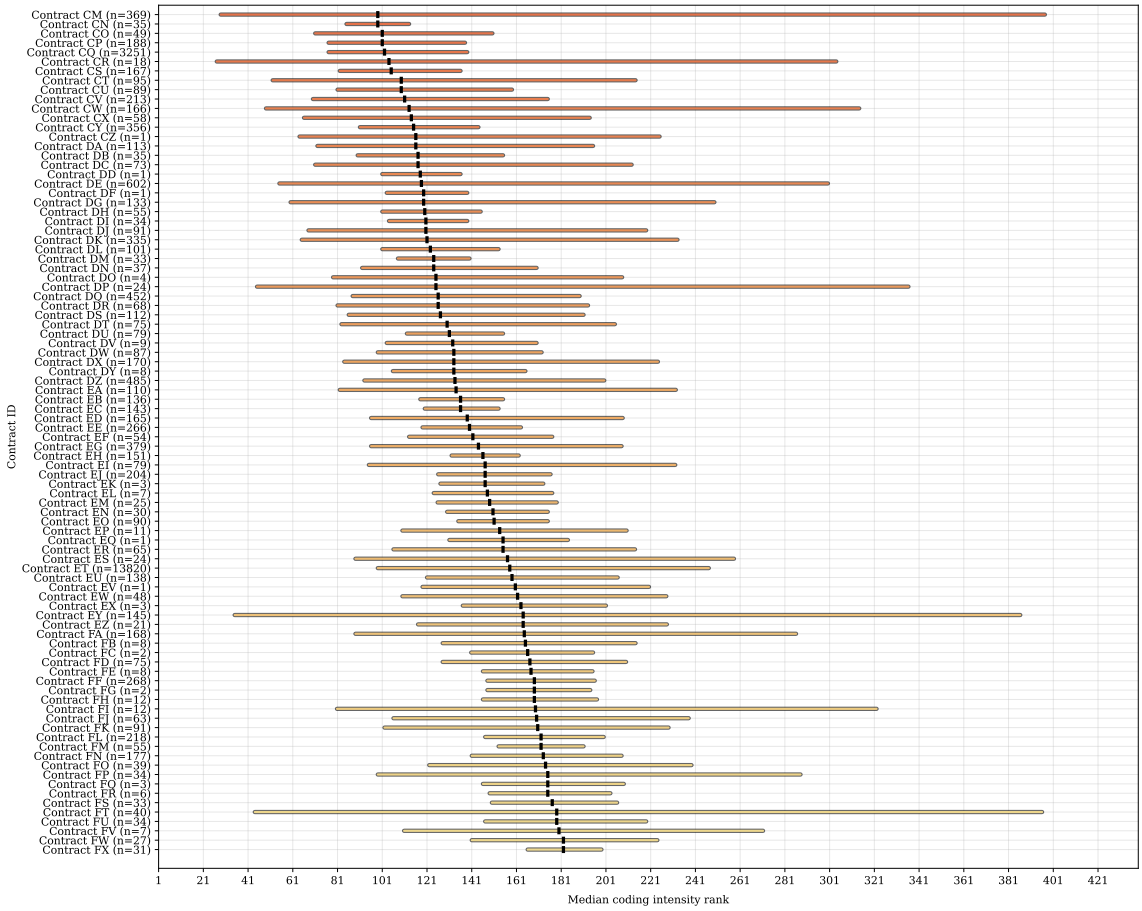


Figure C.27: Contracts ranking 91st to 180th (inclusive) in switchers' cohort, (2018 TM to 2019 MA).

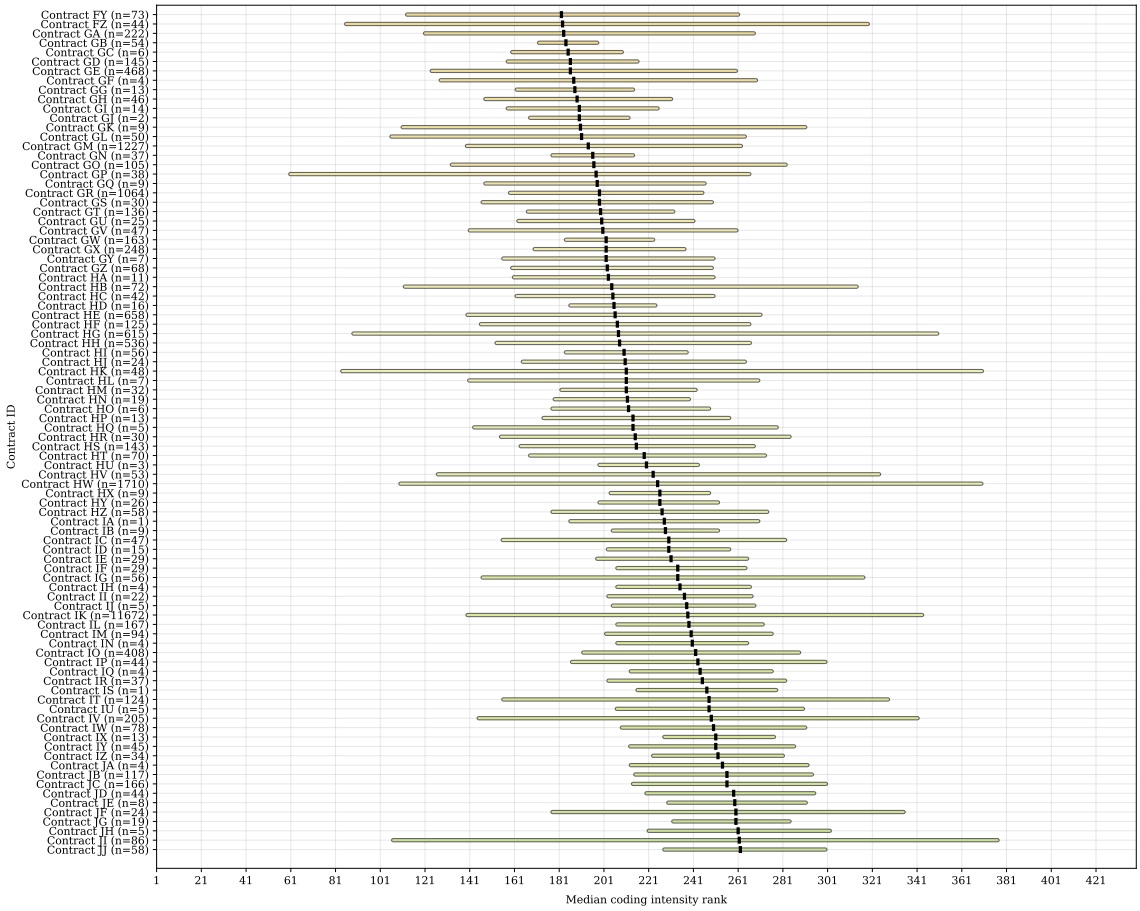


Figure C.28: Contracts ranking 181st to 270th (inclusive) in switchers' cohort, (2018 TM to 2019 MA).

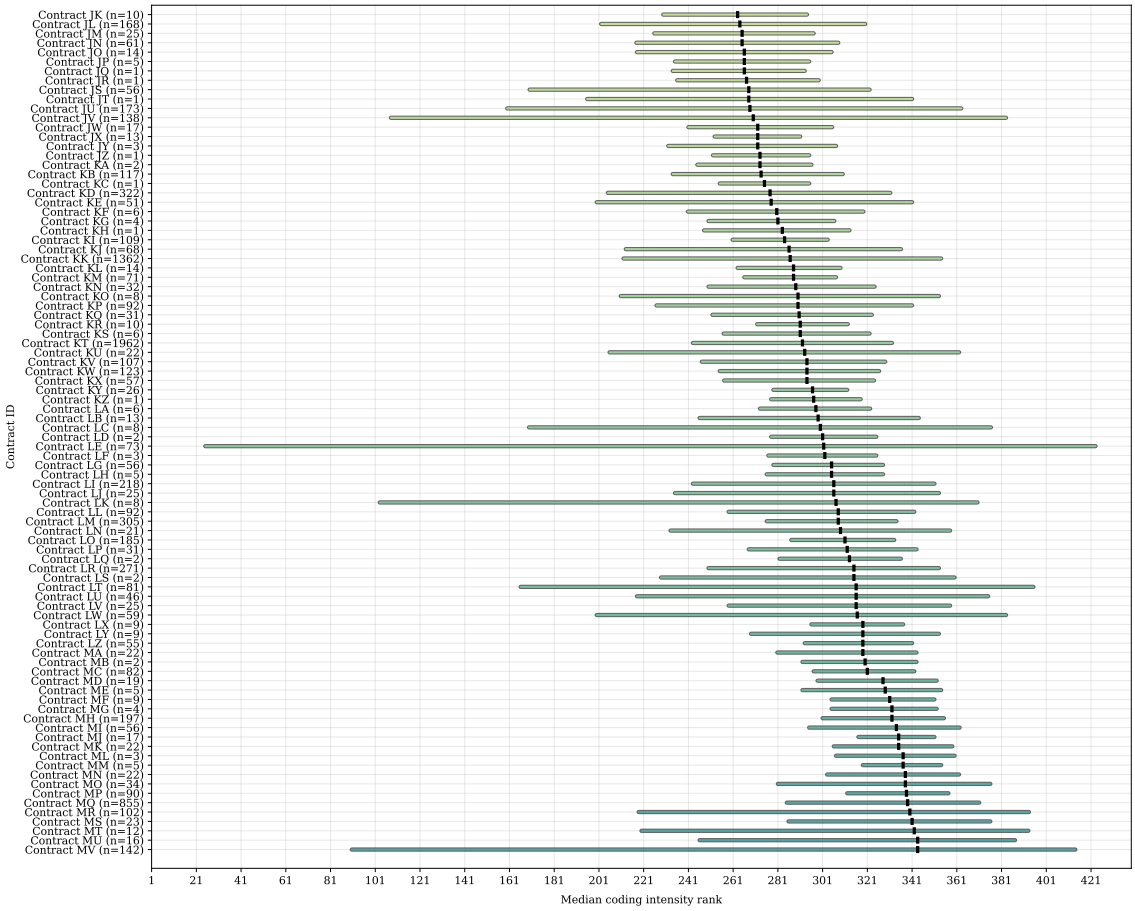


Figure C.29: Contracts ranking 271st to 360th (inclusive) in switchers' cohort, (2018 TM to 2019 MA).

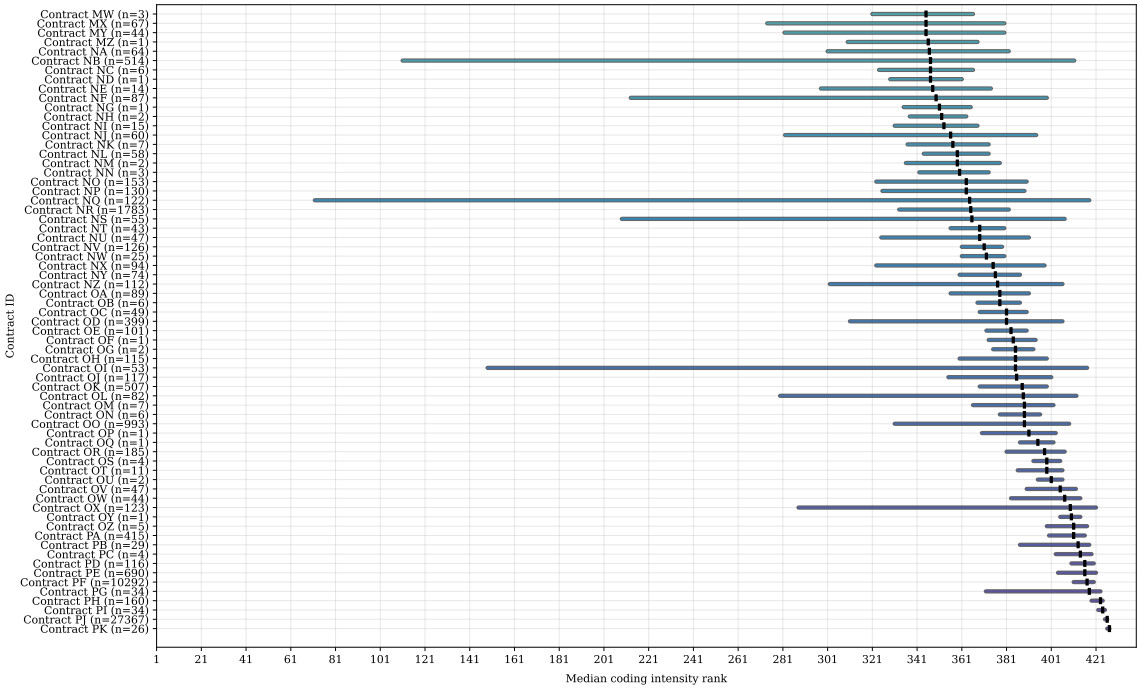


Figure C.30: Contracts ranking 361st to 427th (inclusive) in switchers' cohort, (2018 TM to 2019 MA).

Table C.6: Raw rates of relevant treatments by contract decile for diabetes with complications (HCC18). ATC A10: drugs for diabetes.

Decile	**Treatment rate (ATC A10 prescription)**	**Total enrollment**
1	78.5%	15,496
2	77.5%	13,336
3	75.9%	7,048
4	77.5%	8,286
5	78.4%	3,373
6	79.7%	4,757
7	77.5%	2,348
8	77.9%	3,389
9	80.2%	2,018
10	78.8%	64,049

Table C.7: Raw rates of relevant treatments by contract decile for major depressive, bipolar, and paranoid disorders (HCC58). ATC N06A: antidepressants; RBCS EB: evaluation and management, behavioral health.

Decile	Treatment rate (ATC N06A prescription or RBCS EB code)	Total enrollment
1	78.0%	12,170
2	81.8%	10,966
3	82.0%	10,016
4	86.1%	3,124
5	82.3%	6,045
6	83.5%	3,208
7	86.3%	2,826
8	79.3%	4,411
9	86.0%	15,824
10	87.1%	55,064

Table C.8: Raw rates of relevant treatments by contract decile for specified heart arrhythmias, vascular disease with complications, and vascular disease (HCC96, 107, and 108). ATC B01A: antithrombotic agents; C01B: antiarrhythmics (class I and III); C07A: beta-blocking agents; C04A: peripheral vasodilators; RBCS PC: cardiovascular procedures.

Decile	Treatment rate (ATC B01A, C01B, C07A, C04A or RBCS PC code)	Total enrollment
1	70.1%	27,222
2	71.9%	7,925
3	71.8%	6,238
4	72.3%	2,714
5	74.2%	5,830
6	74.7%	4,621
7	75.2%	2,411
8	73.6%	7,734
9	74.8%	5,682
10	75.3%	53,795

Table C.9: Raw rates of relevant treatments by contract decile for chronic obstructive pulmonary disease (HCC111). ATC R03: drugs for obstructive airway diseases; RBCS TF/TP: pulmonary function tests.

Decile	Treatment rate (ATC R03 prescription or RBCS TF/TP code)	Total enrollment
1	58.9%	14,987
2	56.0%	9,463
3	56.8%	8,974
4	57.1%	6,424
5	56.4%	15,090
6	57.5%	2,666
7	54.0%	2,158
8	58.8%	6,053
9	63.7%	2,569
10	65.9%	55,354

Table C.10: Raw rates of relevant treatments by contract decile for Parkinson's and Huntington's diseases (HCC78). ATC N04: anti-parkinson drugs.

Decile	Treatment rate (ATC N04 prescription)	Total enrollment
1	62.9%	3,308
2	69.7%	29,460
3	72.2%	1,952
4	66.0%	4,211
5	65.2%	1,989
6	61.4%	4,582
7	65.1%	7,822
8	70.6%	18,030
9	71.9%	6,264
10	70.0%	38,561

BIBLIOGRAPHY

[1] Your Medicare coverage choices — medicare.gov. https://www.medicare.gov/what-medicare-covers/your-medicare-coverage-choices. [Accessed 25-04-2024].

[2] Provider of services files, Sep 2022.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] Roy Adams, Katharine E Henry, Anirudh Sridharan, Hossein Soleimani, Andong Zhan, Nishi Rawat, Lauren Johnson, David N Hager, Sara E Cosgrove, Andrew Markowski, et al. Prospective, multi-site study of patient outcomes after implementation of the trews machine learning-based early warning system for sepsis. *Nature Medicine*, pages 1–6, 2022.

[5] C Ambroise and G Govaert. EM algorithm for partially known labels. In *Data Analysis, Classification, and Related Methods*, pages 161–166. Springer, 2000.

[6] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[7] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[8] David Arbour, Drew Dimmery, and Arjun Sondhi. Permutation weighting. In *International Conference on Machine Learning*, pages 331–341. PMLR, 2021.

[9] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242, 2017.

[10] Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*, 2023.

[11] Mohammad Taha Bahadori, Krzysztof Chalupka, Edward Choi, Robert Chen, Walter F. Stewart, and Jimeng Sun. Causal regularization. *arXiv preprint arXiv:1702.02604*, 2017.

[12] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[13] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

[14] Sidhika Balachandar, Nikhil Garg, and Emma Pierson. Domain constraints improve risk prediction when outcome data is missing. In *12th International Conference on Learning Representations*, 2024.

[15] Jane Bambauer and Tal Zarsky. The algorithm game. *Notre Dame L. Rev.*, 94:1, 2018.

[16] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[17] Richard A Bauder and Taghi M Khoshgoftaar. Medicare fraud detection using machine learning methods. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 858–865. IEEE, 2017.

[18] Yahav Bechavod, Katrina Ligett, Steven Wu, and Juba Ziani. Gaming helps! learning from strategic interactions in natural dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 1234–1242. PMLR, 2021.

[19] Theresa A Beery. Gender bias in the diagnosis and treatment of coronary artery disease. *Heart & Lung*, 24(6):427–435, 1995.

[20] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020.

[21] Jessa Bekker, Pieter Robberechts, and Jesse Davis. Beyond the selected completely at random assumption for learning from positive and unlabeled data. In *Machine Learning and Knowledge Discovery in Databases*, pages 71–85, Cham, 2020.

[22] Peter Bergman, Elizabeth Kopko, and Julio E Rodriguez. A seven-college experiment using algorithms to track students: Impacts and implications for equity and fairness. Technical report, National Bureau of Economic Research, 2021.

[23] Jamillah Berry, Kevin Bumpers, Vickie Ogunlade, Roni Glover, Sharon Davis, Margaret Counts-Spriggs, John Kauh, and Christopher Flowers. Examining racial disparities in colorectal cancer care. *Journal of psychosocial oncology*, 27(1):59–83, 2009.

[24] Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. In *International Conference on Machine Learning*, pages 825–836. PMLR, 2021.

[25] Anish N Bhuva, Wenjia Bai, Clement Lau, Rhodri H Davies, Yang Ye, Heeraj Bulluck, Elisa McAlindon, Veronica Culotta, Peter P Swoboda, Gabriella Captur, et al. A multicenter, scan-rescan, human and machine learning cmr study to test generalizability and precision in imaging biomarker analysis. *Circulation: Cardiovascular Imaging*, 12(10):e009214, 2019.

[26] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

[27] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II 9*, pages 405–415, 2017.

[28] Daniel Björkegren, Joshua E Blumenstock, and Samsun Knight. Manipulation-proof machine learning. *arXiv preprint arXiv:2004.03865*, 2020.

[29] Daniel Björkegren and Darrell Grissen. Behavior revealed in mobile phone usage predicts credit repayment. *The World Bank Economic Review*, 34(3):618–634, 2020.

[30] Avrim Blum and Kevin Stangl. Recovering from biased data: Can fairness constraints improve accuracy? In *1st Symposium on Foundations of Responsible Computing*, 2020.

[31] Suresh Bolusani, Mathieu Besançon, Ksenia Bestuzheva, Antonia Chmiela, João Dionísio, Tim Donkiewicz, Jasper van Doornmalen, Leon Eifler, Mohammed Ghannam, Ambros Gleixner, Christoph Graczyk, Katrin Halbig, Ivo Hedtke, Alexander Hoen, Christopher Hojny, Rolf van der Hulst, Dominik Kamp, Thorsten Koch, Kevin Kofler, Jurgen Lentz, Julian Manns, Gioni Mexi, Erik Mühmer, Marc E. Pfetsch, Franziska Schlösser, Felipe Serrano, Yuji Shinano, Mark Turner, Stefan Vigerske, Dieter Weninger, and Lixing Xu. The SCIP Optimization Suite 9.0. Technical report, Optimization Online, February 2024.

[32] Ralph A. Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3–4):324–345, 1952.

[33] Joschka Braun, Dmitrii Krasheninnikov, Usman Anwar, Robert Kirk, Daniel Tan, and David Scott Krueger. A sober look at steering vectors for llms. In *AI Alignment Forum, November 2024b. URL https://www. alignmentforum. org/posts/QQP4nq7TXg89CJGBh/a-sober-look-at-steering-vectors-for-llms*, 2024.

[34] Elizabeth Brondolo, Linda C Gallo, and Hector F Myers. Race, racism and health: disparities, mechanisms, and interventions. *Journal of Behavioral Medicine*, 32(1):1–8, 2009.

[35] Jason Brown, Mark Duggan, Ilyana Kuziemko, and William Woolston. How does risk selection respond to risk adjustment? new evidence from the medicare advantage program. *American Economic Review*, 104(10):3335–3364, 2014.

[36] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[37] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018.

[38] Manuela Cattelan, Cristiano Varin, and David Firth. Dynamic bradley–terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):135–150, 2013.

[39] Centers for Medicare & Medicaid Services. Cms rolls out aggressive strategy to enhance and accelerate medicare advantage audits, 2025.

[40] Centers for Medicare & Medicaid Services. Payment year 2018 medicare advantage contract-specific risk adjustment data validation (radv) audit methods and instructions, 2025.

[41] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011.

[42] Trenton Chang, Michael W. Sjoding, and Jenna Wiens. Disparate censorship & undertesting: A source of label bias in clinical machine learning. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pages 343–390, Aug 2022.

[43] 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. In *Advances in Neural Information Processing Systems*, pages 42311–42348, 2024.

[44] Jian Chen, Zhehao Li, and Xiaojie Mao. Learning under selective labels with heterogeneous decision-makers: An instrumental variable approach. *arXiv preprint arXiv:2306.07566*, 2023.

[45] Kevin Chen, Marco Cusumano-Towner, Brody Huval, Aleksei Petrenko, Jackson Hamburger, Vladlen Koltun, and Philipp Krähenbühl. Reinforcement learning for long-horizon interactive llm agents. *arXiv preprint arXiv:2502.01600*, 2025.

[46] Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. Summscreen: A dataset for abstractive screenplay summarization. *arXiv preprint arXiv:2104.07091*, 2021.

[47] Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020.

[48] Yiling Chen, Yang Liu, and Chara Podimata. Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems*, 33:15265–15276, 2020.

[49] Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance and label-dependent label noise. In *International Conference on Machine Learning*, pages 1789–1799. PMLR, 2020.

[50] Michael E. Chernew. Report to the Congress: Medicare Payment Policy, 2022.

[51] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[52] Danielle Keats Citron and Frank Pasquale. The scored society: Due process for automated predictions. *Wash. L. Rev.*, 89:1, 2014.

[53] Medicare Payment Advisory Commission. A data book: Health care spending and the medicare program, section 9: Medicare advantage, 2025.

[54] BIG-Bench contributors. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2023.

[55] Alicia Curth and Mihaela Van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1810–1818. PMLR, 2021.

[56] Vilsa E. Curto, Eran Politzer, Timothy S. Anderson, John Z. Ayanian, Jeffrey Souza, Alan M. Zaslavsky, and Bruce E. Landon. Coding intensity variation in medicare advantage. *Health Affairs Scholar*, 3(1):qxae176, 2025.

[57] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023.

[58] Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. Proxy Non-Discrimination in Data-Driven Systems, July 2017. arXiv:1707.08120 [cs].

[59] Stacie L Daugherty, Irene V Blair, Edward P Havranek, Anna Furniss, L Miriam Dickinson, Elhum Karimkhani, Deborah S Main, and Frederick A Masoudi. Implicit gender bias and the use of cardiovascular tests among cardiologists. *Journal of the American Heart Association*, 6(12):e006872, 2017.

[60] Daniel De Roux, Boris Perez, Andrés Moreno, Maria del Pilar Villamil, and César Figueroa. Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 215–222, 2018.

[61] Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, May 2021.

[62] Ryan J Delahanty, JoAnn Alvarez, Lisa M Flynn, Robert L Sherwin, and Spencer S Jones. Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. *Annals of Emergency Medicine*, 73(4):334–344, 2019.

[63] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (methodological)*, 39(1):1–22, 1977.

[64] Yuntian Deng, Wenting Zhao, Jack Hessel, Xiang Ren, Claire Cardie, and Yejin Choi. WildVis: Open Source Visualizer for Million-Scale Chat Logs in the Wild. *arXiv preprint arXiv:2409.03753*, 2024.

[65] Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

[66] Georges Dionne, Florence Giuliano, and Pierre Picard. Optimal auditing with scoring: Theory and application to insurance fraud. *Management Science*, 55(1):58–70, 2009.

[67] Benjamin Djulbegovic, Shira Elqayam, Tea Reljic, Iztok Hozo, Branko Miladinovic, Athanasios Tsalatsanis, Ambuj Kumar, Jason Beckstead, Stephanie Taylor, and Janice Cannon-Bowers. How do physicians decide to treat: an empirical evaluation of the threshold model. *BMC Medical Informatics and Decision Making*, 14:1–10, 2014.

[68] Nancy C Dolan, M Rosario Ferreira, Marian L Fitzgibbon, Terry C Davis, Alfred W Rademaker, Dachao Liu, June Lee, Michael Wolf, Brian P Schmitt, and Charles L Bennett. Colorectal cancer screening among African-American and white male veterans. *American Journal of Preventive Medicine*, 28(5):479–482, 2005.

[69] Guanting Dong, Keming Lu, Chengpeng Li, Tingyu Xia, Bowen Yu, Chang Zhou, and Jingren Zhou. Self-play with execution feedback: Improving instruction-following capabilities of large language models. *arXiv preprint arXiv:2406.13542*, 2024.

[70] 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.

[71] Kacie L Dragan, Sunita M Desai, John Billings, and Sherry A Glied. Association of insurance mix and diagnostic coding practices in new york state hospitals. *JAMA Health Forum*, 3(9):e222919–e222919, 2022.

[72] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

[73] Olive Jean Dunn. Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56(293):52–64, March 1961.

[74] Esin Durmus, Alex Tamkin, Jack Clark, Jerry Wei, Jonathan Marcus, Joshua Batson, Kunal Handa, Liane Lovitt, Meg Tong, Miles McCain, Oliver Rausch, Saffron Huang, Sam Bowman, Stuart Ritchie, Tom Henighan, and Deep Ganguli. Evaluating feature steering: A case study in mitigating social biases, 2024.

[75] Edward J. Hu and Yelong Shen and Phillip Wallis and Zeyuan Allen-Zhu and Yuanzhi Li and Shean Wang and Lu Wang and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[76] Tahir Ekin, R Muzaffer Musal, and Lawrence V Fulton. Overpayment models for medical audits: multiple scenarios. *Journal of Applied Statistics*, 42(11):2391–2405, 2015.

[77] Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 30284–30297, 2021.

[78] Nasibeh Zanjirani Farahani, Divaakar Siva Baala Sundaram, Moein Enayati, Shivaram Poigai Arunachalam, Kalyan Pasupathy, and Adelaide M Arruda-Olson. Explanatory analysis of a machine learning model to identify hypertrophic cardiomyopathy patients from EHR using diagnostic codes. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1932–1937, 2020.

[79] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer, 1970.

[80] Lucas M Fleuren, Thomas LT Klausch, Charlotte L Zwager, Linda J Schoonmade, Tingjie Guo, Luca F Roggeveen, Eleonora L Swart, Armand RJ Girbes, Patrick Thoral, Ari Ercole, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive care medicine*, 46(3):383–400, 2020.

[81] Centers for Medicare and Medicaid Services. Fy2024 justification of estimates for appropriations committees, 2024.

[82] Center for Medicare & Medicaid Center. Note to: Medicare advantage organizations, prescription drug plan sponsors, and other interested parties, 2023.

[83] Konrad Furmańczyk, Jan Mielniczuk, Wojciech Rejchel, and Paweł Teisseyre. Joint estimation of posterior probability and propensity score function for positive and unlabelled data. *arXiv preprint arXiv:2209.07787*, 2022.

[84] Jonathan M Gaffin, Nancy Lichtenberg Shotola, Thomas R Martin, and Wanda Phipatanakul. Clinically useful spirometry in preschool-aged children: evaluation of the 2007 american thoracic society guidelines. *Journal of Asthma*, 47(7):762–767, 2010.

[85] Faiz Gani, Joseph K Canner, and Timothy M Pawlik. Assessing coding practices for gastrointestinal surgery over time in the united states. *Surgery*, 164(3):530–538, 2018.

[86] Josh Gardner, Christopher Brooks, and Ryan Baker. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 225–234, 2019.

[87] Arpit Garg, Cuong Nguyen, Rafael Felix, Thanh-Toan Do, and Gustavo Carneiro. Instance-dependent noisy label learning via graphical modelling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2288–2298, 2023.

[88] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

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

[90] Walter Gerych, Thomas Hartvigsen, Luke Buquicchio, Emmanuel Agu, and Elke Rundensteiner. Recovering the propensity score from biased positive unlabeled data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6694–6702, 2022.

[91] Zoubin Ghahramani, Geoffrey E Hinton, et al. The EM algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto, 1996.

[92] Zoubin Ghahramani and Michael Jordan. Supervised learning from incomplete data via an EM approach. *Advances in Neural Information Processing Systems*, 6, 1993.

[93] Shahrzad Gholami, Sara Mc Carthy, Bistra Dilkina, Andrew Plumptre, Milind Tambe, Margaret Driciru, Fred Wanyama, Aggrey Rwetsiba, Mustapha Nsubaga, Joshua Mabonga, et al. Adversary models account for imperfect crime data: Forecasting and planning against real-world poachers. In *International Conference on Autonomous Agents and Multiagent Systems*, 2018.

[94] Judy Wawira Gichoya, Liam G McCoy, Leo Anthony Celi, and Marzyeh Ghassemi. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ health & care informatics*, 28(1):e100289, 2021.

[95] Adam N Glynn and Kevin M Quinn. An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1):36–56, 2010.

[96] Chamal Gomes, Zhuo Jin, and Hailiang Yang. Insurance fraud detection with unsupervised deep learning. *Journal of Risk and Insurance*, 88(3):591–624, 2021.

[97] Chen Gong, Qizhou Wang, Tongliang Liu, Bo Han, Jane Jia You, Jian Yang, and Dacheng Tao. Instance-dependent positive and unlabeled learning with labeling bias estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:4163–4177, 2021.

[98] Luke Guerdan, Amanda Coston, Kenneth Holstein, and Zhiwei Steven Wu. Counterfactual prediction under outcome measurement error. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, page 1584–1598, 2023.

[99] Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein. Ground(less) truth: A causal framework for proxy labels in human-algorithm decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, page 688–704, 2023.

[100] Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate, 2022.

[101] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 8536–8546, 2018.

[102] John T Hancock, Huanjing Wang, Taghi M Khoshgoftaar, and Qianxin Liang. Data reduction techniques for highly imbalanced medicare big data. *Journal of Big Data*, 11(1):8, 2024.

[103] 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.

[104] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[105] Sahand Hariri, Matias Carrasco Kind, and Robert J Brunner. Extended isolation forest. *IEEE transactions on knowledge and data engineering*, 33(4):1479–1489, 2019.

[106] Frank E. Harrell. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543, 1982.

[107] 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.

[108] Thomas Hartvigsen, Cansu Sen, Sarah Brownell, Erin Teeple, Xiangnan Kong, and Elke A Rundensteiner. Early prediction of mrsa infections using electronic health records. In *HEALTHINF*, pages 156–167, 2018.

[109] Qianyu He, Jie Zeng, Qianxi He, Jiaqing Liang, and Yanghua Xiao. From complex to simple: Enhancing multi-constraint complex instruction following ability of large language models. *arXiv preprint arXiv:2404.15846*, 2024.

[110] James Heckman. Varieties of Selection Bias. *The American Economic Review*, 80(2):313–318, 1990. Publisher: American Economic Association.

[111] James J Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4*, pages 475–492. NBER, 1976.

[112] Peter Henderson, Ben Chugg, Brandon Anderson, Kristen Altenburger, Alex Turk, John Guyton, Jacob Goldin, and Daniel E Ho. Integrating reward maximization and population estimation: Sequential decision-making for internal revenue service audit selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5087–5095, 2023.

[113] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2021.

[114] Katharine E Henry, David N Hager, Tiffany M Osborn, Albert W Wu, and Suchi Saria. Comparison of automated sepsis identification methods and electronic health record–based sepsis phenotyping: improving case identification accuracy by accounting for confounding comorbid conditions. *Critical care explorations*, 1(10), 2019.

[115] Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science translational medicine*, 7(299):299ra122–299ra122, 2015.

[116] Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, pages 615–625, 2004.

[117] Francis Heylighen and Jean-Marc Dewaele. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*, 4(1), 1999.

[118] Kelly M Hoffman, Sophie Trawalter, Jordan R Axt, and M Norman Oliver. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences*, 113(16):4296–4301, 2016.

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

[120] Woo Suk Hong, Adrian Daniel Haimovich, and R. Andrew Taylor. Predicting hospital admission at emergency department triage using machine learning. *PLOS ONE*, 13(7):e0201016, July 2018.

[121] Xinting Hu, Yulei Niu, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. On non-random missing labels in semi-supervised learning. In *10th International Conference on Learning Representations*, 2022.

[122] David R. Hunter. Mm algorithms for generalized bradley–terry models. *Annals of Statistics*, 32(1), 2004.

[123] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[124] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[125] John PA Ioannidis, Neil R Powe, and Clyde Yancy. Recalibrating the use of race in medical research. *Jama*, 325(7):623–624, 2021.

[126] Sarah Jabbour, David Fouhey, Ella Kazerooni, Michael W Sjoding, and Jenna Wiens. Deep learning applied to chest x-rays: Exploiting and preventing shortcuts. In *Machine Learning for Healthcare Conference*, pages 750–782. PMLR, 2020.

[127] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[128] Dominik Janzing. Causal regularization. *Advances in Neural Information Processing Systems*, 32, 2019.

[129] Scott Jarvis and Brett James Hashimoto. How operationalizations of word types affect measures of lexical diversity. *International Journal of Learner Corpus Research*, 7(1):163–194, 2021.

[130] Lara Jehi, Xinge Ji, Alex Milinovich, Serpil Erzurum, Brian P Rubin, Steve Gordon, James B Young, and Michael W Kattan. Individualizing risk prediction for positive coronavirus disease 2019 testing: results from 11,672 patients. *Chest*, 158(4):1364–1375, 2020.

[131] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR, 2020.

[132] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *version 0.4). PhysioNet. https://doi. org/10.13026/a3wn-hq05*, 2020.

[133] Alistair EW Johnson, Jerome Aboab, Jesse D Raffa, Tom J Pollard, Rodrigo O Deliberato, Leo A Celi, and David J Stone. A comparative analysis of sepsis identification methods in an electronic database. *Critical Care Medicine*, 46(4):494–499, 2018.

[134] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.

[135] Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal effect inference for structured treatments. *Advances in Neural Information Processing Systems*, 34:24841–24854, 2021.

[136] Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*, 2023.

[137] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd international conference on artificial intelligence and statistics*, pages 2281–2290. PMLR, 2019.

[138] Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *Advances in Neural Information Processing Systems*, 32:3438–3448, 2019.

[139] Fahad Kamran, Shengpu Tang, Erkin Ötleş, Dustin S McEvoy, Sameh N Saleh, Jen Gong, Benjamin Y Li, Sayon Dutta, Xinran Liu, Richard J Medford, Thomas S Valley, Lauren R West, Karandeep Singh, Seth Blumberg, John P Donnelly, Erica S Shenoy, John Z Ayanian, Brahmajee K Nallamothu, Michael W Sjoding, and Jenna Wiens. Early identification of patients admitted to hospital for covid-19 at risk of clinical deterioration: model development and multisite external validation study. *The BMJ*, 376, 2022.

[140] Masahiro Kato, Shuting Wu, Kodai Kureishi, and Shota Yasui. Automatic debiased learning from positive, unlabeled, and exposure data. *arXiv preprint arXiv:2303.04797*, 2023.

[141] Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93, 1938.

[142] Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

[143] Kennes, Lieven Nils, Erhard Cramer, Ralf-Dieter Hilgers, and Nicole Heussen. The impact of selection bias on test decisions in randomized clinical trials. *Statistics in Medicine*, 30(21):2573–2581, September 2011.

[144] Anam Khan, Mao Li, Jessica Finlay, Michael Esposito, Iris Gomez-Lopez, Philippa Clarke, and Megan Chenoweth. National neighborhood data archive (nanda): Health care services by census tractstates, 2003-2017 [data set]. *Inter-University Consortium for Political and Social Research (ICPSR)*, 10:E120907V2, 2020.

[145] Sohaib Kiani, Jared Barton, Jon Sushinsky, Lynda Heimbach, and Bo Luo. Counterfactual prediction under selective confounding. *arXiv preprint arXiv:2310.14064*, 2023.

[146] Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. Abstractive summarization of reddit posts with multi-level memory networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2018.

[147] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. *Institute for Simulation and Training, University of Central Florida*, 1975.

[148] Gary King and Richard Nielsen. Why Propensity Scores Should Not Be Used for Matching. *Political Analysis*, 27(4):435–454, October 2019.

[149] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015.

[150] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2018.

[151] Constantine E Kontokosta and Boyeong Hong. Bias in smart city governance: How socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions. *Sustainable Cities and Society*, 64:102503, 2021.

[152] Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE Samples, Get a Baseline for Free! . *ICLR Deep Reinforcement Learning Meets Structured Prediction Workshop*, 2019.

[153] Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36:47669–47681, 2023.

[154] Richard Kronick, F Michael Chua, Ramona Krauss, Logan Johnson, and Daniel Waldo. Insurer-level estimates of revenue from differential coding in medicare advantage. *Annals of internal medicine*, 178(5):655–662, 2025.

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

[156] Richard Kronick and W. Pete Welch. Measuring coding intensity in the medicare advantage program. *Medicare & Medicaid Research Review*, 4(2):mmrr2014.004.02.a06, 2014.

[157] Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. BOOKSUM: A Collection of Datasets for Long-form Narrative Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, 2021.

[158] 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.

[159] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention, 2023.

[160] Daniel T Lackland. Racial differences in hypertension: implications for high blood pressure management. *The American Journal of the Medical Sciences*, 348(2):135–138, 2014.

[161] Riku Laine, Antti Hyttinen, and Michael Mathioudakis. Evaluating decision makers over selectively labelled data: A causal modelling approach. In *Discovery Science: 23rd International Conference, DS 2020, Thessaloniki, Greece, October 19–21, 2020, Proceedings 23*, pages 3–18, 2020.

[162] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284, 2017.

[163] Benjamin Laufer, Emma Pierson, and Nikhil Garg. End-to-end auditing of decision pipelines. In *ICML Workshop on Responsible Decision-Making in Dynamic Environments.*, pages 1–7, 2022.

[164] Thomas A. LaVeist, Eliseo J. Pérez-Stable, Patrick Richard, Andrew Anderson, Lydia A. Isaac, Riley Santiago, Celine Okoh, Nancy Breen, Tilda Farhat, Assen Assenov, and Darrell J. Gaskin. The Economic Burden of Racial, Ethnic, and Educational Health Inequities in the US. *JAMA*, 329(19):1682, May 2023.

[165] Tosca Lechner and Ruth Urner. Learning losses for strategic classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7337–7344, 2022.

[166] Tosca Lechner, Ruth Urner, and Shai Ben-David. Strategic classification with unknown user manipulations. In *International Conference on Machine Learning*, pages 18714–18732. PMLR, 2023.

[167] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3(2), page 896, 2013.

[168] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1603–1612, 2015.

[169] Elle Lett, Nadia L Dowshen, and Kellan E Baker. Intersectionality and health inequities for gender minority blacks in the us. *American Journal of Preventive Medicine*, 59(5):639–647, 2020.

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

[171] Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *8th International Conference on Learning Representations, ICLR*, 2020.

[172] Junyi Li, Ninareh Mehrabi, Charith Peris, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. On the steerability of large language models toward data-driven personas. *arXiv preprint arXiv:2311.04978*, 2023.

[173] Xuefeng Li, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. Provably end-to-end label-noise learning without anchor points. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6403–6413, 2021.

[174] Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George H Chen. Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12181–12193, 2022.

[175] Judith H Lichtman, Erica C Leifheit, Basmah Safdar, Haikun Bao, Harlan M Krumholz, Nancy P Lorenze, Mitra Daneshvar, John A Spertus, and Gail D'Onofrio. Sex differences in the presentation and perception of symptoms among young patients with myocardial infarction: evidence from the virgo study (variation in recovery: role of gender on outcomes of young ami patients). *Circulation*, 137(8):781–790, 2018.

[176] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[177] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2015.

[178] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[179] Zhi Liu and Nikhil Garg. Equity in resident crowdsourcing: Measuring under-reporting without ground truth data. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 1016–1017, 2022.

[180] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[181] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

[182] Shruti Magesh, Daniel John, Wei Tse Li, Yuxiang Li, Aidan Mattingly-App, Sharad Jain, Eric Y Chang, and Weg M Ongkeko. Disparities in covid-19 outcomes by race, ethnicity, and socioeconomic status: a systematic-review and meta-analysis. *JAMA network open*, 4(11):e2134147–e2134147, 2021.

[183] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

[184] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, March 1947.

[185] Charles F Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

[186] Samuel A McDonald, Richard J Medford, Mujeeb A Basit, Deborah B Diercks, and D Mark Courtney. Derivation with internal validation of a multivariable predictive model to predict covid-19 test results in emergency department patients. *Academic Emergency Medicine*, 28(2):206–214, 2021.

[187] J. Michael McWilliams, Gregory Weinreb, Mary Beth Landrum, and Michael E. Chernew. Use of patient health survey data for risk adjustment to limit distortionary coding incentives in medicare. *Health Affairs*, 44(1):48–57, 2025.

[188] Medicare Payment Advisory Commission. Report to the congress: Medicare payment policy. Technical report, Medicare Payment Advisory Commission, March 2024. Released March 15, 2024; available online.

[189] MedPAC. Estimating medicare advantage coding intensity and favorable selection, 2024.

[190] MedPAC. The medicare advantage program: Status report, 2025.

[191] Aditya Krishna Menon, Brendan Van Rooyen, and Nagarajan Natarajan. Learning from binary labels with instance-dependent noise. *Machine Learning*, 107(8):1561–1595, 2018.

[192] Erik Miehling, Michael Desmond, Karthikeyan Natesan Ramamurthy, Elizabeth M. Daly, Pierre Dognin, Jesus Rios, Djallel Bouneffouf, and Miao Liu. Evaluating the prompt steerability of large language models, 2024.

[193] Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 230–239, 2019.

[194] Supriya Misra, Simona C Kwon, Ana F Abraído-Lanza, Perla Chebli, Chau Trinh-Shevrin, and Stella S Yi. Structural racism and immigrant health in the united states. *Health Education & Behavior*, 48(3):332–341, 2021.

[195] Rajiv Movva, Divya Shanmugam, Kaihua Hou, Priya Pathak, John Guttag, Nikhil Garg, and Emma Pierson. Coarse race data conceals disparities in clinical risk score performance. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, 2023.

[196] Stephanie A Mulherin and William C Miller. Spectrum bias or spectrum effect? subgroup variation in diagnostic test evaluation. *Annals of Internal Medicine*, 137(7):598–602, 2002.

[197] Sendhil Mullainathan and Ziad Obermeyer. Diagnosing physician error: A machine learning approach to low-value health care. *The Quarterly Journal of Economics*, 137(2):679–727, 2022.

[198] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2017.

[199] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi-Phuong-Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. SELF: Learning to Filter Noisy Labels with Self-Ensembling. In *8th International Conference on Learning Representations*, 2020.

[200] Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs. *arXiv preprint arXiv:2407.01082*, 2024.

[201] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

[202] Lucia D Notardonato, Spencer S Langerman, Jifang Zhou, Gregory S Calip, Brian C-H Chiu, and Ben A Derman. Racial Disparities in the Diagnostic Evaluation of Multiple Myeloma. *Blood*, 138(Supplement 1):4116–4116, November 2021.

[203] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[204] Jeeheh Oh, Maggie Makar, Christopher Fusco, Robert McCaffrey, Krishna Rao, Erin E Ryan, Laraine Washer, Lauren R West, Vincent B Young, John Guttag, et al. A generalizable, data-driven approach to predict daily risk of clostridium difficile infection at two large academic health centers. *infection control & hospital epidemiology*, 39(4):425–433, 2018.

[205] OpenAI. Learning to Reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/, 2024.

[206] OpenAI. Openai o3-mini system card. Technical report, OpenAI, February 2025. Accessed: 2025-04-18.

[207] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[208] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web, 1999.

[209] Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via Contrastive Activation Addition. *arXiv preprint arXiv:2312.06681*, 2024.

[210] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[211] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[212] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library, 2019.

[213] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017.

[214] Stephen G Pauker and Jerome P Kassirer. The threshold approach to clinical decision making. *New England Journal of Medicine*, 302(20):1109–1117, 1980.

[215] Nathaniel R. Payne and Susan E. Puumala. Racial Disparities in Ordering Laboratory and Radiology Tests for Pediatric Patients in the Emergency Department:. *Pediatric Emergency Care*, 29(5):598–606, May 2013.

[216] Judea Pearl. *Causality*. Cambridge university press, 2009.

[217] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, July 1900.

[218] 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.

[219] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. What you see is what you get? the impact of representation criteria on human bias in hiring. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 125–134, 2019.

[220] Emma Pierson, Sam Corbett-Davies, and Sharad Goel. Fast threshold tests for detecting discrimination. In *International Conference on Artificial Intelligence and Statistics*, pages 96–105, 2018.

[221] Emma Pierson, David M Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27(1):136–140, 2021.

[222] Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, et al. A large-scale analysis of racial disparities in police stops across the united states. *Nature Human Behaviour*, 4(7):736–745, 2020.

[223] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.

[224] 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.

[225] Eboni G Price-Haywood, Jeffrey Burton, Daniel Fort, and Leonardo Seoane. Hospitalization and mortality among black patients and white patients with Covid-19. *New England Journal of Medicine*, 382(26):2534–2543, 2020.

[226] Anya E. R. Prince and Daniel Schwarcz. Proxy Discrimination in the Age of Artificial Intelligence and Big Data. *Iowa Law Review*, 105(3):1257–1318, 2019.

[227] Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. Infobench: Evaluating instruction following ability in large language models, 2024.

[228] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[229] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[230] Yoshihiko Raita, Tadahiro Goto, Mohammad Kamal Faridi, David F. M. Brown, Carlos A. Camargo, and Kohei Hasegawa. Emergency department triage prediction of clinical outcomes using machine learning models. *Critical Care*, 23(1):64, December 2019.

[231] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, page 1–16. IEEE, November 2020.

[232] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 427–438, 2000.

[233] Ashesh Rambachan and Jonathan Roth. Bias in, bias out? Evaluating the folk wisdom. In *1st Symposium on Foundations of Responsible Computing*, 2020.

[234] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506, 2020.

[235] Matthew A Reyna, Chris Josef, Salman Seyedi, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Ashish Sharma, Shamim Nemati, and Gari D Clifford. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. In *2019 Computing in Cardiology (CinC)*, pages Page–1. IEEE, 2019.

[236] Chanu Rhee and Michael Klompas. Sepsis trends: increasing incidence and decreasing mortality, or changing denominator? *Journal of Thoracic Disease*, 12(Suppl 1):S89, 2020.

[237] Horst W. J. Rittel and Melvin M. Webber. Dilemmas in a general theory of planning. *Policy Sciences*, 4(2):155–169, 1973.

[238] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *9th International Conference on Learning Representations*, 2021.

[239] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, pages 846–866, 1994.

[240] R. Tyrrell Rockafellar. *Convex analysis*. Princeton University Press, Princeton, N. J., 1970.

[241] Casie C Rodenberger. Pre-submission risk adjustment audits: Preventing medicare advantage plans from draining medicare funds dry. *Iowa L. Rev.*, 103:841, 2017.

[242] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[243] Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. Stay on topic with classifier-free guidance. *arXiv preprint arXiv:2306.17806*, 2023.

[244] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. A human-centered review of algorithms used within the u.s. child welfare system. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–15, 2020.

[245] Tobias Schnabel and Jennifer Neville. Prompts as programs: A structure-aware approach to efficient compile-time prompt optimization. *arXiv e-prints*, pages arXiv–2404, 2024.

[246] Chet D. Schrader and Lawrence M. Lewis. Racial Disparity in Emergency Department Triage. *The Journal of Emergency Medicine*, 44(2):511–518, February 2013.

[247] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[248] Kevin A Schulman, Jesse A Berlin, William Harless, Jon F Kerner, Shyrl Sistrunk, Bernard J Gersh, Ross Dube, Christopher K Taleghani, Jennifer E Burke, Sankey Williams, et al. The effect of race and sex on physicians' recommendations for cardiac catheterization. *New England Journal of Medicine*, 340(8):618–626, 1999.

[249] Abigail See, Peter J Liu, and Christopher D Manning. Get to the Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083, 2017.

[250] Christopher W Seymour, Vincent X Liu, Theodore J Iwashyna, Frank M Brunkhorst, Thomas D Rea, André Scherag, Gordon Rubenfeld, Jeremy M Kahn, Manu Shankar-Hari, Mervyn Singer, et al. Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):762–774, 2016.

[251] Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, et al. SCROLLS: Standardized CompaRison Over Long Language Sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12007–12021, 2022.

[252] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[253] Divya Shanmugam, Kaihua Hou, and Emma Pierson. Quantifying disparities in intimate partner violence: a machine learning method to correct for underreporting. *npj Women's Health*, 2(1), 2024.

[254] Han Shao, Avrim Blum, and Omar Montasser. Strategic classification under unknown personalized manipulation. *Advances in Neural Information Processing Systems*, 36, 2024.

[255] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[256] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

[257] Maggie Shi. Monitoring for waste: Evidence from medicare audits. *The Quarterly Journal of Economics*, 139(2):993–1049, 2024.

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

[259] Elaine Silverman and Jonathan S Skinner. Are for-profit hospitals really different? medicare upcoding and market structure, 2001.

[260] Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *Jama*, 315(8):801–810, 2016.

[261] Gopal K Singh, Gem P Daus, Michelle Allender, Christine T Ramey, Elijah K Martin, Chrisp Perry, Andrew A De Los Reyes, and Ivy P Vedamuthu. Social determinants of health in the United States: addressing major health inequality trends for the nation, 1935-2016. *International Journal of MCH and AIDS*, 6(2):139, 2017.

[262] Y. Song, J. Skinner, J. Bynum, J. Sutherland, J. E. Wennberg, and E. S. Fisher. Regional variations in diagnostic practices. *New England Journal of Medicine*, 363(1):45–53, 2010.

[263] Kayte Spector-Bagdady, Shengpu Tang, Sarah Jabbour, W Nicholson Price, Ana Bracic, Melissa S Creary, Sachin Kheterpal, Chad M Brummett, and Jenna Wiens. Respecting autonomy and enabling diversity: The effect of eligibility and enrollment on research data demographics. *Health Affairs*, 40(12):1892–1899, 2021.

[264] Aude Sportisse, Hugo Schmutz, Olivier Humbert, Charles Bouveyron, and Pierre-Alexandre Mattei. Are labels informative in semi-supervised learning? estimating and leveraging the missing-data mechanism. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[265] Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. Does fair ranking improve minority outcomes? Understanding the interplay of human and algorithmic biases in online hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, page 989–999, 2021.

[266] Ravi Sundaram, Anil Vullikanti, Haifeng Xu, and Fan Yao. Pac-learning for strategic classification. *Journal of Machine Learning Research*, 24(192):1–38, 2023.

[267] Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adria Garriga-Alonso, and Robert Kirk. Analyzing the generalization and reliability of steering vectors, 2024.

[268] R. Andrew Taylor, Christopher L. Moore, Kei-Hoi Cheung, and Cynthia Brandt. Predicting urinary tract infections in the emergency department with machine learning. *PLOS ONE*, 13(3):e0194085, March 2018.

[269] R. Andrew Taylor, Joseph R. Pare, Arjun K. Venkatesh, Hani Mowafi, Edward R. Melnick, William Fleischman, and M. Kennedy Hall. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data–Driven, Machine Learning Approach. *Academic Emergency Medicine*, 23(3):269–278, March 2016.

[270] DeepSeek-AI Team. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025.

[271] Qwen Team. Qwen3, April 2025.

[272] Erin Teeple, Thomas Hartvigsen, Cansu Sen, Kajal T Claypool, and Elke A Rundensteiner. Clinical performance evaluation of a machine learning system for predicting hospital-acquired clostridium difficile infection. In *HEALTHINF*, pages 656–663, 2020.

[273] Sharon Tennyson and Pau Salsas-Forn. Claims auditing in automobile insurance: fraud detection and deterrence objectives. *Journal of Risk and Insurance*, 69(3):289–308, 2002.

[274] The pandas development team. pandas-dev/pandas: Pandas, 2020.

[275] Marian Tietz, Thomas J. Fan, Daniel Nouri, Benjamin Bossan, and skorch Developers. *skorch: A scikit-learn compatible neural network library that wraps PyTorch*, July 2017.

[276] Donna Tjandra and Jenna Wiens. Leveraging an alignment set in tackling instance-dependent label noise. In *Proceedings of the Conference on Health, Inference, and Learning*, 2023.

[277] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering Language Models With Activation Engineering. *arXiv preprint arXiv:2308.10248*, 2023.

[278] U.S. Department of Health and Human Services, Office of Inspector General. Medicare advantage compliance audit of specific diagnosis codes that independent health association, inc. (contract h3362) submitted to cms, June 2024.

[279] U.S. Department of Health and Human Services, Office of Inspector General. Medicare advantage compliance audit of specific diagnosis codes that ucare minnesota (contract h2459) submitted to cms, December 2024.

[280] Keyon Vafa, Sarah Bentley, Jon Kleinberg, and Sendhil Mullainathan. What's Producible May Not Be Reachable: Measuring the Steerability of Generative Models. *arXiv preprint arXiv:2503.17482*, 2025.

[281] Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

[282] Stijn Viaene, Richard A Derrig, Bart Baesens, and Guido Dedene. A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *Journal of Risk and Insurance*, 69(3):373–421, 2002.

[283] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[284] Robin Vogel, Aurélien Bellet, and Stephan Clémençon. Learning fair scoring functions: Bipartite ranking under ROC-based fairness constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 784–792, 2021.

[285] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

[286] Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms, 2020.

[287] Stefan Wager. Stats 361: Causal inference. Technical report, Stanford University, 2020.

[288] Yoav Wald and Suchi Saria. Birds of an odd feather: Guaranteed out-of-distribution (OOD) novel category detection. In *Uncertainty in Artificial Intelligence*, pages 2179–2191, 2023.

[289] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 526–536, 2021.

[290] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 526–536, 2021.

[291] Xutao Wang, Hanting Chen, Tianyu Guo, and Yunhe Wang. Pue: Biased positive-unlabeled learning enhancement by causal inference. *Advances in Neural Information Processing Systems*, 36, 2024.

[292] Christopher Weaver, Tom McGinty, Anna Wilde Matthews, and Mark Maremont. Insurers pocketed $50 billion from medicare for diseases no doctor treated. *Wall Street Journal*, 2024.

[293] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[294] H. G. Welch, S. M. Sharp, D. J. Gottlieb, J. S. Skinner, and J. E. Wennberg. Geographic variation in diagnosis frequency and risk of death among medicare beneficiaries. *JAMA*, 305(11):1113–1118, 2011.

[295] Jenna Wiens, Melissa Creary, and Michael W Sjoding. AI models in health care are not colour blind and we should not be either. *The Lancet Digital Health*, 4(6):e399–e400, June 2022.

[296] Erik Wijmans, Brody Huval, Alexander Hertzberg, Vladlen Koltun, and Philipp Krähenbühl. Cut your losses in large-vocabulary language models. *arXiv preprint arXiv:2411.09009*, 2024.

[297] Songhua Wu, Mingming Gong, Bo Han, Yang Liu, and Tongliang Liu. Fair classification with instance-dependent label noise. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 927–943, Apr 2022.

[298] Hongzuo Xu, Guansong Pang, Yijie Wang, and Yongjun Wang. Deep isolation forest for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[299] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A Survey on Deep Semi-Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954, September 2023.

[300] Yu Yao, Tongliang Liu, Mingming Gong, Bo Han, Gang Niu, and Kun Zhang. Instance-dependent label-noise learning under a structural causal model. In *Advances in Neural Information Processing Systems*, volume 34, pages 4409–4420, 2021.

[301] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[302] Dylan Zapzalka, Trenton Chang, Lindsay Warrenburg, Sae-Hwan Park, Daniel K. Shenfeld, Ravi B. Parikh, et al. Disentangling misreporting from genuine adaptation in strategic settings: A causal approach. In *Proceedings of the NeurIPS 2025 Workshop on Causal & Scientific Learning (CauScien)*, 2025.

[303] Hanrui Zhang and Vincent Conitzer. Incentive-aware pac learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):5797–5804, 2021.

[304] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. In *ICLR*, 2021.

[305] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 8792–8802, 2018.

[306] Ganlong Zhao, Guanbin Li, Yipeng Qin, Feng Liu, and Yizhou Yu. Centrality and consistency: Two-stage clean samples identification for learning with instance-dependent noisy labels. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 21–37, 2022.

[307] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT Interaction Logs in the Wild. *arXiv preprint arXiv:2405.01470*, 2024.

[308] Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7, 2019.

[309] Qihuang Zhong, Kang Wang, Ziyang Xu, Juhua Liu, Liang Ding, Bo Du, and Dacheng Tao. Achieving >97% on GSM8k: Deeply understanding the problems makes llms perfect reasoners. *arXiv preprint arXiv:2404.14963*, 2024.

[310] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models. *arXiv preprint arXiv:2311.07911*, 2023.

[311] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Center for Automated Learning and Discovery, Carnegie Mellon University, 2002.

[312] Zhaowei Zhu, Yiwen Song, and Yang Liu. Clusterability as an alternative to anchor points when learning with noisy labels. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12912–12923, 2021.