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

Trenton Chang CTRENTON@UMICH.EDU
Division of Computer Science and Engineering
University of Michigan
Ann Arbor, MI, USA

Michael W. Sjoding MSJODING@MED.UMICH.EDU
Department of Internal Medicine
University of Michigan
Ann Arbor, MI, USA

Jenna Wiens WIENSJ@UMICH.EDU
Division of Computer Science and Engineering
University of Michigan
Ann Arbor, MI, USA

Abstract

As machine learning (ML) models gain traction in clinical applications, understanding the impact of clinician and societal biases on ML models is increasingly important. While biases can arise in the labels used for model training, the many sources from which these biases arise are not yet well-studied. In this paper, we highlight *disparate censorship* (*i.e.,* differences in testing rates across patient groups) as a source of label bias that clinical ML models may amplify, potentially causing harm. Many patient risk-stratification models are trained using the results of clinician-ordered diagnostic and laboratory tests of labels. Patients without test results are often assigned a negative label, which assumes that untested patients do not experience the outcome. Since orders are affected by clinical and resource considerations, testing may not be uniform in patient populations, giving rise to disparate censorship. Disparate censorship in patients of equivalent risk leads to *undertesting* in certain groups, and in turn, more biased labels for such groups. Using such biased labels in standard ML pipelines could contribute to gaps in model performance across patient groups. Here, we theoretically and empirically characterize conditions in which disparate censorship or undertesting affect model performance across subgroups. Our findings call attention to disparate censorship as a source of label bias in clinical ML models.

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 (Buolamwini and Gebru, 2018; Obermeyer et al., 2019; Pierson et al., 2021). Biases in training data resulting from spurious correlations between the inputs and outputs have received much attention (*i.e.*, "shortcuts" Geirhos et al. (2020); Jabbour et al. (2020)). However, biases can also arise in the labels used for model training. Obermeyer et al. (2019) highlighted one such type of *label bias* in equating healthcare need

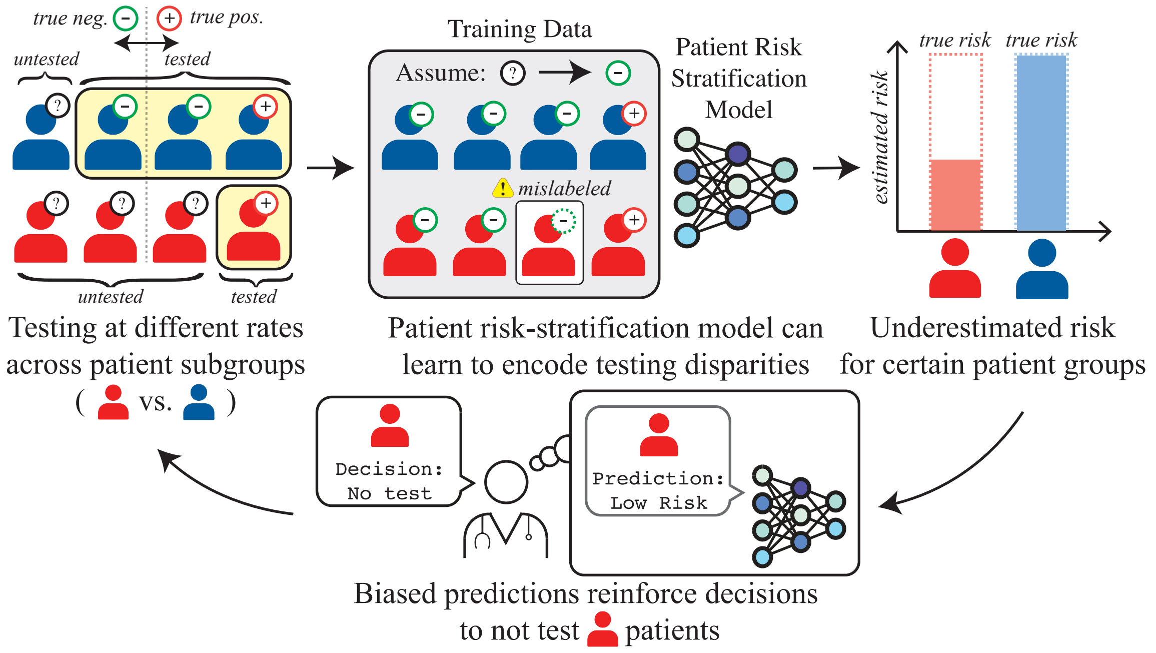
true neg.
true pos.
untested
tested
untested
tested
Testing at different rates across patient subgroups (vs.)
Training Data
Assume: ? → −
mislabeled
Patient Risk Stratification Model
Patient risk-stratification model can learn to encode testing disparities
estimated risk
true risk
true risk
Underestimated risk for certain patient groups
Decision: No test
Prediction: Low Risk
Biased predictions reinforce decisions to not test patients

Figure 1: Disparate censorship can cause a harmful feedback loop in clinical ML workflows. Clockwise from top left: **a)** disparate censorship occurs when a patient subgroup is tested for some condition at a lower rate compared to other groups. **b)** When untested patients are assumed negative, if patients of equivalent risk are subject to disparate censorship (*i.e.*, undertesting), standard ML training may learn to encode label bias (*i.e.*, missed positives by clinicians) from undertesting. **c)** Such models may underestimate risk for certain patients. **d)** Acting on such predictions perpetuates undertesting—further harming already-underserved populations.

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. (Rhee and Klompas, 2020; Seymour et al., 2016; Henry et al., 2019)), 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 (Bekker and Davis, 2020)). For example, many sepsis prediction models derive labels from laboratory test-based definitions, such that untested patients are negative by definition (Adams et al., 2022; Henry et al., 2015; Fleuren et al., 2020; Reyna et al., 2019). Beyond sepsis, this is also the case when building models to predict healthcare-associated infections (Oh et al., 2018; Teeple et al., 2020; Hartvigsen et al., 2018). 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 (Dolan et al., 2005), or biases in cardiac evaluations for women (Schulman et al., 1999). 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 (Schulman et al., 1999; Daugherty et al., 2017), different levels of healthcare access or consent (Spector-Bagdady et al., 2021), or different test performance across groups (Gaffin et al., 2010).

While the immediate harm of undertesting and missed diagnoses is clear (Magesh et al., 2021; Berry et al., 2009), 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 (Figure 1), in which models reinforce biased "do not test" decisions.

In this paper, 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 (Singh et al., 2017). Racial disparities in biomarkers indicating COVID-19 severity have also been documented (Price-Haywood et al., 2020). 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 (Lichtman et al., 2018).

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.

Generalizable Insights about Machine Learning in the Context of Healthcare

This paper 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 prove 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.

2. Problem Setup & Definitions

In this section, we propose a causal model of disparate censorship (Section 2.1), and outline three settings in which we study the impacts of disparate censorship (Section 2.2).

2.1. Causal Model of Disparate Censorship

We formalize disparate censorship using a causal directed acyclic graph (DAG; Figure 3). 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) encodes two additional assumptions. First, we assume clinician decisions to test, *i.e.*, T, depend on both X and A. 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) only, 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 $\tilde{y}_i$.

[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.

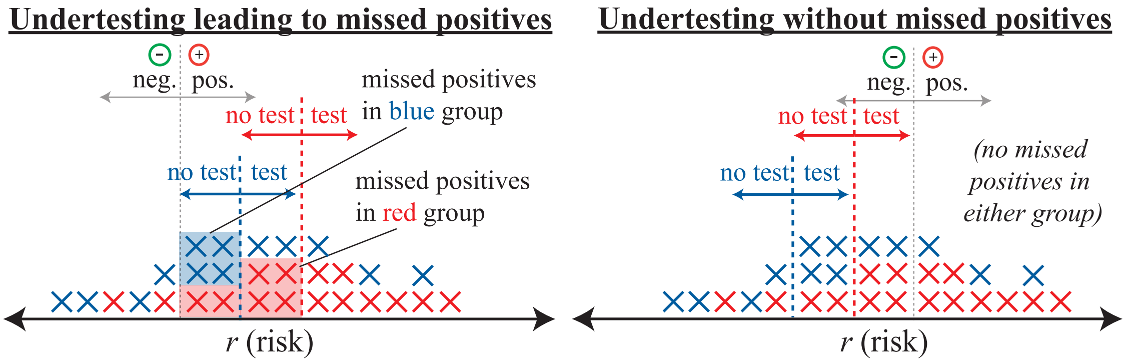

Figure 2: Stylized undertesting patterns. In both examples, the red group is undertested (higher risk r required for a test). Left: undertesting is harmful due to the disproportionate rates of missed diagnoses. Right: the absence of missed diagnoses means that undertesting has no disproportionate impact. In this paper, we focus on the scenario on the left.

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 the 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. To understand when testing disparities have the potential for harm, we focus on undertesting such that one group suffers disproportionately high rates of missed diagnoses/positives compared to the other group (Figure 2, left). Undertesting raises fewer concerns if it occurs in patients without the condition of interest, since undertesting such patients does not result in missed positives (Figure 2, 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 arguments are based on arguments about label noise. 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.

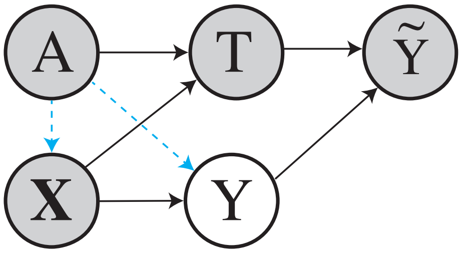

Figure 3: Causal DAG used in our analysis. (Un)shaded variables are (un)observed. Patient risk-stratification models aim to predict y given x, but only $\tilde{y}$ is observed. Disparate censorship arises when a affects t. Also, a may affect $\mathbf{x}$ and y (dashed arrows).

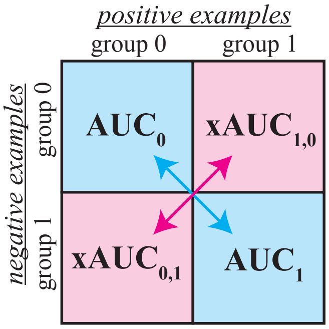

Figure 4: Decomposition of overall AUC into within-group AUCs (cyan) and xAUCs (magenta). Zero ranking performance gap between groups a requires parity for both within-group AUCs (cyan arrow) and xAUCs (magenta arrow).

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 Vyas et al. (2020); Ioannidis et al. (2021).[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 between patient subgroups. Examples of this setting arise in the study of social determinants of health (Singh et al., 2017), disparities in hypertension rates (Lackland, 2014), or multicenter datasets (Bhuva et al., 2019). In

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

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

addition, differences in healthcare utilization (alluded to in Price-Haywood et al. (2020)) and/or consent (Spector-Bagdady et al., 2021) 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 (Lichtman et al., 2018).

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. 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, Kallus and Zhou (2019)).

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, Kallus and Zhou (2019)).

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 4, cyan). The xAUC has a similar interpretation: $\text{xAUC}_{a,a'}$ is is the probability that a random positive patient in group a was ranked above a negative patient from group a' (Figure 4, 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 4, cyan arrow), while cross-group ranking performance should be symmetric (Figure 4, magenta arrow). This yields the following performance gap metrics—ΔAUC and ΔxAUC (lower is better):

$$\Delta\text{AUC} \triangleq |\text{AUC}_1 - \text{AUC}_0| \tag{1}$$

$$\Delta\text{xAUC} \triangleq |\text{xAUC}_{0,1} - \text{xAUC}_{1,0}| \tag{2}$$

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 C.

3.2. Impact of distributional differences on performance gaps

Here, we prove conditions that lead to zero $\Delta\text{AUC}, \Delta\text{xAUC}$ in the presence of disparate censorship and/or undertesting for the three distributional settings studied.

3.2.1. 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, $\Delta\text{AUC}, \Delta\text{xAUC}$ converge to zero.

3.2.2. 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 (Pauker and Kassirer, 1980)). 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 5: note that the censorship boundaries are "parallel" to the true decision boundary. Concretely, we assume that there exist $\tau_a \in \mathbb{R}$ for $a \in \{0, 1\}$, which act as censorship thresholds, and $b \in \mathbb{R}$, which acts as a decision boundary, such that $t = \mathbb{1}[s(\mathbf{x}) > \tau_a \vee p = 1]$, where $p \sim \mathit{Bernoulli}(c)$ for a suitably small $c \in (0, 1]$ for each group a and $y = \mathbb{1}[s(\mathbf{x}) > b]$.[6] 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 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 $z \mid a = 0 \sim \mathcal{N}(\mu_0, \sigma^2), z \mid a = 1 \sim \mathcal{N}(\mu_1, \sigma^2)$.*

[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.

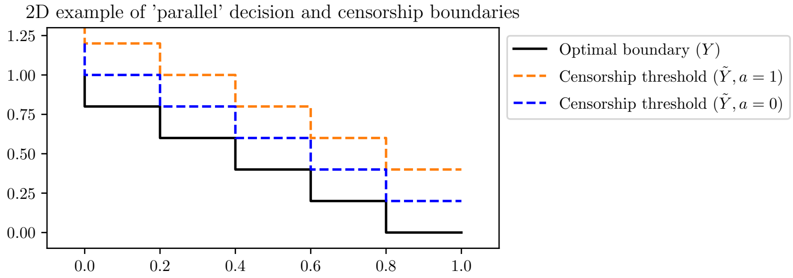

Figure 5: Example decision (black) and censorship boundaries (blue, orange) in 2D. In our noise model, we assume that the clinician "test decision" (t) and true "condition status" (y) decision boundaries are based on a threshold with the same functional form (*i.e.* "parallel").

Without loss of generality, suppose that $\mu_1 \geq \mu_0$. Then if $\tau_1 \leq \tau_0$ or no positives are missed, the $\Delta AUC, \Delta xAUC$ 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 raise no concerns. 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" (Menon et al., 2018). 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.

3.2.3. 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 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 $\mathbf{x}$ *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"

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

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.

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.

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).

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 6 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

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

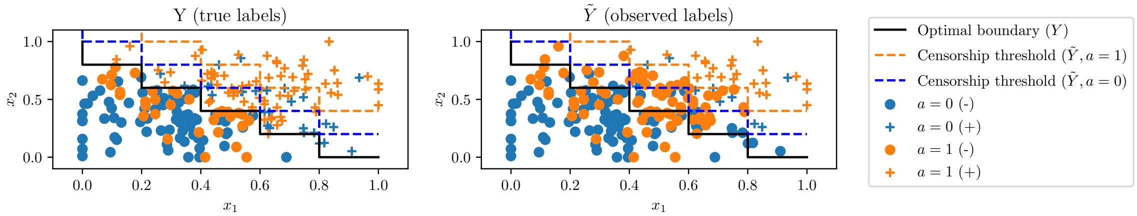

Figure 6: Example of disparate censorship in 2D with groups $a = 0$ (in blue) and $a = 1$ (in orange), and decision boundary in black. All blue positives ($a = 0$, $+$) below the blue dashed line are censored; likewise for orange positives ($a = 1$, $+$).

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 B.

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.

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 $\phi \in [0, 360)$.

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 D.

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 5.1)

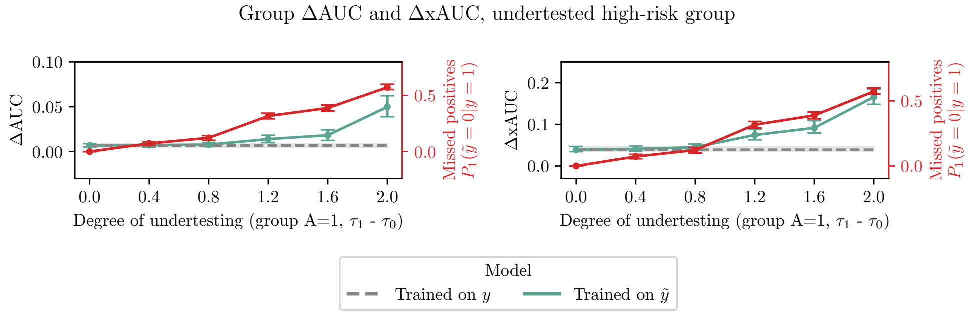

Figure 7: ΔAUC (left) and ΔxAUC (right) with 95% empirical CIs for the model trained on y (gray) and the model trained on $\tilde{y}$ (green). As the degree of undertesting experienced by the high-risk group $a = 1$ $(\tau_1 - \tau_0)$ increases, ΔAUC, ΔxAUC (green) as the missed positive rate (red) sharply rises in group $a = 1$ (note that $P_0(\tilde{y} = 0 \mid y = 1) = 0$—no missed positives in group $a = 0$).

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

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.*, 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, so does the rate of missed positives in group $a = 1$, and performance gaps grow as expected. Consistent with Theorem 3, ΔAUC, ΔxAUC increase when the high-risk group is undertested (Figure 7).We also examine increasing τ_0

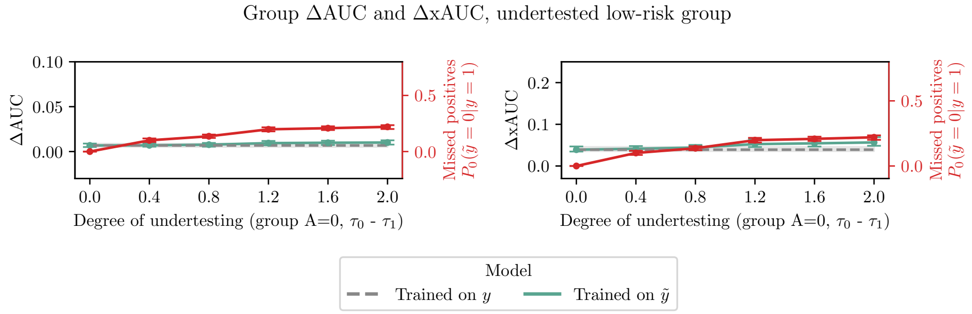

Figure 8: ΔAUC (left) and ΔxAUC (right) with 95% empirical CIs for the model trained on y (gray) and the model trained on $\tilde{y}$ (green). Even as the degree of undertesting experienced by the low-risk group $a = 0$ $(\tau_0 - \tau_1)$ increases, all metrics stay near oracle performance while the missed positive rate plateaus (red) in group $a = 0$ (note that $P_1(\tilde{y} = 0 \mid y = 1) = 0$—there are no missed positives in group $a = 1$).

from 5, inducing missed positives in both groups; the correlation between the performance gap and the value of $\tau_1 - \tau_0$ still holds (Appendix E).

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, limiting the amount of harm that undertesting the low-risk group can cause. Thus, when the low-risk group is undertested, training with labels $\tilde{y}$ versus y does not significantly affect ΔAUC and ΔxAUC, even as the missed positive rate rises slightly. Figure 8 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 $\Delta\text{AUC}, \Delta\text{xAUC}$) 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 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 E).

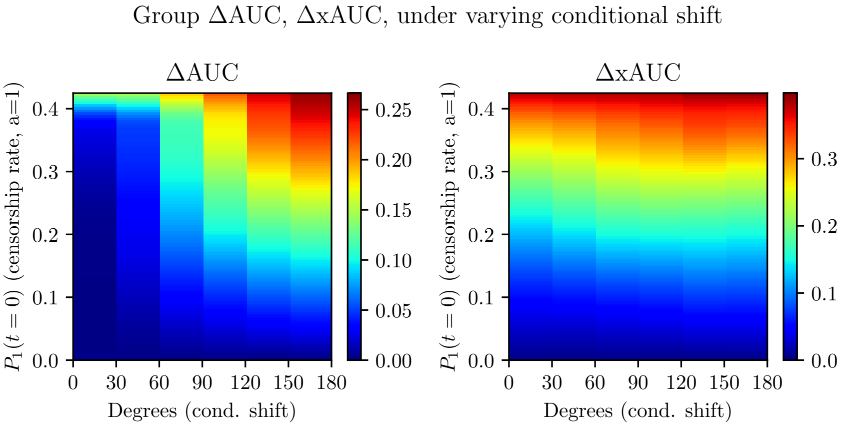

Figure 9: Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (ϕ; x-axis) and censorship rate in $a = 1$ ($P_1(t = 0)$, y-axis); 4 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger gaps are in dark red. Performance gaps widen as conditional shift or disparate censorship intensify.

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 1 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 E). We plot median $\Delta\text{AUC}, \Delta\text{xAUC}$ 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 $\Delta\text{AUC}, \Delta\text{xAUC}$ (Figure 9). 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 $\Delta\text{AUC}, \Delta\text{xAUC}$ 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.

6. Practical Concerns & Guidance for Addressing Disparate Censorship

Our empirical results demonstrate that, in some settings, disparate censorship may lead to performance gaps. Here, we show the extent to which disparate censorship occurs in common laboratory/diagnostic tests in MIMIC-IV, a popular dataset used in ML for healthcare, and suggest ways to address disparate censorship.

Disparate censorship in the MIMIC-IV dataset. We validate the existence of disparate censorship in laboratory/diagnostic tests in the MIMIC-IV dataset (Johnson et al.,

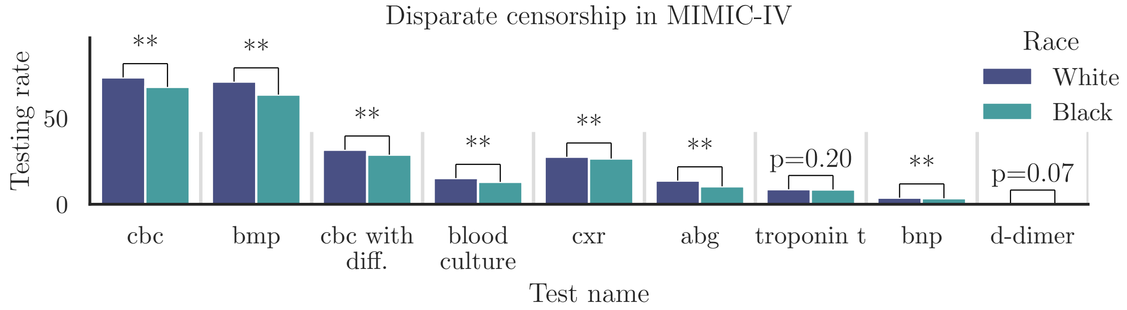

Figure 10: Disparate censorship in terms of the testing rate disparity ($P(T = 1 \mid \text{admission for White patient}) - P(T = 0 \mid \text{admission for Black patient})$) in common laboratory tests in MIMIC-IV. Testing rates plotted by race (White = indigo, Black = teal). "**" denotes a statistically significant difference ($\alpha = 1.1 \times 10^{-3}$ post-Bonferroni); p-value noted otherwise.

2020) (version 1.0), a widely-used clinical dataset consisting of electronic health record data from hospital admissions to emergency/intensive care units at the Beth Israel Deaconess Medical Center (BIDMC). We limit our analysis to "White" and "Black/African-American" (category names from MIMIC-IV) patients, since they constitute the two most prevalent racial groups in the dataset. This yields a sample of 337630 admissions corresponding to White patients and 80293 admissions corresponding to "Black/African-American" patients (referred to here as "Black" patients).

We test for disparate censorship via two-sample z-tests (1% significance threshold with Bonferroni correction; $\alpha = 1.1 \times 10^{-3}$) in standard laboratory/diagnostic tests such as complete blood counts (CBC, with and without differential), base metabolic panels (BMP), blood cultures, chest X-ray orders (CXR), arterial blood gas tests (ABG), troponin T tests, brain natriuretic peptide (BNP) tests, and d-dimer tests. These tests are chosen since they may be used to help diagnose certain conditions (*e.g.*, anemia or infection from CBCs, kidney injury from BMPs) or feature directly in clinical definitions (*e.g.*, blood culture orders and sepsis), which may impact label definitions for training downstream ML models. Figure 10 shows that statistically significant disparate censorship occurs in CBCs (with and without diff.), BMPs, blood cultures, CXRs, ABGs, and BNP tests.

Importantly, Black patients are significantly less likely to be tested in instances of disparate censorship. Even though these results do not definitively prove undertesting, the consistently lower testing rates in Black patients raise concerns about whether ML models trained on such data could encode such testing disparities, leading to the negative impacts we highlight in our study. This is particularly concerning due to the wide usage of MIMIC-IV in clinical ML. More detailed results can be found in Appendix E.

Addressing disparate censorship and undertesting. Recall that, if the marginal risk distribution differs (Condition 1, Figure 11) and the high-risk group is undertested leading to missed positives (Condition 2, Figure 11), performance gaps may emerge. If the conditional risk distribution differs instead (Condition 3, Figure 11) and the decision and censorship boundaries are non-parallel (Condition 4, Figure 11), model performance gaps may also emerge. Identifying settings in which disparate censorship and undertesting can have adverse effects could inform interventions in health policy, clinical care delivery, or even computational solutions. Here, we suggest potential methods for detecting and mitigating disparate censorship and undertesting.

Via our causal model (Figure 3), in the presence of disparate censorship (*i.e.*, $A \to T$), removing the dependence between A and X in Setting 2 (*i.e.*, $P_0(\mathbf{x}) \overset{d}{\neq} P_1(\mathbf{x})$) and/or the dependence between A and Y in Setting 3 (*i.e.*, $P_0(y \mid \mathbf{x}) \overset{d}{\neq} P_1(y \mid \mathbf{x})$) would transform instances of Settings 2/3 into instances of Setting 1, making the patient groups A indistinguishable. These approaches could potentially mitigate performance gaps, since a model trained on such data would behave identically across groups. Identifying whether differences in the marginal and conditional risk distributions (Settings 2 and 3, respectively) can provide further information for model design decisions.

To check whether $P_0(\mathbf{x}) \overset{d}{=} P_1(\mathbf{x})$ (Condition 1 of Figure 11/Setting 2), standard hypothesis tests for distributional equality such as Kolmogorov-Smirnov (Massey Jr, 1951) can be used for each covariate. Non-parametric distributional distances (*e.g.*, 2-Wasserstein, MMD (Gretton et al., 2012)) could help quantify to what extent $P_0(\mathbf{x}) \overset{d}{=} P_1(\mathbf{x})$ holds. For distinguishing the high-risk group in particular, beyond hypothesis testing, incorporating domain knowledge on health disparities in the relevant covariates may be necessary. To determine if the high-risk group is undertested (and therefore, performance gaps may arise), one may estimate τ_a (Condition 2, Figure 11) via threshold tests (Simoiu et al., 2017; Pierson et al., 2018; Patel et al., 2021). For mitigation, the condition $P_0(\mathbf{x}) \overset{d}{=} P_1(\mathbf{x})$ is reminiscent of covariate shift in domain adaptation. Hence, standard approaches (*e.g.*, reweighing methods (Jiang, 2008), optimal transport (Courty et al., 2017), feature augmentation (Daumé III, 2009)) may apply.

Checking whether $P_0(y \mid \mathbf{x}) \overset{d}{=} P_1(y \mid \mathbf{x})$ holds (Condition 3 of Figure 11/Setting 3) is less straightforward, as we observe y with potentially varying noise rates in each group. Metrics of conditional distributional similarity (*i.e.*, Bregman correntropy (Yu et al., 2020)) have been proposed, but determining the presence of undertesting remains an open question for Setting 3, as the covariates most predictive of y may vary by group. Positive-unlabeled (PU) learning is a promising direction here, but instance-dependent/group-wise PU learning remains underexplored. We highlight Gong et al. (2021) as a potential approach. Lastly, resolving the "parallel boundaries" assumption (Condition 4, Figure 11) requires modeling both boundaries, so the same limitations for checking Condition 3 apply. Using domain knowledge may be most practical for verifying Conditions 3 and 4.

7. Discussion

We investigate the impact of disparate censorship and undertesting across different settings. Recall that disparate censorship (or lack thereof) can lead to *undertesting* if differences in testing lead to disproportionately frequent missed positives in certain groups. We theoretically show that when the marginal and conditional distributions of covariates are the same across groups, undertesting raises no concerns. However, when the marginal risk distribution differs, performance gaps may arise if certain patient subgroups are undertested such

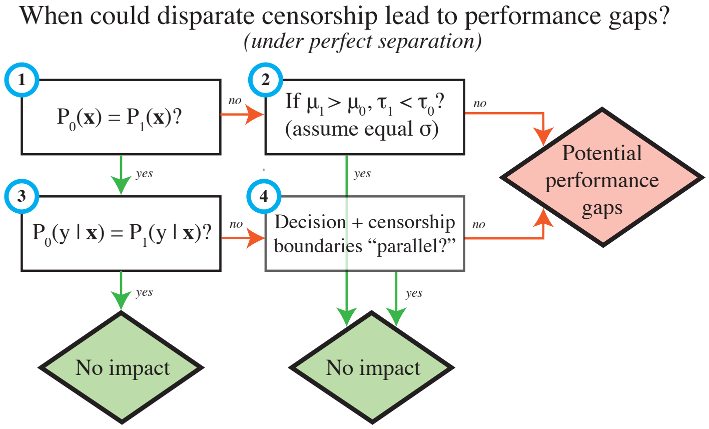
When could disparate censorship lead to performance gaps?
(under perfect separation)
1
$P_0(\mathbf{x}) = P_1(\mathbf{x})$?
no
2
If $\mu_1 > \mu_0$, $\tau_1 < \tau_0$?
(assume equal σ)
no
Potential performance gaps
yes
yes
3
$P_0(y \mid \mathbf{x}) = P_1(y \mid \mathbf{x})$?
no
4
Decision + censorship boundaries "parallel?"
no
yes
yes
No impact
No impact

Figure 11: 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 ($\Delta\text{AUC}, \Delta\text{xAUC} > 0$).

that one group experiences a disproportionate amount of missed positives. To better understand when disparate censorship may lead to model performance gaps due to undertesting, we conduct a simulation study. We find that when either the marginal or conditional risk distributions differ, undertesting can result in performance gaps. We then identify 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. Our findings encourage viligance to harmful undertesting—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: Obermeyer et al. (2019); Pierson et al. (2021) 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 mitigates some of the harm. Such solutions may be less applicable to *disparate censorship*: here, the labels correspond to the outcome of interest, but are (partially) observed at different rates for each group.

Addressing adverse effects of disparate censorship and undertesting requires thoroughly understanding one's problem setting. Disparate censorship and undertesting raise few concerns 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 (Jiang and Nachum, 2020; Cheng et al., 2020; Berthon et al., 2021; Wang et al., 2021). Beyond ML methods, modeling techniques in the presence of censored/missing data include inverse probability weighting-based methods in the epidemiology/causal inference literature (Hernán et al., 2004) or the Heckman correction in the quantitative social sciences (Heckman, 1976).

Beyond 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, such 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 Latinx patients (Brondolo et al., 2009), immigrant communities (Misra et al., 2021), and Black gender minorities intersectionally (Lett et al., 2020). 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 Dolan et al. (2005); Schulman et al. (1999)). Thus, while we expect our assumptions to partially hold, it is unclear to what extent they hold in practice.

Nevertheless, this paper provides a foundation for a deeper exploration of the impacts of disparate censorship and undertesting. We suggest a plausible mechanism of dataset bias and conditions under which gaps in model performance (ΔAUC, ΔxAUC) are likely to arise. Our theoretical results highlight when zero performance gap is feasible, and also suggest when nonzero performance gaps may occur. Our simulation study supports these theoretical insights. We further identify disparate censorship in the form of disproportionately low rates of laboratory/diagnostic test orders for Black patients in MIMIC-IV, a concerning finding due to the wide usage of MIMIC-IV in clinical ML. 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 negative impacts of disparate censorship.

Acknowledgments

We thank (in alphabetical order) Donna Tjandra, Erkin Ötleş, Fahad Kamran, Jiaxuan Wang, Jung Min Lee, Meera Krishnamoorthy, Sarah Jabbour, Shengpu Tang, and Stephanie Shepard for helpful discussions. We also thank Harry Rubin-Falcone, Maggie Makar, Michael Ito for their feedback in preparing the paper presentation. Special thanks to Sarah Jabbour for assisting with chest X-ray data processing. This work was supported by the Agency of Health Research and Quality (grant no. 5R01HS027431-03) and the National Heart Lung and Blood Institute (grant no. 1R01HL158626-01). The views and conclusions in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Agency of Health Research and Quality or the National Heart Lung and Blood Institute.

References

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.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020.

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.

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.

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.

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.

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.

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.

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.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*, 30, 2017.

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.

Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

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.

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.

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.

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.

Chen Gong, Qizhou Wang, Tongliang Liu, Bo Han, Jane J You, Jian Yang, and Dacheng Tao. Instance-dependent positive and unlabeled learning with labeling bias estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

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.

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.

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.

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.

Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, pages 615–625, 2004.

John PA Ioannidis, Neil R Powe, and Clyde Yancy. Recalibrating the use of race in medical research. *Jama*, 325(7):623–624, 2021.

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.

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.

Jing Jiang. A literature survey on domain adaptation of statistical classifiers. *URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey*, 3(1-12):3, 2008.

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.

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.

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.

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.

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.

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.

Frank J Massey Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

Aditya Krishna Menon, Brendan Van Rooyen, and Nagarajan Natarajan. Learning from binary labels with instance-dependent noise. *Machine Learning*, 107(8):1561–1595, 2018.

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.

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.

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.

Birju S Patel, Ethan Steinberg, Stephen R Pfohl, and Nigam H Shah. Learning decision thresholds for risk stratification models from aggregate clinician behavior. *Journal of the American Medical Informatics Association*, 28(10):2258–2264, 2021.

Stephen G Pauker and Jerome P Kassirer. The threshold approach to clinical decision making. *New England Journal of Medicine*, 302(20):1109–1117, 1980.

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. PMLR, 2018.

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.

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.

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.

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.

Chanu Rhee and Michael Klompas. Sepsis trends: increasing incidence and decreasing mortality, or changing denominator? *Journal of Thoracic Disease*, 12(Suppl 1):S89, 2020.

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.

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.

Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

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.

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.

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.

Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms, 2020.

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.

Shujian Yu, Ammar Shaker, Francesco Alesiani, and Jose C Principe. Measuring the discrepancy between conditional distributions: Methods, properties and applications. *arXiv preprint arXiv:2005.02196*, 2020.

Appendix A. Proofs

A.1. Preliminaries

First, we restate the definition of a boundary-consistent noise (BCN) model from Menon et al. (2018), which is useful for our proofs.

Definition 5 (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 Menon et al. (2018) with some notation adapted to our setting:

Theorem 6 (Theorem 2 of Menon et al. (2018).) *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,*

$$reg_{rank}(s; D) \leq C \cdot \frac{\bar{\pi}(1 - \bar{\pi})}{\pi(1 - \pi)} \cdot reg_{rank}(s; \bar{D}) \tag{3}$$

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 Menon et al. (2018) 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 7 (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. 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

$$f_1(z) = \begin{cases} 1 - c & z < \tau_1 \\ \frac{p_a \cdot \exp\left(-\frac{(z-\mu_0)^2}{2\sigma^2}\right)}{p_a \cdot \exp\left(-\frac{(z-\mu_0)^2}{2\sigma^2}\right) + (1-p_a)\cdot(1-c)\cdot\exp\left(-\frac{(z-\mu_1)^2}{2\sigma^2}\right)} & \tau_1 \leq z < \tau_0 \\ 0 & z \geq \tau_0 \end{cases}. \tag{4}$$

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 $a(z)$.[9] We rewrite the portion of f_1 for $\tau_1 \leq z < \tau_0$ as a sigmoid function:

$$a(z) = \frac{p_a \cdot \exp\left(-\frac{(z-\mu_0)^2}{2\sigma^2}\right)}{p_a \cdot \exp\left(-\frac{(z-\mu_0)^2}{2\sigma^2}\right) + (1 - p_a) \cdot (1 - c) \cdot \exp\left(-\frac{(z-\mu_1)^2}{2\sigma^2}\right)} \tag{5}$$

$$= \frac{1}{1 + \exp\left(\log\frac{(1-p_a)(1-c)}{p_a} + \frac{(z-\mu_0)^2 - (z-\mu_1)^2}{2\sigma^2}\right)} \tag{6}$$

$$= \frac{1}{1 + \exp\left(\log\frac{(1-p_a)(1-c)}{p_a} + \frac{1}{2\sigma^2}\left[(2z - \mu_0 - \mu_1)(\mu_1 - \mu_0)\right]\right)}. \tag{7}$$

We further rewrite Eq. 7 in the form $\sigma(g(z))$, where

$$g(z) = \log\frac{p_a}{(1 - p_a)(1 - c)} + \frac{(2z - \mu_0 - \mu_1)(\mu_0 - \mu_1)}{2\sigma^2}. \tag{8}$$

[9] 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.

Then, taking derivatives:

$$\frac{d}{dz}a(z) = \sigma(g(z))(1-\sigma(g(z)))\frac{d}{dz}g(z) = \sigma(g(z))(1-\sigma(g(z)))\left((\mu_0-\mu_1)/2\sigma^2\right) \leq 0, \tag{9}$$

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

$$\tau_1 \leq \log\frac{(1-c)^2(1-p_a)}{c \cdot p_a} \cdot \frac{2\sigma^2}{\mu_0-\mu_1} + \frac{\mu_0+\mu_1}{2}, \tag{10}$$

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 8 *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 9 *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 $[\tau_1, \tau_0)$. 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.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

$$f_1(\cdot) = \begin{cases} c & \boldsymbol{\theta}^\top\mathbf{x} + \beta \leq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{11}$$

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 $0 < \boldsymbol{\theta}_a^\top \mathbf{x} + b_a < \boldsymbol{\theta}_a^\top \mathbf{x}' + b_a$.[10] Then:

$$\begin{aligned}
\boldsymbol{\theta}_a^\top (\mathbf{x} - \mathbf{x}') > 0 &\iff \delta \boldsymbol{\theta}_a^\top (\mathbf{x} - \mathbf{x}') > 0 && (12)\\
&\iff \boldsymbol{\theta}^\top (\mathbf{x} - \mathbf{x}') > 0 && (13)\\
&\iff \boldsymbol{\theta}^\top \mathbf{x} + \beta > \boldsymbol{\theta}^\top \mathbf{x}' + \beta. && (14)
\end{aligned}$$

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 6, Lemma 7 concludes the proof. ■

Remark 10 *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 11 *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. 14, 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.*

Appendix B. 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:

$$\begin{aligned}
&\mathbf{x} \sim \max(0, \min(1, \mathcal{N}(\mu_a \mathbf{1}, \sigma^2 \mathbf{I}))) && (15)\\
&y = \mathbb{1}\left[s_a(\mathbf{x}) > 5\right] && (16)\\
&t \sim \max\left(\mathbb{1}\left[s_a(\mathbf{x}) > \tau_a \vee p = 1\right]\right), \quad p \sim \mathit{Bernoulli}(c) && (17)\\
&\tilde{y} = y \cdot t && (18)
\end{aligned}$$

[10] 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 $\theta_a^\top \mathbf{x} + b_a > 0$.

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	$\mu_0 = \mu_1 = 0.45$	$s_0(\mathbf{x}) = s(\mathbf{x}), s_1(\mathbf{x}) = s(\mathit{Rot}(\mathbf{x}; \phi, d', \mathbf{x}_0))$

Table 1: 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. 19.

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 15). 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 6 for a 2D example):

$$s_0(\mathbf{x}) = s_1(\mathbf{x}) = \frac{1}{5}\left(\sum_{i=1}^{10} \lceil 5x_i \rceil\right). \tag{19}$$

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. 16). 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. 17). If the risk for a patient in group a lies above τ_a, they 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. 18). 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 1. We define *Rot* as

$$\mathit{Rot}(\mathbf{x}; \phi, d', \mathbf{x}_0) = \begin{pmatrix} R(-\phi) & & d'/2 \text{ times} & \\ \vdots & \ddots & & \\ 0 & \dots & R(-\phi) & 0 \\ 0 & \dots & 0 & \mathbf{I}_{10-d'} \end{pmatrix} (\mathbf{x} - \mathbf{x}_0) + \mathbf{x}_0, \tag{20}$$

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.

Appendix C. 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 4, cyan). The xAUC has a similar interpretation: $\mathrm{xAUC}_{a,a'}$ is the probability that a randomly chosen positive example from group a is scored above a randomly chosen negative example from group a' (Figure 4, magenta). We defer to Kallus and Zhou (2019) for details on xAUC.

This yields a decomposition of the overall AUC in terms of within-group AUC and xAUC (Figure 4). 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:

$$\begin{aligned}\text{Overall AUC} = {} & p_0(0) \cdot p_1(0) \cdot \mathrm{AUC}_0 + p_0(1) \cdot p_1(1) \cdot \mathrm{AUC}_1 \\ & + p_0(0) \cdot p_1(1) \cdot \mathrm{xAUC}_{1,0} + p_0(1) \cdot p_1(0) \cdot \mathrm{xAUC}_{0,1}.\end{aligned} \tag{21}$$

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. 21 to hold.

Appendix D. Model Details

We provide all settings used for model training here.

D.1. Model Training

We train probabilistic kernel support vector machines (SVMs) (Platt et al., 1999), 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.

D.2. Model Hyperparameters

As the focus of this paper is on evaluation, we keep all default parameters for the SVM; that is, regularization weight $C = 1$ and $\gamma = (d \cdot Var(\mathrm{vec}(\mathbf{x})))^{-1}$ 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`).

D.3. Software

We use `scikit-learn` for the SVM implementation, which is built on LIBSVM (Chang and Lin, 2011).

Appendix E. 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.

E.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.[11] 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.

[11] 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 2: 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.

E.2. Setting 2: Difference in marginal risk distributions

We provide full results with empirical 95% confidence intervals for ΔAUC (Table 3) and ΔxAUC (Table 4) 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.

E.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 12
- $d' = 4$: Figure 13
- $d' = 6$: Figure 14
- $d' = 8$: Figure 15
- $d' = 10$: Figure 16

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

- $d' = 2$: AUC (Table 5), xAUC (Table 6)
- $d' = 4$: AUC (Table 7), xAUC (Table 8)
- $d' = 6$: AUC (Table 9), xAUC (Table 10)
- $d' = 8$: AUC (Table 11), xAUC (Table 12)

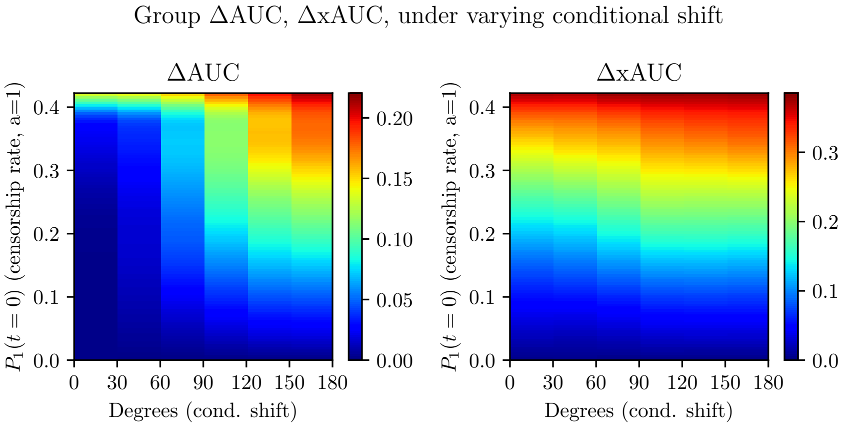

Figure 12: Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (x-axis) and censorship rate in $a = 1$ (y-axis); 2 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger performance gaps are in dark red.

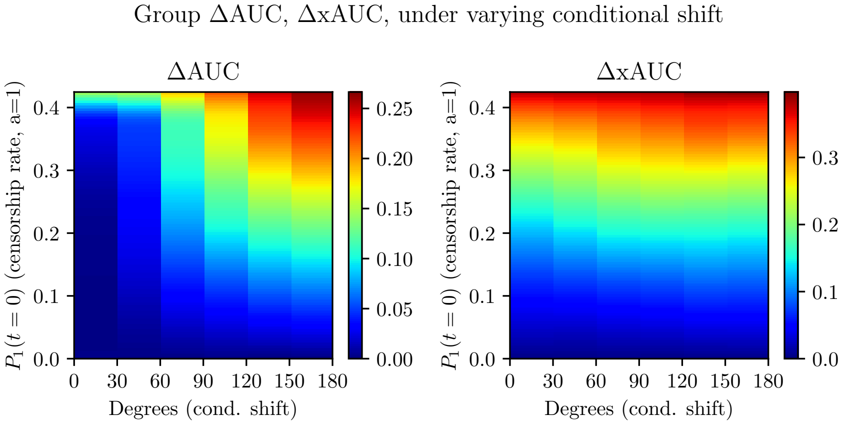

Figure 13: Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (ϕ; x-axis) and censorship rate in $a = 1$ ($P_1(t = 0)$, y-axis); 4 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger performance gaps are in dark red.

- $d' = 10$: AUC (Table 13), xAUC (Table 14)

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.

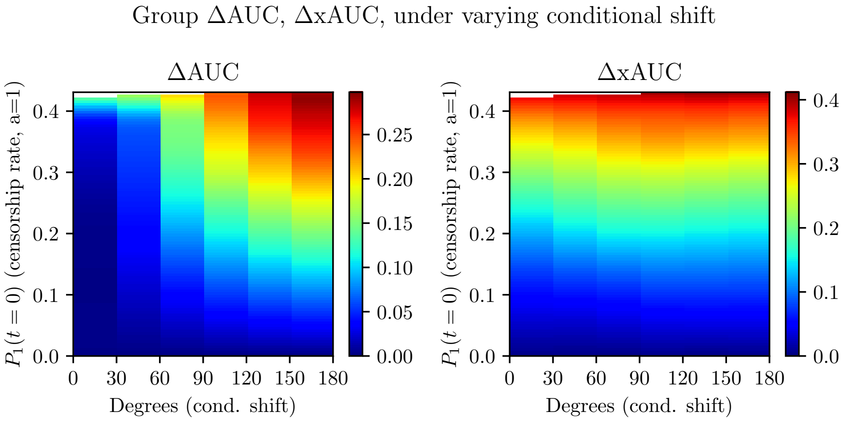

Figure 14: Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (ϕ; x-axis) and censorship rate in $a = 1$ ($P_1(t = 0)$, y-axis); 6 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger performance gaps are in dark red.

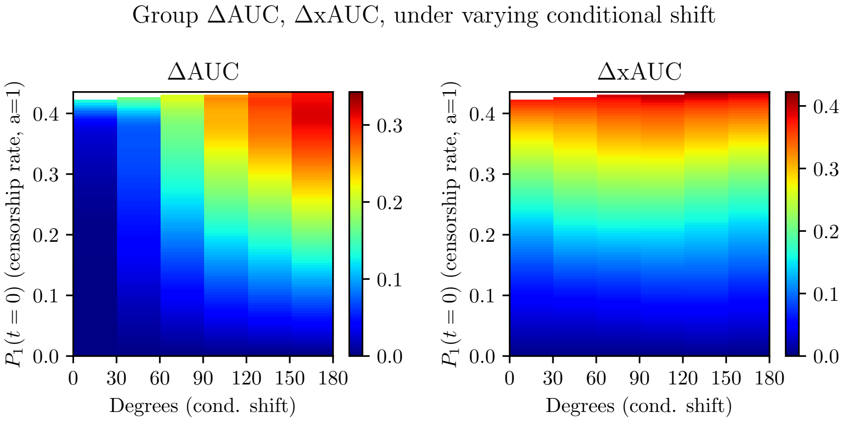

Figure 15: Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (ϕ; x-axis) and censorship rate in $a = 1$ ($P_1(t = 0)$, y-axis); 8 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger performance gaps are in dark red.

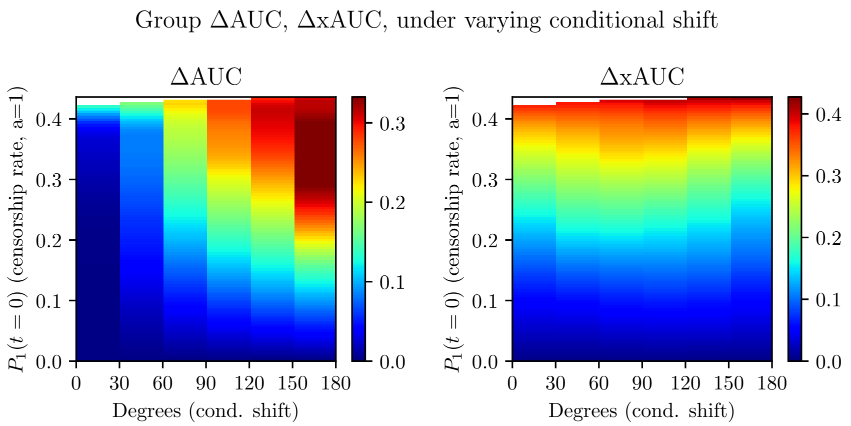

Figure 16: Heatmap showing median ΔAUC (left) and ΔxAUC (right) at varying levels of conditional shift (ϕ; x-axis) and censorship rate in $a = 1$ ($P_1(t = 0)$, y-axis); 10 dimensions rotated. Regions with smaller performance gap are in dark blue, while larger performance gaps are in dark red.

E.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.

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 17, 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 18 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.

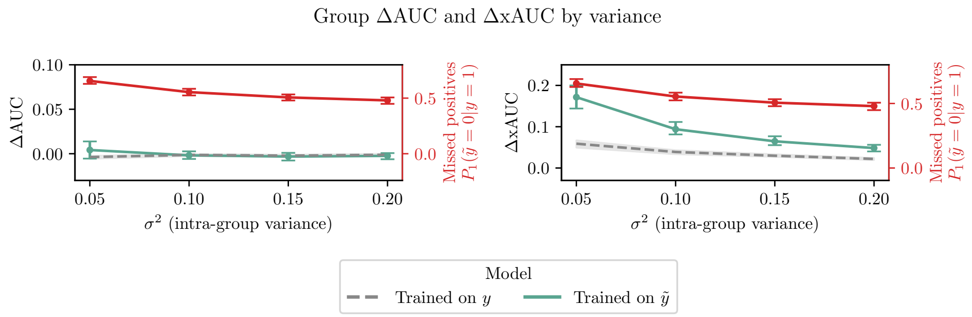

Figure 17: ΔAUC (left) and ΔxAUC (right) with 95% empirical CIs for the model trained on y (gray) and on $\tilde{y}$ (green). As σ^2 increases to the right, increasing overlap between the distributions, the number of missed positives in group $a = 1$ increases, and the AUC and xAUC gaps *narrow*. Note that $P_0(\tilde{y} = 0 \mid y = 1) = 0$—there are no missed positives in group $a = 0$.

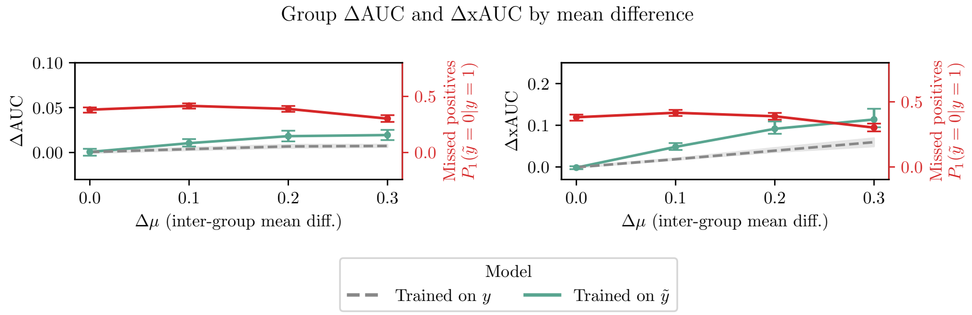

Figure 18: ΔAUC (left) and ΔxAUC (right) with 95% empirical CIs for the model trained on y (gray) and on $\tilde{y}$ (green). As $\Delta\mu$ increases to the right, the AUC and xAUC gaps *widen*. Note that $P_0(\tilde{y} = 0 \mid y = 1) = 0$—there are no missed positives in group $a = 0$. The number of missed positives in group $a = 1$ (red) first increases, then decreases, since sufficiently high $\Delta\mu$ means that untested patients' risk largely lies in one tail of the risk distribution.

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 15 (differences in $\Delta\mu$) and Table 16 (differences in σ^2).

Appendix F. 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 3: 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 4: 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 5: 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 6: 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 7: 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 8: 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 9: 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 10: 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 11: 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 12: 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 13: 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 14: 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 15: 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 16: AUC, xAUC with empirical 95% CIs at varying settings of σ^2, $\tau_0 = 5, \tau_1 = 6.6$.