Fairness Evaluation in Presence of Biased Noisy Labels

Riccardo Fogliato
Carnegie Mellon University
Partnership on AI

Max G'Sell
Carnegie Mellon University

Alexandra Chouldechova
Carnegie Mellon University
Partnership on AI

Abstract

Risk assessment tools are widely used around the country to inform decision making within the criminal justice system. Recently, considerable attention has been devoted to the question of whether such tools may suffer from racial bias. In this type of assessment, a fundamental issue is that the training and evaluation of the model is based on a variable (arrest) that may represent a noisy version of an unobserved outcome of more central interest (offense). We propose a sensitivity analysis framework for assessing how assumptions on the noise across groups affect the predictive bias properties of the risk assessment model as a predictor of reoffense. Our experimental results on two real world criminal justice data sets demonstrate how even small biases in the observed labels may call into question the conclusions of an analysis based on the noisy outcome.

1 Introduction

The goal of recidivism risk assessment instruments (RAI's) is to estimate the likelihood that an individual will reoffend at some future point in time, such as while on release pending trial, on probation or parole (Desmarais and Singh, 2013). Risk assessment tools have long been used in the criminal justice system to guide interventions aimed at reducing recidivism risk (James, 2015). More recently they have received considerable attention as major components of broader pretrial reform efforts seeking to reduce unnecessary pretrial detention without compromising public safety. From a public safety standpoint, society incurs a cost when a crime is committed, irrespective of whether the crime results in an arrest. The relevant fairness question in this context is thus whether a tool provides an "unbiased" prediction of who goes on to commit future crimes. However, because offending is not directly observed, risk assessment models are trained and evaluated on data where the target variable is rearrest, reconviction, or reincarceration.

While these observed proxies for offending may be of interest in their own right, they are problematic as a basis for predictive bias assessment, particularly with respect to race. Racial disparities in rearrest rates may stem from two separate causes: differential *involvement* in crime, and differential law enforcement practices, also known as differential *selection* (Piquero and Brame, 2008). Rearrest is a result of not only an individual's actions, but also of law enforcement practices affecting the likelihood of getting arrested for crimes committed (or even for crimes not committed). The limited evidence that exists suggests that differential law enforcement is not a major factor in arrests for violent crimes (Piquero, 2015). Problematically, though, for lower level offenses, which form the majority of arrests in existing data, there is reason to believe that the likelihood of getting arrested for a committed offense does differ across racial groups. Evidence of differential selection is strongest in the case of drug crimes, where surveys suggest that whites are at least as likely as blacks to sell or use drugs; yet blacks are more than twice as likely to be arrested for drug-related offenses (Rothwell, 2014). This *racially differential* discrepancy between the unobservable outcome Y^* (reoffense) and the noisy observed variable Y (rearrest) poses a critical challenge when evaluating RAI's for racial predictive bias. In this paper, we will refer to such differential discrepancy as *target variable bias* (TVB). As we show, in the presence of TVB, a model that appears to be fair with respect to rearrest could be an unfair predictor of reoffense.

We develop a statistical sensitivity analysis framework for evaluating RAI's according to several of the most common fairness metrics, including calibration,

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

predictive parity, and error rate balance. Our approach is conceptually inspired by sensitivity analysis approaches widely used in causal inference studies (Rosenbaum, 2014). When presenting analytic results it is common to report not only point estimates and confidence intervals, but also a parameter Γ reflecting the magnitude of unobserved confounding that would be sufficient to nullify the observed results. In this work we introduce a similar parameter, α, that governs the level of label bias in the observed data. Our methods characterize how the fairness properties of a model vary with α, and can be used to determine the level of label noise sufficient to contradict the observed findings about those properties. We illustrate our approach through a reanalysis of the fairness properties of the COMPAS RAI used in the ProPublica debate, and a risk assessment tool developed on data provided by the Pennsylvania Commission on Sentencing.

1.1 Related work

What we call target variable bias is often referred to as *differential* outcome measurement bias or differential outcome misclassification bias in the statistics and epidemiology literature on measurement error (Carroll et al., 2006; Grace, 2016). Most of the measurement error literature is concerned with the problem of *non*-differentially mismeasured exposure (treatment), covariates, and outcomes. That is, while this form of data bias has a name, it has received little attention relative to other measurement issues. The work of Imai and Yamamoto (2010) is a notable exception. They do consider the setting of differential measurement error, but their goal is different from ours in that they are seeking to estimate a causal effect parameter.

In the machine learning literature, our setting is known as *censoring positive and unlabeled (PU) learning* (Menon et al., 2015). This literature differs from the current work in two key ways. First, while the case of feature-independent noise has been widely studied (Elkan and Noto, 2008; Scott and Blanchard, 2009; Du Plessis et al., 2014; Liu and Tao, 2016; Menon et al., 2015), our work contributes to the nascent literature on feature-dependent noise (Menon et al., 2016; Bekker and Davis, 2018; Scott, 2018; Bootkrajang and Chaijaruwanich, 2018; Cannings et al., 2018; He et al., 2018). We believe our paper is among the first to consider issues of fairness in the context of PU learning.

There are also connections between the goal of our work and causal approaches to algorithmic bias that have recently been proposed in the fairness literature (Kusner et al., 2017; Loftus et al., 2018; Kilbertus et al., 2017; Nabi and Shpitser, 2018). These works provide an approach to addressing biases in the observed data by attempting to directly model the causal structure governing the data generating process. Problematically, the underlying assumptions are often not empirically testable, and when violated may result in incorrect inference.

Lastly, label noise has been briefly mentioned in prior work as a potential concern in the training and evaluation of RAI's (Johndrow and Lum, 2017; Corbett-Davies et al., 2017; Corbett-Davies and Goel, 2018). However, none of these works undertake a formal analysis of how label noise affects training or evaluation.

2 Problem setup

We denote the observed noisy outcome (e.g., rearrest) by Y, the true unobserved outcome (e.g., reoffense) by Y^*, the set of covariates (e.g. age, criminal history) by X, the group indicator (race) by $A \in \{b, w\}$, and the risk score (our RAI) by $S = S(X, A)$. The risk score $S(x, a)$ can be thought of as an empirical estimate of $\mathbb{E}[Y|X = x, A = a]$. When discussing binary classification metrics, we will set a risk threshold s_{HR} applied to S to obtain the classifier $\hat{Y} = \mathbb{1}_{S>s_{HR}}$. The discrepancy between the observed and true outcome is captured in the *noise rate* function $\gamma(x, a, y) \coloneqq \mathbb{P}(Y = 1 - y|X = x, A = a, Y^* = y)$. **A central aim of this work is to characterize what can be learned about the predictive bias properties of S as a predictor of the true unobserved outcome Y^* under assumptions on the magnitude but not the structure of the noise.**

We make two simplifying assumptions that, while implausible in practice, greatly simplify exposition in the main manuscript and reduce the notational overhead. First, we assume that the noise is one-sided, which rules out the case of "false arrests."

Assumption 1. $\gamma(x, a, 0) = 0$ *for all* x *and* a.

This allows us to drop the dependency on Y^* in the notation of γ, and rewrite $\mathbb{E}[Y|X = x, A = a]$ as $(1 - \gamma(x, a))\mathbb{E}[Y^*|X = x, A = a]$. That is, the discrepancy between Y and Y^* is due to the presence of "hidden recidivists". Table 1 describes the general setup for this setting. The left table represents the observed confusion matrix expressed in terms of the cell frequencies $p_{ij} = \mathbb{P}(Y = i, \hat{Y} = j)$; the right table introduces the parameters $\alpha_j \coloneqq \mathbb{P}(Y^* = 1, Y = 0, \hat{Y} = j)$. Large values of α_1 indicate that hidden recidivists are *more* likely to be classified as high risk, while large values of α_0 indicate that hidden recidivists are *less* likely to be classified as high risk. We also define $\alpha \coloneqq \alpha_0 + \alpha_1 = \mathbb{E}Y^* - \mathbb{E}Y$ that corresponds to the overall proportion of "hidden recidivists" in the observed data.

Second, in the main paper we suppose that one of the

	$\hat{Y}=0$	$\hat{Y}=1$
$Y=0$	p_{00}	p_{01}
$Y=1$	p_{10}	p_{11}

	$\hat{Y}=0$	$\hat{Y}=1$
$Y^*=0$	$p_{00}-\alpha_0$	$p_{01}-\alpha_1$
$Y^*=1$	$p_{10}+\alpha_0$	$p_{11}+\alpha_1$

Table 1: Observed (left) and true (right) confusion matrices for arrest/offense and predicted risk.

groups is being observed *without bias*.

Assumption 2. $\gamma(x, b, 1) = 0$ *for all* x.

That is, for $A = b$ we assume that $Y^* = Y$. In the running COMPAS example, this amounts to operating as though we observed the true offenses for the black population. One could also think of γ as capturing the *additional* degree of hidden recidivism in the white population relative to the black population. Again, this assumption is made solely to simplify exposition, and it does not qualitatively affect the presented results. [1] In Supplement §B.3 we show how all results are readily extensible to the case where this assumption is removed.

As we shall show next in Section 3, most of the bounds in our sensitivity analysis correspond to the case where the hidden recidivists correspond to the highest/lowest-scoring (α_0 or $\alpha_1 = 0$) defendants for whom we observed $Y = 0$. While these extreme cases may seem unlikely in practice, they generally cannot be ruled out on the basis of the observed data alone without further assumptions. In such settings, existing methods typically (1) assume some data generating mechanism to conduct sensitivity analysis (Heckman, 1979; Little and Rubin, 2019; Robins et al., 2000; Molenberghs et al., 2014), (2) assume parametric models and estimate the noise by EM algorithms (Rubin, 1976; Bekker and Davis, 2018), or (3) impose stronger conditions on the noise processes. For instance, γ may be assumed to depend only on a subset of X (Bekker and Davis, 2018) or be a monotonic function of $\mathbb{E}[Y|X = x]$ (Menon et al., 2016; Scott, 2018).

In this paper we are primarily interested in what can be said about the predictive bias properties of an RAI without untestable structural assumptions on the noise process. We note, however, that our results can be adapted to incorporate structural assumptions when reasonable ones are available. For instance, an assumption tailored to our setting might be $Y \perp\!\!\!\perp X \mid (Y^*, A)$.[2] This would assume that the noise process is constant within groups. Such an assumption probabilistically rules out extreme cases for α_0 and α_1, and, as we show in Supplement §A.2.2, it allows us to obtain tighter estimation results. There we also demonstrate how a range of results from the label-dependent noise literature can be easily adapted to our setting.

2.1 Data and background

In May 2016 an investigative journalism team at ProPublica released a report on a proprietary risk assessment instrument called COMPAS, developed by Northpointe Inc (now Equivant)(Angwin et al., 2016). The investigation found that the COMPAS instrument had significantly higher false positive rates and lower false negative rates for black defendants than for white defendants. This evidence led the authors to conclude that COMPAS is biased against black defendants. The report was met with a critical response challenging its central conclusion (Flores et al., 2016; Dieterich et al., 2016; Corbett-Davies et al.). Error rate imbalance, critics argued, is not an indication of racial bias. Instead, RAI's should be assessed for properties such as predictive parity (Dieterich et al., 2016) and calibration(Flores et al., 2016), which COMPAS was shown to satisfy. A series of papers reflecting on the debate showed that when recidivism prevalence varies across groups, as is observed to be the case in ProPublica's Broward County data, a tool cannot simultaneously satisfy both predictive parity (calibration) and error rate balance (resp. balance for the positive and negative class) (Kleinberg et al., 2016; Chouldechova, 2017; Berk et al., 2017).

One popular interpretation of such "impossibility results" is that error rate imbalance is a (perhaps inconsequential) artifact of differences in recidivism (rearrest) prevalence across groups. That is, if one were to assess the instrument on a population where prevalence was equal, the RAI could (might be expected to) achieve parity on all of the metrics simultaneously. Applying our framework to reanalyse the data in the setting where true offense rates are assumed to be the same across groups, we show that disparities with respect to Y^* (reoffense) may in fact be *greater* than those observed for Y (rearrest).

We also analyze a second private data set provided by the Pennsylvania Sentencing Commission for the purpose of research. This dataset contains information on all offenders sentenced in the state's criminal courts between 2004-2006. In reports published by the Commission, they observe that the risk assessment tool they constructed appeared to overestimate risk for white offenders. While we do not have access to their tool, the tool we construct by applying regularized logistic regression to their data evidences the

[1] For this reason, in the paper we typically denote $\alpha := \mathbb{E}[Y^*|A = w] - \mathbb{E}[Y|A = w] := \alpha^w$.

[2] This is a slight modification of label-dependent noise, or noise at random. In the PU learning and missing data literature, the latter is known as *selected at random (SAR)* (Bekker and Davis, 2018) and *missing not at random (MNAR)* (Rubin, 1976) respectively.

same miscalibration issues. Our empirical results are based on applying this score to a held out set of 55031 offenders, of whom 65.4% are white.

3 Sensitivity analysis under target variable bias

This section presents our main technical results, coupled with experiments that demonstrate how the results may be used in practice. All proofs are contained in Supplement §B.1. Given observations (Y, S) and a classification threshold s_{HR}, we want to understand how the relationship between the observed (M) and unobserved (M^*) performance metrics depends on the noise level α in the problem setup outlined in Section 2. Superscripts w and b denote within-race group estimates. We present sensitivity analysis results for predictive parity, error rate balance (aka equalized odds (Hardt et al., 2016)), accuracy parity, and two tests of differential calibration. Supplement §C presents experiments on the COMPAS data set for two fairness-promoting algorithms. All code is available at https://github.com/ricfog/Fairness-tvb.

3.1 Error rate balance and predictive parity

We begin by presenting results for the false positive rate (FPR), the false negative rate (FNR), and the positive predicted value (PPV). Our first result shows that the observed values FPR and FNR impose constraints on the true error rates even if no assumptions are made on the magnitude of the noise.

Proposition 3.1. *Suppose that $1 - FPR < FNR$. Then $FNR \leq FNR^*$ and $FPR \geq FPR^*$ cannot both hold. If $1 - FPR > FNR$, then the opposite inequalities can not both hold.*

Proposition 3.1 permits us to rule out one of the possible relations between observed and true error rates based solely on observed quantities.

Example: COMPAS. In ProPublica's COMPAS analysis, we observe that $FPR^w = 0.23$ and $FNR^w = 0.48$. We are thus in the case where $1 - FPR > FNR$, and therefore either $FNR^w = 0.48 \leq FNR^{*w}$ or $FPR^w = 0.23 \geq FPR^{*w}$, or both.

The next set of results directly relate the observed metrics M to the target quantities M^* based on the noise level α. Table 1 summarizes the relationship between the observed and target confusion tables used to derive these relationships. While a version of the FPR results was previously reported in (Claesen et al., 2015), the case of PPV and FNR are novel.

Theorem 3.2. *Under the setup of Table 1, the target values FPR^*, FNR^*, and PPV^* can be sharply related to observed quantities as follows:*

$$\frac{p_{01} - \alpha}{p_{00} + p_{01} - \alpha} \leq FPR^*(\alpha_0, \alpha_1) \leq \frac{p_{01}}{p_{00} + p_{01} - \alpha} \quad (1)$$

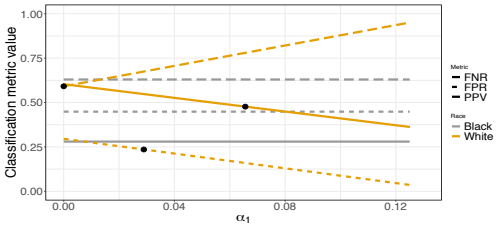
$$\frac{p_{10}}{p_{10} + p_{11} + \alpha} \leq FNR^*(\alpha_0, \alpha_1) \leq \frac{p_{10} + \alpha}{p_{10} + p_{11} + \alpha} \quad (2)$$

$$PPV \leq PPV^*(\alpha_0, \alpha_1) \leq \frac{p_{11} + \alpha}{p_{01} + p_{11}} \quad (3)$$

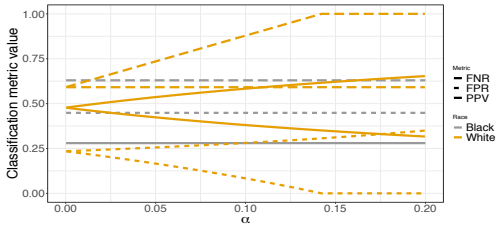
Example: COMPAS. This result allows us to reanalyse ProPublica's COMPAS data to answer the question: *If the reoffense rate was equal across races, would disparities disappear?* Figure 1a shows the possible values of $PPV(\alpha_0, \alpha_1)$, $FPR(\alpha_0, \alpha_1)$, and $FNR(\alpha_0, \alpha_1)$ for fixed $\alpha = 0.12$. At this choice of α, the true reoffense rate among white defendants is assumed equal to the rate observed for black defendants. Since α is fixed, $\alpha_0 = 0.12 - \alpha_1$ and hence the metrics are a function of just α_1. We see that *for most values of α_1 disparities are even greater than what is observed.* Furthermore, while there exist values of α_1 under which the true metric for white defendants would equal the observed (and assumed true) metric for black defendants, the equalizing value of α_1 differs across the metrics.

Figure 1b shows the theoretical bounds (orange lines) provided by Theorem 3.2 as functions of α for the white population, and the observed metrics for the black population (grey lines) on the COMPAS data. We highlight the regions highlighted in red, which indicate areas where the true disparity in metrics could be of a different sign than what is observed. This plot also shows that parity on the true FPR and FNR is infeasible in this data at the given choice of classification threshold.

As a corollary of this result we can also study the question: *Under what level of label noise could we expect disparities on a given metric to be smaller in truth than what was observed?* First, note that when the observed recidivism rate is greater in group b than w, as in the case of the COMPAS example, we will generally observe $FPR^w \leq FPR^b$ and $FNR^w \geq FNR^b$. A necessary condition for the disparity between the true error rates to be no larger than that for the observed rates is thus that $FNR^w \geq FNR^{*w}(\alpha_0, \alpha_1)$ and $FPR^w \leq FPR^{*w}(\alpha_0, \alpha_1)$. The following corollary characterizes when this occurs.



(a) α fixed at 0.12 to equalize reoffense rates across groups. Black dots shown indicate the value (α_1, M^*) for which $M^*(\alpha_1) = M$ (observed equals true).



(b) Bounds as described in Theorem 3.2 in terms of α. Red area corresponds to region where $\text{sign}(M^{*w} - M^b) \neq \text{sign}(M^w - M^b)$.

Figure 1: Analysis of predictive parity and error rates for COMPAS across different TVB scenarios. Orange lines show values of FPR^{*w}, FNR^{*w}, and PPV^{*w}. Grey lines show corresponding values for the black population.

Corollary 3.2.1. *In the notation of Theorem 3.2,*

$$FPR \geq \frac{\alpha_1}{\alpha} \iff FPR \leq FPR^*(\alpha, \alpha_1) \tag{4}$$

$$FNR \geq \frac{\alpha_0}{\alpha} \iff FNR \geq FNR^*(\alpha, \alpha_1), \tag{5}$$

with equality on LHS iff there is equality on RHS.

The condition in (5) turns out to be equivalent to the odds ratio:[3]

$$\frac{\mathbb{P}(\hat{Y} = 1 | Y^* = 1, Y = 0) / \mathbb{P}(\hat{Y} = 0 | Y^* = 1, Y = 0)}{\mathbb{P}(\hat{Y} = 1 | Y = 1) / \mathbb{P}(\hat{Y} = 0 | Y = 1)} \geq 1. \tag{6}$$

This condition tells us that (5) holds precisely when the odds of correctly classifying a *hidden* recidivist to $\hat{Y} = 1$ are greater than the odds of correctly classifying an *observed* recidivist, which seems unlikely to hold in practice. A similar interpretation can be derived for FPR: condition (4) holds when the odds of misclassifying a *hidden* recidivist to $\hat{Y} = 0$ are higher than those of correctly classifying an *observed* non-recidivist.

Example: COMPAS. Conditions (4) and (5) in Corollary 3.2 require $\alpha_1 \leq 0.3\alpha_0$ and $\alpha_1 \geq 1.09\alpha_0$ respectively. Note, however, that both conditions cannot simultaneously hold, as formally shown in Proposition 3.1.

In practice, if the predicted risk for hidden recidivists was generally low, condition (6) would likely not hold. Consequently, we would thus have $FNR^w - FNR^b \leq FNR^{*w}(\alpha, \alpha_1) - FNR^b$, which says that the true FNR disparity between groups would be greater than the observed FNR disparity.

[3](Kallus and Zhou, 2018) obtain similar expressions in their study of "residual unfairness" in the context of a related data bias problem. They consider the setting where we fail to observe outcomes entirely for a fraction of the population (e.g., defendants who are not released on bail, and thus do not have the opportunity to recidivate). When viewed as functions of the underlying classification threshold s_{HR}, these odds ratios are interpreted in (Kallus and Zhou, 2018) as a type of stochastic dominance condition.

3.2 Accuracy equity

In their response to the ProPublica investigation, Dieterich et al. (2016) demonstrated that COMPAS satisfies predictive parity (equality of PPV and NPV across groups), and what they term *accuracy equity* (equality of AUC). Menon et al. (2015) and Jain et al. (2017) previously considered estimation of the AUC under label noise, but in the simpler setting of label-dependent noise. Here we obtain bounds for the true AUC in the general instance-dependent noise setting through its relation to the Mann-Whitney U-statistic.

Let $n_y = \#\{Y_i = y\}$ denote the number of observations with outcome $Y = y \in \{0, 1\}$. We will assume that there are $k = \lceil n\alpha \rceil$ hidden recidivists present in the observed data, with $k < \min(n_0, n_1)$. Let r_i denote the adjusted[4] rank of observation i when ordered in ascending order of the score S. Lastly, let $R_1 = \sum_{i:Y_i=1} r_i$ denote the sum of the ranks for observations in class $Y = 1$. In this notation, the observed AUC of S is given by

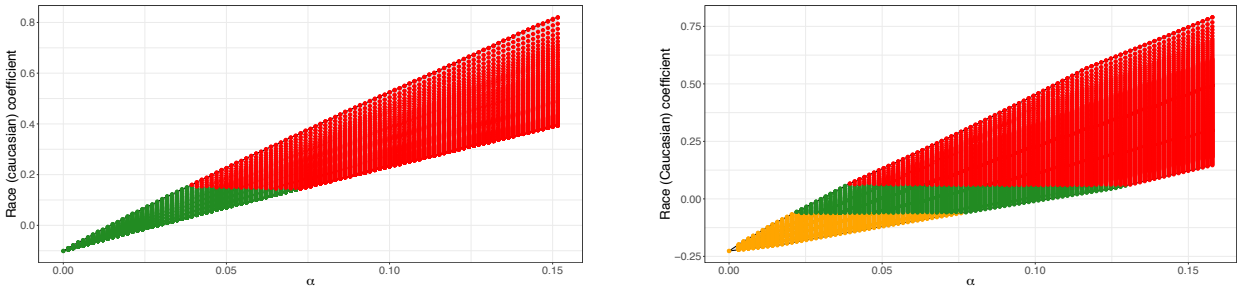
$$AUC = \frac{R_1}{n_1(n - n_1)} - \frac{n_1 + 1}{2(n - n_1)} \tag{7}$$

Let $L_{0,k}$ denote the indexes of the lowest-ranked (i.e., lowest-scoring) observations in class $Y = 0$. Likewise, let $H_{0,k}$ denote the indexes of the highest-ranked (i.e., highest-scoring) observations in class $Y = 0$.

Proposition 3.3. *In the presence of k hidden recidivists, the target value AUC is bounded as follows:*

$$\frac{R_1 + \sum_{i \in L_{0,k}} r_i - \beta_k}{(n_0 - k)(n_1 + k)} \leq AUC^* \leq \frac{R_1 + \sum_{i \in H_{0,k}} r_i - \beta_k}{(n_0 - k)(n_1 + k)} \tag{8}$$

[4]In the case of ties among the scores, the U-statistic is calculated using fractional ranks.



(a) Calibration analysis on COMPAS data.

(b) Calibration analysis on Sentencing comm. data.

Figure 2: Sensitivity analysis for the race coefficient in a logistic regression test of calibration as described in Section 3.3. Green region indicates the race coefficient is not statistically significant for testing calibration wrt *offense*. Red (resp., orange) region indicates statistically significant bias against the black (resp., white) group.

where $\beta_k = (n_1 + k)(n_1 + k + 1)/2$.

It is easy to see that the upper and lower bounds correspond to the settings where the hidden recidivists are, respectively, the highest and lowest scoring defendants with $Y = 0$. This result tells us, for instance, that if the hidden recidivists are more likely to have high scores, then the true AUC will be greater than the observed AUC. One key difference between the AUC result and the previous analysis of error metrics is that now the impact of label noise depends on the ranks of the hidden recidivists, and not only on the dichotomized version of the risk score.

Example: COMPAS. The observed AUC for both the black and white defendant population is around 0.69. Evaluating the bounds from the proposition for the white population, we find that for $\alpha = 0.05$ and $\alpha = 0.12$, the AUC^{*w} is bounded between $[0.63, 0.76]$ and $[0.51, 0.84]$, respectively. These bounds are very wide, but they can be narrowed if we are willing to make further assumptions on the likely ranks of the hidden recidivists.

3.3 Calibration testing via logistic regression

One of the most common metrics for assessing predictive bias of RAI's is a test of *calibration* or *differential prediction* (Skeem and Lowenkamp, 2015). Formally, we say that a risk score S is well-calibrated with respect to A if

$$\mathbb{E}[Y \mid S = s, A = w] = \mathbb{E}[Y \mid S = s, A = b]. \qquad (9)$$

for all values of S. This is equivalent to requiring that $Y \perp\!\!\!\perp A \mid S$. Typically calibration is assessed by running a logistic regression and testing for statistical significance of A in $Y \sim S$ vs. $Y \sim S + A$ or $Y \sim S + A + SA$ using a Wald or likelihood ratio test.[5]

Other covariates are occasionally also included in the regression. When the coefficients of A are not statistically significant, S is deemed to be well-calibrated with respect to A. This approach was taken by Flores et al. (2016) to confirm racial calibration for the COMPAS RAI. Note that in the presence of TVB, such tests provide evidence that S is well-calibrated as a predictor of Y (rearrest). We wish to understand what this means about S as a predictor of the true outcome Y^* (reoffense). Our main result is as follows.

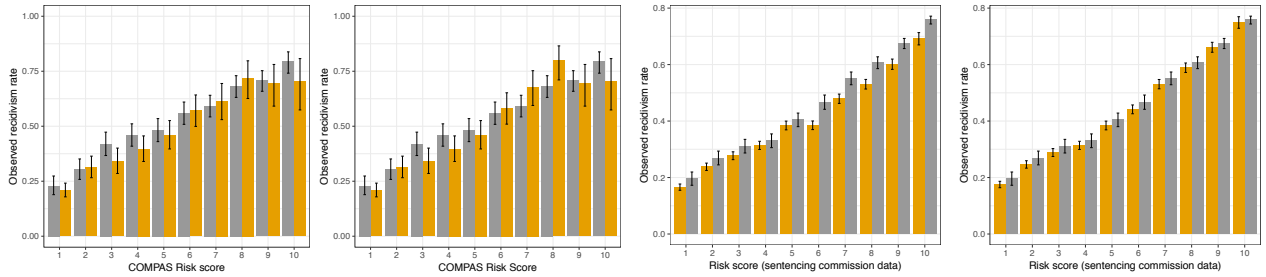
Proposition 3.4. *Under a mild technical assumption on the design matrix,*[6] *for a logistic regression model of the form* $Y \sim S + A$*, for fixed* α*, the bounds for the coefficients of* S *and* A *are achieved when the* $\lceil n_w \alpha \rceil$ *white defendants with the highest and lowest values of* S *are hidden recidivists.*

This result allows us to answer the question: *What level of label noise* α *is sufficient to contradict the observed findings that an RAI is (or is not) well-calibrated across groups?* We provide two illustrative examples, one where the RAI is observed to be well-calibrated as a predictor of arrest, and the other where it is not.

Example: COMPAS. Figure 2 (a) shows the feasible values for the coefficient of $A = w$ in the COMPAS data for $0 \leq \alpha \leq 0.16$. The green and red areas correspond, respectively, to regions where the race coefficient *is not* and *is* statistically significant. Recall that non-significance of the race coefficient indicates that the model is well-calibrated. In this analysis, we find that a TVB level as low as $\alpha = 0.07$ might be sufficient

[5] We adopt the shorthand $Y \sim X_1 + X_2 + \cdots X_p$ to refer to the logistic regression model $\log(p(X)/(1 - p(X)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$, where $p(X) = \mathbb{P}(Y = 1 \mid X)$.

[6] The explanation of the assumption is deferred to the proof in the Supplement. While the assumption needs to be empirically verified case by case, in the COMPAS dataset it holds at every level of α^w that we considered.



(a) Observed COMPAS. $T = 9.36$, p-value$= 0.49$.

(b) Minimal shift to break calibration under proportionality constraint. $N_h = 30$, $T = 19.4$, p-value $= 0.035$.

(c) Observed SC score. . $T = 164$, p-value ≈ 0

(d) Minimal shift to achieve calibration, unconstrained. $N_h = 1001$, $T = 18.2$, p-value $= 0.051$

Figure 3: Sensitivity analysis for the race coefficient in a chi-squared nonparametric test of calibration as described in Section 3.4. Orange bars show white defendant data; grey bars are black defendant data. Error bars show 95% confidence intervals. Plots (a) and (b) correspond to the COMPAS data example where a small amount of TVB is sufficient to lead to miscalibration. Plots (c) and (d) are the sentencing commission (SC) example where a small amount of TVB can account for observed miscalibration in predicting arrest.

for COMPAS to fail the calibration test across *all* possible noise realizations of that magnitude. At a noise level of only $\alpha = 0.04$, calibration might also fail for *some* noise realizations of this magnitude. Note that our analytic results present bounds not just on the race coefficient but also on the score coefficient in the model. We present the two-dimensional bounds for the COMPAS tool in Supplement §B.4.

Example: Sentencing commission. Figure 2 (b) shows the results of the same experiment on sentencing commission (SC) data described in Section 2.1. In absence of TVB, Figure 2 (b) shows that this tool, unlike the COMPAS RAI, is *not* observed to be well-calibrated across groups. Indeed, the coefficient for $A = w$ is statistically significantly negative, indicating the RAI overestimates risk for white offenders. Our analysis showns that TVB as low as $\alpha = 0.03$ is sufficient to admit calibration. More generally, we see that for $0.03 \leq \alpha \leq 0.13$ calibration might be possible for some realizations of the noise process. For a larger magnitude of TVB, the coefficient might be significant and positive; in other words, it would be possible for the instrument to underestimate the reoffense risk for the white population.

3.4 Calibration testing via chi-squared test

We also consider the general test of conditional independence $Y \perp\!\!\!\perp A \mid S$ in the setting where S is either assumed to be discrete, or has been binned for the purpose of analysis. When S is categorical, testing the saturated logistic model $Y \sim S + A + SA$ vs. $Y \sim S$ is precisely testing the conditional independence of $Y \perp\!\!\!\perp A \mid S$. This section thus extends the analysis from the previous section beyond the (likely misspecified) simple shift-alternative considered therein. There are several asymptotically equivalent tests that can be applied to test this hypothesis (Hinkley and Cox, 1979). We use the Pearson chi-squared test, as it is the most straightforward to analyse.

The general setup for assessing the sensitivity of the chi-squared conditional independence test to TVB is described by Table 2. Our goal is to understand the behavior of the chi-squared test statistic,

$$T(h) = \sum_{k=1}^{|S|} \sum_{a,y} \left(O_{ay}^{(k)} - E_{ay}^{(k)} \right)^2 / E_{ay}^{(k)} \qquad (10)$$

as a function of the hidden recidivist counts $h = (h_1, \ldots, h_{|S|})$. The notations O and E denote the "observed" and "expected" cell counts for calculating the chi-squared statistic. Expected counts are estimated from the data assuming the null hypothesis $Y^* \perp\!\!\!\perp A \mid S$ is true. These quantities evaluate to

$$O_{ay}^{(k)} = n_{ay}^{(k)} + h_k \mathbb{1}_{a=w}(2y-1), \text{ and}$$
$$E_{ay}^{(k)} = \left(n_{wy}^{(k)} + n_{by}^{(k)} + (2y-1)h_k \right) \left(n_{a0}^{(k)} + n_{a1}^{(k)} \right) / n^{(k)}.$$

The key observation is that, when viewed as a function of h_k, the numerator terms $(O_{ay}^{(k)} - E_{ay}^{(k)})^2$ are convex quadratics in h_k, and the denominator terms $E_{ay}^{(k)}$ are linear functions in h_k, constrained to be positive.

We address two basic questions: (1) *When S appears racially well-calibrated for the observed Y, how large would N_h, the number of hidden recidivists, have to be for S to fail the calibration test for Y^*?* (2) *When S appears to underestimate risk for the one racial*

$S = k$	$Y^* = 0$	$Y^* = 1$
$A = w$	$n_{w0}^{(k)} - h_k$	$n_{w1}^{(k)} + h_k$
$A = b$	$n_{b0}^{(k)}$	$n_{b1}^{(k)}$

Table 2: Contingency table for rearrest outcome in score level $S = k \in \{1, \ldots, |S|\}$ for testing $H_0 : Y^* \perp\!\!\!\perp A \mid S$ with the chi-squared test. Here h_k denotes the number of "hidden recidivists" in the white defendant population in score level $S = k$.

group, how large would N_h have to be for S to appear racially well-calibrated for Y^*? Answering (1) entails maximizing the test statistic T over h subject to $\sum h_k \leq N_h$. Answering (2) entails minimizing the test statistic. Note that each inner summand of equation (10) is a quadratic-over-linear function, which is strongly convex (Boyd and Vandenberghe, 2004). The test statistic T as a function of h thus has the form $T(h) = \sum_{k=1}^{|S|} f_k(h_k)$, where each f_k is a strongly convex function. Since $T(h)$ is a strongly convex separable function of the h_k's, the minimization can be performed with a numerical convex solver. Note that it is also straightforward to incorporate convex constraints into the optimization. The maximization task is a case of a separable nonlinear optimization problem, for which general tools exist. For our analysis we instead present a practical greedy algorithm in Supplement §B.1.4.

Example: COMPAS. Figure 3(a) shows the observed recidivism rates for black and white defendants across the range of the COMPAS decile score. When we apply the chi-squared test to test for calibration, we find that the COMPAS instrument appears well-calibrated with respect to race ($T = 9.36, p$-value $= 0.49$). However, applying our method to *maximize* the test statistic, we find that the presence of just $N_h = 20$ hidden recidivists is sufficient to break calibration. This is achieved when all $N_h = 20$ hidden recidivists are located in score level 8. Looking at the data, this is unsurprising. Score level 8 already has the largest observed discrepancy with the black defendant recidivism rate. Pushing this discrepancy further will rapidly cause the test to reject. Figure 3(b) shows the minimal shift necessary to break calibration when we impose a *proportionality constraint* that prohibits allocations that concentrate too much on a single bin. Specifically, we require that $h_k \leq \epsilon n_{w1}^{(k)}$. This ensures that the proportion of true recidivists that are hidden *in any score bin* is no greater than ϵ. For our experiment we take $\epsilon = 0.1$. Under this constraint, we find that $N_h = 30$ are sufficient to break calibration. These are allocated as $h = (0, 0, 0, 0, 0, 12, 9, 9, 0, 0)$.

Example: Sentencing commission. The right panel of Figure 3 shows the observed recidivism rates for black and white defendants across the range of the decile score we constructed based on the sentencing commission data. Unlike in the COMPAS example, we find that the SC score shows clear evidence of poor calibration ($T = 164, p$-value ≈ 0). The RAI underestimates risk of rearrest for white offenders relative to black offenders across the range of score levels. This effect is especially pronounced in the highest scores. Applying our method to *minimize* the test statistic, we find that just $N_h = 1001$ hidden recidivists are sufficient to achieve calibration. While this may seem like a large number, there are $n_w = 31607$ white offenders in the data, of which $n_{w1} = 13552$ are observed to reoffend. Thus the minimizing allocation requires only that $1001/(13552 + 1001) = 6.9\%$ of all true recidivists go unobserved. The minimizing allocation, represented in the left panel of Figure 3, is $h = (41, 35, 46, 0, 0, 224, 186, 197, 170, 102)$.

4 Conclusion

When target variable bias is a concern, the sensitivity analysis framework presented in this paper can be used to quantify the level of bias sufficient to call into question conclusions about the fairness of a model obtained from biased observed data. In the sentencing commission example, for instance, we find that a small gap in the likelihood of arrest could fully account for the observed miscalibration. Such observations may help inform deliberations of whether to correct for observed predictive bias when doing so would further increase outcome disparities. Furthermore, as our reanalysis of the ProPublica COMPAS data shows, the racial disparity story goes deeper than an imbalance in observed recidivism rates. Even if offense rates are equal across groups, the disparities could be worse with respect to offense than what is observed for arrest.

The sensitivity analysis approach outlined in this work has generally avoided making assumptions about how the likelihood of getting caught might depend on observable features, at a cost of producing fairly wide bounds. Existing work on self-report studies, wrongful arrests, and wrongful convictions may provide some insight into reasonable structural assumptions that may be incorporated to further refine the analysis (Huizinga and Elliott, 1986; Hindelang et al., 1979; Gilman et al., 2014).

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data under the selected at random assumption. *arXiv preprint arXiv:1808.08755*, 2018.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 0(0):0049124118782533, 2017. doi: 10.1177/0049124118782533. URL https://doi.org/10.1177/0049124118782533.

Jakramate Bootkrajang and Jeerayut Chaijaruwanich. Towards instance-dependent label noise-tolerant classification: a probabilistic approach. *Pattern Analysis and Applications*, pages 1–17, 2018.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Timothy I Cannings, Yingying Fan, and Richard J Samworth. Classification with imperfect training labels. *arXiv preprint arXiv:1805.11505*, 2018.

Raymond J Carroll, David Ruppert, Ciprian M Crainiceanu, and Leonard A Stefanski. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 2017.

Marc Claesen, Jesse Davis, Frank De Smet, and Bart De Moor. Assessing binary classifiers using only positive and unlabeled data. *arXiv preprint arXiv:1504.06837*, 2015.

Stéphan Clémençon, Gábor Lugosi, Nicolas Vayatis, et al. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 797–806, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3098095. URL http://doi.acm.org/10.1145/3097983.3098095.

Sarah Desmarais and Jay Singh. Risk assessment instruments validated and implemented in correctional settings in the united states. 2013.

William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. 2016.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

Marthinus C Du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *Advances in neural information processing systems*, pages 703–711, 2014.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220. ACM, 2008.

Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.". *Unpublished manuscript*, 2016.

Amanda B Gilman, Karl G Hill, BK Elizabeth Kim, Alyssa Nevell, J David Hawkins, and David P Farrington. Understanding the relationship between self-reported offending and official criminal charges across early adulthood. *Criminal behaviour and mental health*, 24(4):229–240, 2014.

Y Yi Grace. *Statistical Analysis with Measurement Error Or Misclassification*. Springer, 2016.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

Fengxiang He, Tongliang Liu, Geoffrey I Webb, and Dacheng Tao. Instance-dependent pu learning by bayesian optimal relabeling. *arXiv preprint arXiv:1808.02180*, 2018.

James J Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

Michael J Hindelang, Travis Hirschi, and Joseph G Weis. Correlates of delinquency: The illusion of discrepancy between self-report and official measures. *American sociological review*, pages 995–1014, 1979.

David Victor Hinkley and DR Cox. *Theoretical statistics*. Chapman and Hall/CRC, 1979.

David Huizinga and Delbert S Elliott. Reassessing the reliability and validity of self-report delinquency measures. *Journal of quantitative criminology*, 2(4): 293–327, 1986.

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

Shantanu Jain, Martha White, and Predrag Radivojac. Recovering true classifier performance in positive-unlabeled learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Nathan James. *Risk and Needs Assessment in the Criminal Justice System*, volume 44087. Washington, DC: Congressional Research Service, 2015.

James E. Johndrow and Kristian Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *arXiv preprint arXiv:1703.04957*, 2017.

Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. *arXiv preprint arXiv:1806.02887*, 2018.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. Wiley, 2019.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2016.

Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*, pages 125–134, 2015.

Aditya Krishna Menon, Brendan Van Rooyen, and Nagarajan Natarajan. Learning from binary labels with instance-dependent corruption. *arXiv preprint arXiv:1605.00751*, 2016.

Geert Molenberghs, Garrett Fitzmaurice, Michael G Kenward, Anastasios Tsiatis, and Geert Verbeke. *Handbook of missing data methodology*. Chapman and Hall/CRC, 2014.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access, 2018.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204, 2013.

Alex R Piquero. Understanding race/ethnicity differences in offending across the life course: Gaps and opportunities. *Journal of developmental and life-course criminology*, 1(1):21–32, 2015.

Alex R Piquero and Robert W Brame. Assessing the race–crime and ethnicity–crime relationship in a sample of serious adolescent delinquents. *Crime & Delinquency*, 54(3):390–422, 2008.

James M Robins, Andrea Rotnitzky, and Daniel O Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. Springer, 2000.

Paul R Rosenbaum. Sensitivity analysis in observational studies. *Wiley StatsRef: Statistics Reference Online*, 2014.

Jonathan Rothwell. How the war on drugs damages black social mobility. *The Brookings Institution, published Sept*, 30, 2014.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Clayton Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Artificial Intelligence and Statistics*, pages 838–846, 2015.

Clayton Scott. A generalized neyman-pearson criterion for optimal domain adaptation. *arXiv preprint arXiv:1810.01545*, 2018.

Clayton Scott and Gilles Blanchard. Novelty detection: Unlabeled data definitely help. In *Artificial Intelligence and Statistics*, pages 464–471, 2009.

Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference On Learning Theory*, pages 489–511, 2013.

Jennifer L Skeem and Christopher T Lowenkamp. Risk, race, & recidivism: Predictive bias and disparate impact. *Available at SSRN*, 2015.