# Data-Centric Factors in Algorithmic Fairness

Nianyun Li
ETH Zurich
Zürich, Switzerland
nianyun.li@gess.ethz.ch

Naman Goel
University of Oxford
Oxford, UK
naman.goel@cs.ox.ac.uk

Elliott Ash
ETH Zurich
Zürich, Switzerland
elliott.ash@gess.ethz.ch

## ABSTRACT

Notwithstanding the widely held view that data generation and data curation processes are prominent sources of bias in machine learning algorithms, there is little empirical research seeking to document and understand the specific data dimensions affecting algorithmic unfairness. Contra the previous work, which has focused on modeling using simple, small-scale benchmark datasets, we hold the model constant and methodically intervene on relevant dimensions of a much larger, more diverse dataset. For this purpose, we introduce a new dataset on recidivism in 1.5 million criminal cases from courts in the U.S. state of Wisconsin, 2000-2018. From this main dataset, we generate multiple auxiliary datasets to simulate different kinds of biases in the data. Focusing on algorithmic bias toward different race/ethnicity groups, we assess the relevance of training data size, base rate difference between groups, representation of groups in the training data, temporal aspects of data curation, including race/ethnicity or neighborhood characteristics as features, and training separate classifiers by race/ethnicity or crime type. We find that these factors often do influence fairness metrics holding the classifier specification constant, without having a corresponding effect on accuracy metrics. The methodology and the results in the paper provide a useful reference point for a data-centric approach to studying algorithmic fairness in recidivism prediction and beyond.

## CCS CONCEPTS

• **Human-centered computing**; • **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → **Machine learning**; • **Information systems** → *Decision support systems*;

## KEYWORDS

Algorithmic Fairness, Datasets, Recidivism Prediction, Machine Learning

**ACM Reference Format:**
Nianyun Li, Naman Goel, and Elliott Ash. 2022. Data-Centric Factors in Algorithmic Fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3514094.3534147



*AIES '22, August 1–3, 2022, Oxford, United Kingdom.*

ACM ISBN 978-1-4503-9247-1/22/08.
https://doi.org/10.1145/3514094.3534147

## 1 INTRODUCTION

Undesirable bias in artificial intelligence systems has been repeatedly demonstrated in prior research. For instance, decision support systems for credit loan applications were found to favor certain sociodemographic groups [30, 33]. Natural language understanding and translation systems often exhibit undesired gender stereotypes [16]. Facial recognition applications usually don't perform well for people with dark skin color as they are trained and evaluated on data from white people [15]. Systems for automated criminal risk scoring discriminate against some racial groups [20].

A central goal of "fair machine learning" systems is to prevent such disparate harm across groups [9]. In practice, these subgroups are most often linked to protected demographic attributes, e.g., gender, age, and race/ethnicity. Simply omitting sensitive attributes from the system does not solve this problem because correlated non-sensitive attributes can act as proxies for the protected attribute (e.g., salary as a proxy of gender).

Bias in artificial intelligence systems can come in at any stage (from task definition to model deployment). This paper's focus is on the data related factors. That includes both training data as well as target population related factors.

It is widely acknowledged that data influences the fairness of the algorithms that are trained on it [9, 11, 23, 53]. But most of the algorithmic fairness papers use small, partial, unrepresentative datasets. These datasets include the "benchmark" datasets like the COMPAS dataset assembled by ProPublica[48], the census income dataset [45] (also known as the ADULT dataset), the German credit dataset [35], etc. These datasets do not provide sufficiently robust understanding of the extensive variation in real-world contexts.

These limitations and the importance of data-centric research [55] has been increasingly recognized in the community. The issue is not limited to algorithmic fairness; it affects empirical evaluation of several other robustness properties of machine learning algorithms as well. As a result, NeurIPS (one of the most important machine learning conferences and publication venues) launched a new track in 2021 called the "Datasets and Benchmarks Track".

Yet the existence of multiple datasets and their use in benchmarking is not the only sign of progress. We also need to understand what underlying factors in a given dataset make an algorithm behave in a certain way. In this paper, we empirically explore this dimension of the data-centric research for the case of algorithmic fairness. We introduce a new dataset and provide empirical insights about the fairness of basic classification algorithms on this new dataset. We show how researchers and practitioners can use large datasets such as this to rigorously understand the effect of data related factors on the predictive performance and fairness of algorithms.

We focus on the problem of recidivism risk prediction [20]. We collected a large-scale dataset from Wisconsin circuit courts. From those records, we have information on defendants' demographics

such as gender, race and age, the characteristics of cases such as charges and severity, and the outcomes of those cases. After preprocessing, we obtain a dataset for recidivism prediction with around 1.5 million convicted criminal cases from 2000 to 2018.

We show how one can use this main dataset to generate multiple auxiliary datasets to simulate different types of biases in the data. These auxiliary datasets can then be used to understand the effect of different data characteristics and different data curation processes on algorithmic fairness for five different identity groups (african american, caucasian, native american, asian and hispanic). More specifically, we study the effect of training data size, base rate difference across groups, proportion of different groups, temporal aspects of data curation, type of crime, data from different counties with different types of biases, availability of additional attributes etc. In addition to AUC and accuracy for predictive performance, we consider demographic parity, equal error rates (false positive and false negative rates) and bias amplification as fairness notions.

We report a rich sequence of results on how data matters for the fairness of recidivism predictions. We find, first, that more training data does not necessarily lead to a fairer model when the data generating process stays the same. Second, base rates and group sizes are not the only determinants of unfairness; the disparity does not necessarily decrease when we try to balance these two factors between races. Third, temporal factors are necessary to consider when designing and evaluating the models; depending on the time of training data and when the model is applied, fairness evaluation varies significantly. Fourth, adding race as an attribute may increase unfairness without increasing accuracy, but adding neighborhood characteristics increases fairness in our experiments. Fifth, for some types of crime, fairness is much worse than on other types of crime. Sixth, if we train separate models for different races, it is not always better for the minority groups.

More generally, in the above experiments, we find that there is always a significant discrepancy in fairness and accuracy estimates when the datasets available for training and evaluation vary, even when the target distribution does not vary. In particular, in several experiments, the intervention affects fairness metrics but not accuracy metrics. This highlights the more general issue with training and evaluating machine learning models under distribution shift.

These rich results demonstrate the promise of a data-centric approach to algorithmic fairness. This idea can also be extended to other data related factors not studied in this paper e.g. bias due to selective labeling, geographical factors, judge characteristics in past decisions etc. It may also be used to understand how various data related factors interact with common algorithmic design choices, for example, whether the classifier is a deterministic or a randomized one, the type of loss function, type of fairness metric the classifier implements, fairness intervention approach (pre-processing, in-processing or post-processing) being used, and more. We hope this research helps motivate these future explorations.

The rest of the paper is organized as follows. In Section 2, we discuss works that are most closely related to ours. In Section 3, we describe how the dataset was collected from the Wisconsin Circuit Courts Access (WCAA) and how different variables in the dataset were inferred from the raw data. In Section 4, we discuss the role of various data related factors on predictive performance and fairness of algorithms. We conclude and discuss future work in Section 5.

## 2 RELATED WORK

In the vast and growing literature on algorithmic fairness, our paper fits into the sub-field on data-centric issues. The closest paper is Ding et al. [26]. They take a data-centric approach to fairness in a set of prediction tasks using U.S. Census data. They construct five datasets related to income, employment, health coverage, commute time, and housing. They show that many results are contingent on the data. Several other works [58, 62, 63, 65, 66] point out limitations of using a few specific datasets for benchmarking ML algorithms.

Fabris et al. [27] focuses specifically on benchmark datasets used in algorithmic fairness. A number of other papers [9, 11, 23, 38, 40, 50, 53, 56, 61] provide overviews of the various factors in the data collection and curation process that may affect fairness. These include, for example, representation in terms of size, biases due to missingness, geographical and temporal differences, etc.

Our paper complements this work by exploring datasets related to recidivism prediction. In the field of criminal justice algorithms, Bao et al. [8] highlight several problems with the COMPAS dataset for recidivism prediction problems. The authors argue that "reviewers should encourage well-designed simulations" if benchmarking on a real dataset like COMPAS alone is not useful.

In this vein, Lum and Isaac[52] use simulations to investigate how feedback loops may increase the unfairness of a predictive policing algorithm. Other work on simulations includes D'Amour et al.[24], who use synthetic simulations (not based on datasets) for evaluating the effect of fairness interventions in dynamic populations. Our paper independently shows how more of such simulations can be designed using a large dataset for better understanding of data-centric factors that influence algorithmic fairness. L. Cardoso et al.[46] use synthetic biased datasets that are similar to real world data, created by a Bayesian network approach to benchmark discrimination-aware models. Another interesting work in this line has appeared since the acceptance of our paper for publication; [1] on "stress-testing" fairness algorithms under different data biases.

This is an active research area. Chen et al.[17] decompose discrimination metrics into bias, variance and noise, and show that unfairness due to inadequate samples or predictive variables should be addressed through data collection. Słowik and Bottou[64] study the relation between distributionally robust optimization (DRO) and data curation. The authors conclude that there is no universally robust training set or a universal way to setup a DRO problem to ensure desired fairness results. [7, 49] analyze the issue of bias amplification in context of various data related factors. [68] revisit the fairness-accuracy trade-off under label bias, and show that fairness and accuracy may not necessarily be in conflict if label bias is taken into account during model evaluation.

Several works also study in more detail specific data-centric factors. For example, [67] focus on the issue of gender bias in deep image representations and show that even models trained on balanced datasets amplify the association between labels and gender, as much as if data had not been balanced. [71] consider the problem of bias amplification in visual semantic role labeling, and show how injecting corpus-level constraints can mitigate that problem. [69] propose filtering and balancing the ImageNet dataset [25] in order to improve the fairness of computer vision algorithms trained on it. [60] provide arguments for and against a more careful data curation

approach for building more reliable and bias-free natural language processing capabilities.

[32, 54] show how to combine data from different sources or human labelers with different costs and accuracies, while guaranteeing fairness with respect to different groups. [14] observe that balanced samples improve fairness of not only machine classifiers but also of human labelers in crowdsourcing. [70] propose a Shapley value based method to attribute unfairness to data and algorithm. [42] show the effect of counterfactually augmented data on the performance of sentiment analysis classifiers. [29] discuss the tension between different fairness definitions depending on the worldview one assumes from the observed data. [10] study the empirical effect on accuracy of using race and zipcode as predictors in forecasting failure on probation or parole. [13] propose an adversarial training procedure to remove information about the sensitive attribute from the latent representation learned by a neural network and empirically study how different data distributions use in the adversarial learning affect the resulting fairness of the model. [12] also explore how fairness depends on difference in data distributions.

In the case of recidivism prediction, many models are biased by selective labeling. Several works [22, 31, 41] have looked at the effect of selective labels in the training data on algorithmic fairness. [59] derive the conditions under which a bias reversal phenomenon may occur, i.e. the more biased a past decision-maker is against a group, the more the algorithm favors that group when trained on the resulting data. [47] propose a method called contraction that harnesses the heterogeneity of human decision-makers for evaluation of predictive models in the presence of selective labels.

## 3 THE WISCONSIN CIRCUIT COURTS DATASET

We collected data on criminal cases through the API service of Wisconsin Circuit Court Access (WCCA). WCCA was created in 1999, and it contains public case records and docket information from the 72 county courts of the U.S. state of Wisconsin. The original data consists of criminal case dockets from 1970 to 2020. There are around 11 million records, out of which around 2.5 million are criminal. The original data records are public and can be accessed from the Wisconsin Circuit Courts Access (WCCA) web site, https://wcca.wicourts.gov.

The case records include the charges in current offense, the outcomes and sentences of cases and the defendants' demographic information (e.g. sex, race, address, and date of birth). We will use these as model features.

There is additional information available about various events (e.g. hearings, bail decisions, bail amounts etc) for every case. We also have information on the associated attorneys and government officials involved, including prosecutors and judges. This additional information has not been used in this paper's analysis.

### Pre-Processing

We exclude cases that only have forfeiture charges (non-crime), and we exclude the new data after 2018 as we need a 2 year follow-up period to observe recidivism. After pre-processing, we obtain a dataset for recidivism prediction with around 1.5 million convicted criminal cases from 2000 to 2018.

### Base Attributes (Type of Offense, Sex and Race)

From the raw case records, we create a main dataset for recidivism prediction, similar to the COMPAS dataset created by ProPublica that is widely used in the fairness literature but is much smaller in size. Attributes that are directly available through WCCA API are the type of offense (felony, misdemeanor and criminal traffic), defendant sex, and race.

We construct the rest of the attributes and the outcome for prediction from the information that is indirectly available in the case records. We use a combination of first name, last name, and date of birth as a unique identifier for a defendant. This identifier allows us to conduct a search in the database of case records to match the defendant across multiple cases and construct the additional variables. The process of constructing additional variables and the design choices are described as follows.

### Additional Attributes (Prior Criminal Count, Age and Zipcode Level Data)

Using the database search, we obtain the prior count of each of the three crime types - felony, misdemeanor and criminal traffic - of the defendant for each of the cases. We were able to collect cases from as early as 1970. We use all these case records for constructing the prior criminal count. However, the records in earlier years tend to be incomplete and compared to later years, the number of cases are much smaller. We analyze the impact of this issue further in Section 4.4. We also infer age at judgment and age at first offense for each case. Age at judgment is calculated based on the date of birth of the defendant and judgment disposition date of the case. Age at first offense for each case is the age when the defendant committed the first crime found in the database. We further merge 9 local demographics variables to our data on zipcode from a zipcode level dataset processed from 2010 census data [6], including population density, proportion who attended college, proportion eligible for food stamp, African American population share, Hispanic population share, proportion of male, proportions who live in rural and urban area and median household income.

### Target Variable (Outcome)

The target variable for prediction (or the outcome variable) of interest is whether the defendant recidivates or not. However, the performance of the machine learning models (including fairness) is sensitive to the precise definition of the outcome variable [26]. Therefore, it is important for the decision support system's designer to select the right prediction task with domain knowledge that reflects the real goal of the tool. While defining the outcome variable, we face a set of design choices that are discussed below.

***Follow-up Period.*** We had to decide on a follow-up period within which committing a new offense is deemed as recidivism. ProPublica [39] defined recidivism as a new offense within a two year period, mainly because Northpointe, the company that designed the COMPAS tool, indicated that its recidivism score was based on that timeline. Further, a study [36] by the U.S. Sentencing Commission showed that most recidivists reoffend within two years after release (if they reoffend at all). In this paper, we follow this choice of 2 year follow-up period. But with our data, it is possible

**Table 1: Summary of Wisconsin Circuit Courts Dataset**

| | Full sample | Caucasian | African American | Hispanic | Native American | Asian |
|---|---|---|---|---|---|---|
| *Sample size* | 1,476,967 | 964,922 | 333,036 | 101,607 | 63,862 | 13,540 |
| *Sample share* | | 65.33% | 22.55% | 6.88% | 4.32% | 0.92% |
| Recidivism (if observed) | 42.21% | 40.34% | 46.43% | 38.76% | 56.47% | 37.80% |
| *Sex* | | | | | | |
| Male | 80.40% | 79.05% | 83.47% | 88.88% | 69.65% | 87.57% |
| *Age* | | | | | | |
| Below 30 | 51.38% | 49.45% | 54.13% | 56.91% | 53.71% | 68.60% |
| 30 to 60 | 47.44% | 49.09% | 45.17% | 42.61% | 45.58% | 30.85% |
| *Case type* | | | | | | |
| Felony | 32.18% | 30.76% | 39.98% | 21.09% | 29.80% | 36.39% |
| Misdemeanor | 43.04% | 43.67% | 43.14% | 34.12% | 47.55% | 40.89% |
| Criminal Traffic | 24.78% | 25.57% | 16.88% | 44.79% | 22.66% | 22.73% |

to define other follow up periods that can be used to study the implications for fairness.

***Decision Making Stage.*** Another aspect we considered is the stage at which the risk estimate is intended to be used. Risk assessment tools can be applied at all stages in the criminal justice process: from pretrial, sentencing to parole planning. The prediction task for each stage can be different. For this paper, we assume that the recidivism risk score is intended to inform the judges at the stage of sentencing, therefore we decide that the task is to predict whether the defendant will commit a crime within two years *from the date of judgment*. ProPublica assumed a pre-trial stage [39].

***Missing Outcomes due to Incapacitation.*** Finally, we had to define which cases have missing outcomes in our dataset. It is often true that the observed data is a consequence of previous human decisions. In our case, whether we observe recidivism or not is affected by the decisions of judges. In some settings, it is relatively straightforward to identify cases with a missing outcome from the original data. For example, drivers that are not searched by the police or defendants who are denied bail. However, in our prediction task, identifying cases that have a missing outcome is not trivial because defendants serve different sentence lengths. Assigning a missing outcome to every case with a sentence throws away a lot of useful data.[1] Yet extending the follow up period for two years after the assigned sentence period instead of the judgment date is also problematic because defendants often serve more or less than the assigned sentence. Since there is not a comparable data source that has the exact jail record of every defendant in Wisconsin, we don't observe the actual sentence length. Moreover, the sentence itself could affect probability of recidivism. Further, the defendants who receive sentences are a selected group, so there is the issue of selective labeling explored by Lakkaraju et al. [47].

There is no consensus in the literature about how to deal with this problem. We assess the importance of these decisions as follows. We use a cutoff for sentence length, of 180 days, such that we don't have to throw away a lot of useful data and still leave enough time in the follow-up period for the defendant to reveal crime potential. Above this cutoff, we treat the defendants' outcome as missing (and hence dropped from the dataset) even if they do not reoffend within the follow-up period. We explore this issue further in Section 4.1.

[1]To see this, consider three defendants who stay in jail for one year, for 22 months, and for 2 years after judgment. The first defendant has 1 year left to reoffend in the follow-up period, the second has only 2 months, and for the third, we can not observe recidivism in the follow-up period of 2 years at all.

## Descriptive Statistics

We have case records from as early as 1970, but the records in earlier years tend to be incomplete and the number of cases much smaller. Therefore, we only keep the cases from 2000 for further analysis. It only means that the rows in the dataset that we use for training and testing machine classifiers are only those cases that appear in the courts from 2000. The pre-2000 information for such cases is still included in the form of prior criminal count of defendants. Given the 2 year follow-up period, we exclude cases that are disposed after 2018 since there is not enough time to observe recidivism. We also excluded dismissed cases that do not result in conviction. We also had to delete records of defendants from the main dataset that do not have sex and/or race data available. Finally, we exclude cases that only have forfeiture (non-crime) charge. Table 1 presents summary statistics of the main dataset thus constructed with around 1.5 million cases from 2000 to 2018. There are five race groups in the dataset, with around 65% Caucasian, 23% African American and 7% Hispanic. The 'Asian or Pacific Islander' and 'American Indian or Alaskan Native' groups are abbreviated as 'Asian' and 'Native American' respectively. The proportion of male criminals (80%) is significantly higher than female, and most cases are committed at a younger age (below 60). The recidivism rate is the highest (~56%) with Native American. Misdemeanors are the most frequent crime type except for Hispanic with criminal traffic (45%) being the most common crime type.

## Other Datasets in Criminal Justice

Our dataset provides a valuable complement to the standard datasets used in the literature on algorithmic fairness of recividism predictions. The standard COMPAS data set used in algorithmic fairness literature, assembled by ProPublica [39], has 7000 observations from a single court (Broward County, Florida) over two years (2013 and

2014). It is limited to the set of defendants assessed with COMPAS at the pre-trial stage. The dataset does not include information on judges or other officials.

Another set of papers have used datasets from the court systems of large cities, but they have only rarely been used in algorithmic fairness literature. [2] examine racial disparities in pretrial bail decisions using data on 163K cases from Philadelphia, 2010-2014, and 93K cases from Miami-Dade, 2006-2014. The dataset in [43] includes all arrests made in New York City, 2008-2013, adding up to 758K observations. [3] also use a dataset from NYC, with 595K cases from 2008 through 2013. The outcome in these three data sets are pretrial misconducts such as failure to appear or new arrest before case disposition, whereas ours are recidivism after case disposition.

There are other datasets in the broader context of criminal justice used to examine the behavior of judges. [51] analyze a dataset of all convicted felony crime cases in Texas, 2004 to 2014, with around 440K cases. They supplement this dataset with judge's ethnicity, gender and partisanship to study how those characteristics may affect sentencing decisions. [4] analyze 5 million criminal case records from 2010-2018 in Indian criminal courts to examine the in-group bias of judges. This dataset includes not only convicted cases but also acquitted cases. Finally, [5] use a dataset with 1 million criminal sentencing decisions from U.S District Courts, 1992-2011. These data sets do not code recidivism measures.

# 4 ANALYSIS

This section analyses the relevance of various data related factors in the performance of unconstrained (fairness unaware) machine learning classifiers, from both accuracy and fairness perspectives.

There are many classifiers that could be used for prediction. We consider two popular classifiers: the logistic regression (LR) classifier and the XGBoost classifier. These models represent linear and non-linear base classifiers respectively.

For evaluation metrics, we use average accuracy as well as area under the ROC curve (AUC). While accuracy depends on the decision threshold applied to the probabilistic risk scores, the AUC metric is independent of threshold.

We use four fairness metrics: false positive rate (FPR) difference, false negative rate (FNR) difference, positive rate (PR) difference, and bias amplification. FPR difference quantifies the difference in FPR for two groups in the population – i.e., the difference in the fraction of people who were classified as high risk but didn't recidivate. Similarly, FNR difference quantifies the difference in FNR for two groups in the population. Higher FPR and/or lower FNR for one group compared to the other is considered unfair. PR difference quantifies the difference in the fraction of people who are classified as high risk in the two groups. This definition ignores actual recidivism.

Finally, bias amplification quantifies the difference in the PR difference (with sign) between two groups in the decisions of the classifier and the recidivism rate (or base rate) difference (with sign) of two groups in the data. It is often measured w.r.t base rate difference in the training data. However, generally, train and test data are assumed to be from same distribution. We don't make such assumption in our analysis and therefore, always measure bias amplification w.r.t. base rate difference in the test distribution. The formal details of all these standard metrics are provided in Appendix A.

Even though the fairness gap is higher for other pairs of identity groups in our dataset, for brevity and following earlier works, we focus mostly on the African American and Caucasian groups in further sections. These are also the two biggest racial groups in the the dataset. For FPR difference, we report the FPR for African American minus FPR for Caucasian. For FNR difference, we report the FNR for Caucasian minus FNR for African American. For PR difference, we report PR for African American minus PR for Caucasian. Thus, positive values of these three metrics imply African American is the disadvantaged group. Unless otherwise stated, accuracy and AUC reported in various graphs in the paper are not group specific – they are for all five groups in the data combined.

## 4.1 Recidivism Prediction Performance on Wisconsin Dataset

The predictors included in the models were sex, type of offense, prior criminal count (for each type), and age (at judgment and at first offense). LR and XGBoost classifiers were implemented with Python's scikit-learn [57] and XGBoost [18] packages respectively. LR learns a linear model while XGBoost learns a decision tree ensemble. We split the entire data into 70% train and 30% test, and the results are reported on the test data. For LR classifier, we include L2 regularization and select the regularization parameter via 10-fold cross-validation. For XGBoost, we include both L1 and L2 regularization and tune the hyperparameters via grid search and 5-fold cross-validation. 95% confidence intervals are constructed by multiple train and test splits.

Table 2 shows the performance of the LR classifier and the XGBoost classifier on this dataset across all five identity groups. We observe from the table that Native American is the most disadvantaged group in the sense that it has the highest FPR and PR, and the lowest FNR. The largest FPR difference is around 13%, the largest FNR difference is around 18%, and the largest PR difference is around 20%, all between Native American and Hispanic with XGBoost classifier. Hispanic and Caucasian groups receive the most favorable decisions (1-PR) in both models, followed by Asian and African American groups. The overall accuracy of XGBoost and LR is not very different but the FNR, FPR and PR differences between groups is higher for XGBoost.

We also examined an alternate way of defining the target variable (outcome) by varying the sentence length cut-off from 180 days to 2 years, and extending the follow-up period of 2 years by adding the sentence length. We didn't find much difference in the recidivism rates by group or the classifiers' performance. However, we include more details and all the results in Appendix B.

## 4.2 The Role of Training Data Size

***Motivation***. More data is often thought to be the solution for many problems in machine learning. We study to what extent, only collecting more training samples can reduce unfairness in decisions of unconstrained classifiers.

***Experiment Design/Settings***. We reserve 20% of the data and from the remaining 80% data, we create training datasets of sizes

**Table 2: Recidivism Prediction Performance on Wisconsin Dataset, with 95% Confidence Intervals**

| | Overall | Caucasian | African American | Hispanic | Native American | Asian |
|---|---|---|---|---|---|---|
| *XGBoost* | | | | | | |
| Accuracy | 0.6588 ± 0.0018 | 0.6648 ± 0.0020 | 0.6459 ± 0.0034 | 0.6567 ± 0.0055 | 0.6303 ± 0.0076 | 0.6760 ± 0.0108 |
| AUC | 0.7039 ± 0.0018 | 0.7044 ± 0.0023 | 0.7033 ± 0.0035 | 0.6719 ± 0.0068 | 0.6878 ± 0.0090 | 0.7079 ± 0.0122 |
| FPR | 0.2244 ± 0.0041 | 0.2159 ± 0.0042 | 0.2454 ± 0.0048 | 0.2016 ± 0.0073 | 0.3272 ± 0.0145 | 0.2203 ± 0.0147 |
| FNR | 0.5008 ± 0.0040 | 0.5113 ± 0.0045 | 0.4792 ± 0.0067 | 0.5674 ± 0.0096 | 0.4025 ± 0.0106 | 0.4944 ± 0.0225 |
| PR | 0.3405 ± 0.0035 | 0.3261 ± 0.0038 | 0.3734 ± 0.0041 | 0.2910 ± 0.0060 | 0.4800 ± 0.0100 | 0.3283 ± 0.0128 |
| *Logistic Regression* | | | | | | |
| Accuracy | 0.6461 ± 0.0013 | 0.6560 ± 0.0020 | 0.6206 ± 0.0027 | 0.6535 ± 0.0048 | 0.6026 ± 0.0057 | 0.6735 ± 0.0116 |
| AUC | 0.6822 ± 0.0013 | 0.6825 ± 0.0022 | 0.6806 ± 0.0030 | 0.6558 ± 0.0056 | 0.6746 ± 0.0068 | 0.6946 ± 0.0136 |
| FPR | 0.1532 ± 0.0022 | 0.1479 ± 0.0026 | 0.1667 ± 0.0031 | 0.1292 ± 0.0051 | 0.2395 ± 0.0081 | 0.1386 ± 0.0118 |
| FNR | 0.6282 ± 0.0029 | 0.6334 ± 0.0035 | 0.6244 ± 0.0039 | 0.6903 ± 0.0085 | 0.5189 ± 0.0080 | 0.6350 ± 0.0199 |
| PR | 0.2456 ± 0.0021 | 0.2363 ± 0.0024 | 0.2638 ± 0.0023 | 0.1991 ± 0.0049 | 0.3761 ± 0.0065 | 0.2243 ± 0.0106 |

1000, 10000, 100000 and 500000 using random sampling with replacement.[2] We train machine learning models on each of the sampled training datasets and report the models' performance on the reserved set. The predictors are sex, type of offense, prior criminal count (for each offense type) and age (at judgment and first offense).

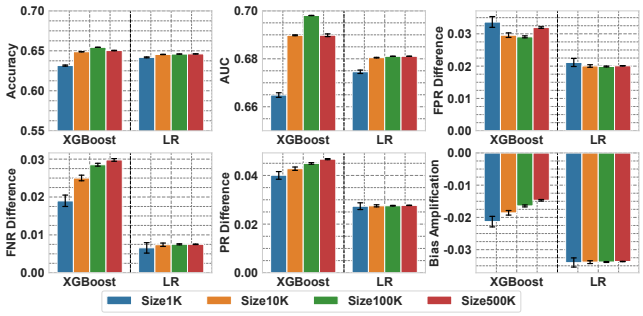


**Figure 1: Role of Training Data Size: X axis shows the two classifiers, Y axis corresponds to the mean of each metric from 300 random samples with error bars denoting 95% CI.**

***Results***. Figure 1 shows that increasing the size of the training data alone does not change anything for the LR classifier (neither predictive performance nor fairness). For XGBoost, the smallest train size (1000) leads to highest FPR difference but also the smallest FNR and PR differences. This can perhaps be attributed to the marginal change in overall predictive performance of XGBoost with training data size. Any differences observed are usually less than 1%. Negative values of bias amplification mean that the PR rate difference (with sign) in classifier's decisions is lower than the base rate difference (with sign) in the test data. In most of our results in the paper, both PR rate difference and base rate difference is positive (i.e. African Americans are the disadvantaged group). In this case, a negative bias amplification value means that the classifier reduces the bias against African Americans but the direction of bias does not change to be against Caucasians. By decomposing cost-based metrics into bias, variance and noise, [17] analyze when bigger training data size helps; but as our results indicate, bigger training data doesn't always lead to fairer models.

[2]We also performed same experiments using random sampling without replacement, and obtained similar results.

## 4.3 The Role of Group Proportions and Base Rate Difference

***Motivation***. The training datasets are often imbalanced in the sense that they have different proportions of the demographic groups. This may be either due to imbalance in the underlying demographic distribution or it may be a result of data collection or sampling strategy. Using imbalanced data for machine learning can disproportionately affect different groups for a number of reasons. For example, the average loss that the classifier optimizes may be dominated by the majority group or the data in the minority group may not be enough to learn the decision boundaries for that group as good as for the majority group. A related factor is the observed difference in base rates of the two groups i.e. the rate at which people in two groups recidivate in the training data. Again, this can be due to a number of factors in the real world or it may be a result of data collection or sampling strategy. Various theoretical results point out the implications of base rates being different for fair decision-making [20, 21, 44]. We empirically study the extent to which the imbalance in group proportions and base rates have an effect on the fairness and accuracy of unconstrained classifiers.

***Experiment Design/Settings***. We reserve 20% of the original data. From the remaining 80% data, we construct five settings of data that follow five different distributions w.r.t. group proportions and base rate difference between African American and Caucasian groups. We randomly remove some data samples to achieve the desired distributions. The random removal of data is repeated 30 times in each setting in order to estimate the variance due to randomness. The predictors included in the classifiers were same as in Section 4.2.

The five distribution settings that we simulate are as follows:

**1. Fewer Caucasian:** The number of samples from the Caucasian group are 32% of the number of samples from the African American

group in new training data, which is the opposite relationship compared to the original data. This is achieved by randomly removing some of the data points from the Caucasian group.
**2. Equal Size**: The number of samples from the Caucasian group are equal to the number of samples from the African American group in new training data. This is achieved by randomly removing some of the data points from the Caucasian group.
**3. Equal Base Rates:** The recidivism rate is the same for Caucasian and African American groups. This is achieved by randomly removing some of the data points from the African American group that were observed to recidivate in the dataset.
**4. Higher Base Rate Difference** The base rate difference between Caucasian and African American groups is increased to 12% in the new data, compared to 6% in the original data. This is achieved by randomly removing some of the data points from the Caucasian group that were observed to recidivate in the dataset.
**5. Balanced Outcomes:** For both recidivism or non-recidivism outcomes, the number of Caucasian and African American are the same. This is achieved by randomly removing some of the data points from the Caucasian group with both kinds of outcomes.

***Results***. We report the results for two types of test data. The first type is when the test data follows the same (S) distribution as the train data (using 70/30 split). The second type is when the test data follows the target (T) distribution – i.e., the 20% data that we had reserved before applying any of the random data removal settings described above. The reason for keeping this 20% data in evaluation is to simulate the real-world settings in which a data scientist may control the data curation process for the training data but that may not change anything in the target population where the model will eventually be applied. This crucial evaluation aspect is often neglected and evaluation is only done on test data that comes from the same distribution as training.

Figure 2 shows fairness and performance metrics for the five training data settings described above. The first observation here is that the trend of XGBoost being more unfair without being significantly more accurate, holds across all settings. Second, the unfairness in the balanced outcome is the highest. This is perhaps surprising because balanced outcomes ensures the base rates and the sizes of Caucasian and African American groups are equal. Other training sets also influence fairness in interesting ways. It would be interesting to theoretically analyze these in further detail in future work. For FPR, FNR, and overall AUC/accuracy, the results appear robust across S and T test distributions. Not surprisingly, PR (and bias amplification) is different between S and T as the base rates in the test distribution also change. Readers may want to use Table 2 for comparison with baseline (full) sampling. Bias amplification and overall accuracy plots are in Figure 9 in Appendix C.

## 4.4 The Role of Time

***Motivation***. The nature of data changes significantly over time due to various developments in the real world. The feature distributions, the outcome distribution, and/or the conditional distribution of outcomes given the features may change over time.

Moreover, the data curation process itself may create an unnatural shift. For example, we have case records from as early as 1970 until 2020 but earlier case records tend to be incomplete. Due to

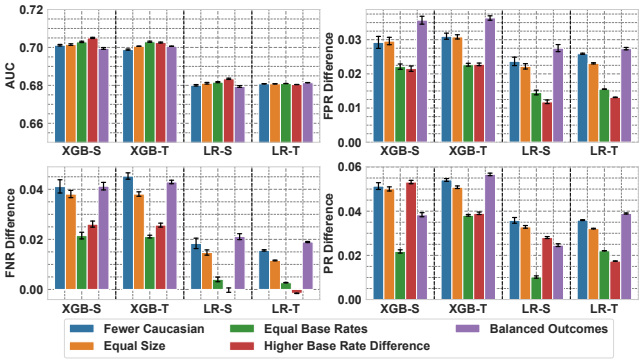


**Figure 2: The Role of Group Proportions and Base Rate Difference in Training Data: X axis shows the model and test distribution pair, where S denotes the same (as train) distribution and T denotes the target distribution. Y axis shows mean of each metric from 30 random samples (with 95% CI).**

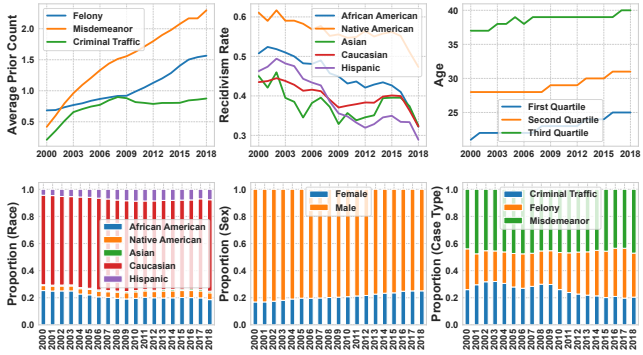


**Figure 3: Change in Data Distribution with Time**

this, the prior criminal count of the defendants that we inferred from the original database may be underestimated for earlier cases in our dataset. Figure 3 shows the average prior count of the defendants in each year for each type of offense. There is about 4 x difference in the average prior count in the years 2000 and 2018. This doesn't necessarily reflect a shift in the prior count in the real world but perhaps the fundamental limitation of the way such datasets can be constructed. Overall, the performance of machine learning models trained on this dataset will be constrained due to apparently changing relationships between prior count and the outcome variable.

To illustrate such changes, Figure 3 also shows the recidivism rate for different races over the years. Overall recidivism rate and the recidivism rate difference between groups is decreasing over time. We also observed marginal shifts in the distribution of offense types, group proportions, sex, and age. This motivates us to study the role of temporal factors in the dataset on accuracy and fairness of the unconstrained classifiers.

***Experiment Design/Settings***. We train the model on data from two consecutive years and test it on the data from the subsequent two years (with a two year gap between train and test, since we

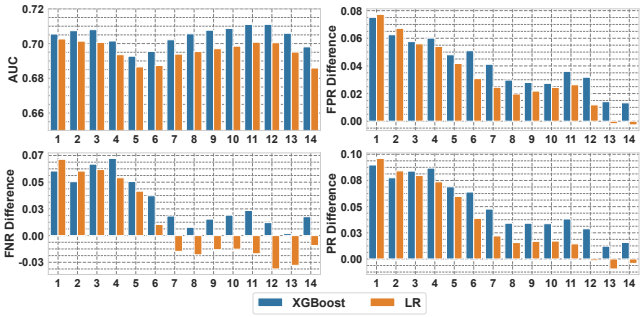


**Figure 4: The Role of Time: X axis corresponds to training datasets from two consecutive years between 2000 and 2018 (e.g., "1" on x axis denotes the training data from the years 2000 & 2001, "2" denotes the training data from the years 2001 & 2002 and so on. The test data comes from the next two years after a two year gap (e.g. if training data is from 2000 & 2001, test comes from 2003 & 2004.)**

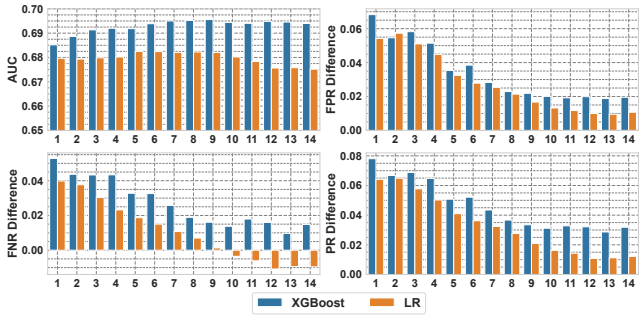


**Figure 5: The Role of Time: X axis corresponds to training datasets, as described in the caption of Figure 4 caption. Main difference is that the test data in this figure is the reserved data from all the years between 2000 and 2018.**

need two years to observe the outcome). For example, if we train on the data from 2010 and 2011, we test the model on the data from 2013 and 2014. This simulates the real-world setting where data from the past is used for training a model, which is then applied to make decisions in the future. The reasons for limiting past data to two years are, for example, limiting the influence of older data that may not be relevant anymore, and unavailability of older data.

We repeat this approach in a moving-window manner between 2000 and 2018, and thus obtain 14 different training sets. We train 14 different models for both LR and XGBoost. In addition to reporting models' performance on the data from two subsequent years, we also report the performance when we apply each of these models to a reserved data (20% of entire data) that includes all years, to see the difference. This latter case simulates the settings where a model may be trained on data from a specific period of time and applied across different time periods. The predictors are the same as in Section 4.2.

***Results***. The results are reported in Figures 4 and 5 for the two test settings respectively described above. In both cases, we observe that while overall AUC is stable across years, the fairness metrics change significantly. For example, FPR difference in Figure 4 goes from 8% in 2000 to less than 2% in 2018. This shows how unfair machine decisions can be in different points in time. In Figure 4, when the test data distribution is fixed and only training data changes, we observe that models trained on older data are much more unfair. This is true not only for FPR and FNR differences, but also for PR difference. It is also interesting to note that for the LR classifier, the FNR difference changes sign in later years (going against the Caucasian group).

The results show that temporal factors in data generation and curation should be more explicitly considered in empirical research on algorithmic fairness. For completeness, bias amplification and overall accuracy plots are also in Figures 10 and 11 in Appendix D.

## 4.5 The Role of Race and Zipcode Level Data

***Motivation***. Race is the protected attribute according to which we define fairness. It is often the case that such protected attributes are hard to obtain in the data (at training and/or test time) or are simply not allowed to be used in the decision-making process. Previous works have shown that excluding race from the machine learning model is not enough to ensure fairness due to correlation with other attributes. On the contrary, it is known [21] that when base rates differ, information about the sensitive attribute is necessary for optimal decision making under fairness constraints.

Further, neighborhood and zipcode level data are often correlated with race. We are interested in understanding whether including detailed zipcode level data could in fact improve fairness (by removing confounders for example) in the base classifiers. As already discussed in Section 3, these zipcode level variables include population density, proportion who attended college, proportion eligible for food stamps, African American population share, Hispanic population share, proportion male, proportions who live in rural and urban areas, and median household income.

***Experiment Design/Settings***. In all experiments, we include sex, type of offense, prior criminal count (for each offense type), age at judgment, and age at first offense as predictors for both LR and XGBoost classifiers. Depending on the experiment setting, we also include race and/or zipcode level variables in the set of predictors, and train the model to observe the effect on accuracy and fairness. We thus have four combinations of predictors depending whether we include race, zipcode level variables, neither, or both.

***Results***. Figure 6 shows the fairness and performance metrics for each set of predictors. We can make two interesting observations. First, including zipcode level variables improves fairness for all four fairness metrics and both types of classifiers. This is different from the finding in Jabri [37], where in the COMPAS dataset including neighborhood information reduced fairness.

Second, including race as a predictor has a significant negative effect on fairness. In our dataset, giving the model information on race increases the proportion of errors that disadvantage black defendants. We note, however, that this observation should be interpreted with caution. Including race as a predictor may not always

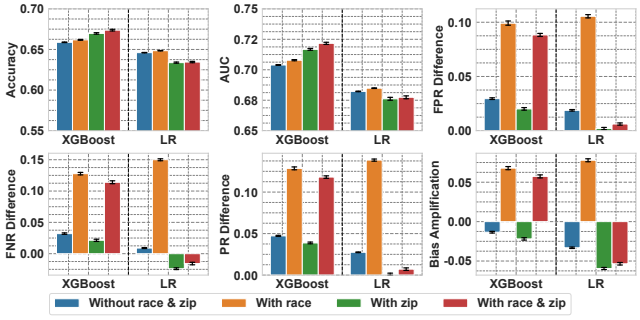


**Figure 6: The Role of Including Race and Zipcode Level Data as Additional Predictors in LR and XGBoost. 95% confidence intervals are constructed by multiple train and test splits.**

lead to more unfairness, as it depends on the data generation mechanism (see e.g. [19] for a causal graphs based analysis of this issue). For overall accuracy and AUC, there is a marginal improvement in the case of XGBoost with more predictors.

## 4.6 Separate Classifiers Trained on Data from Different Races and Offenses

***Motivation.*** Instead of merely adding race as one of the predictors, one could train separate classifiers for each racial group, using only data from the respective racial group. On one hand, separate classifiers get more flexibility and can minimize loss for each group separately (following a best-effort principle). This can benefit minority groups compared to a joint classifier that minimize loss for the overall population dominated by the majority group. On the other hand, by training separate classifiers, we also reduce the amount of data that is used for training each of the separate classifiers. This may hurt both groups compared to a joint classifier trained on overall more data, since there could be shared predictive relationships across the groups.

Similar arguments apply to the idea of training different classifiers by offense type.

***Experiment Design/Settings.*** We train separate classifiers for each race and for each offense type. Before dividing data by race/offense type, we reserve 20% of the overall dataset that represents the general distribution. We divide the remaining data by race/offense type. For each of these new datasets, we then have a 70/30 train and test split. We thus have five separate classifiers for each of the five race groups, and three classifiers for each of the three offense types. For race-specific classifiers, we include sex, offense type, prior criminal count (for *each* offense type), age at judgment, and age at first offense as predictors. For offense-specific classifiers, we don't need offense type as a predictor. Race is not included as a predictor in this section, but race-specific classifiers already include race information since the classifiers are different for the two races.

***Results.*** *1) Race-Specific Classifiers:* Table 3 shows the performance of the two classifiers trained on data from African American and Caucasian groups, respectively. Table 6 in Appendix E shows the same table with 95% confidence interval. These classifiers were applied on test datasets from their respective groups.[3] For comparison, we include the performance metrics of the joint classifier that was trained on data from all groups, and applied on data from each group. For brevity, we only show the performance metrics on the two groups. The joint classifier did not include race as a predictor.

Compared to the joint classifier, race-specific classifiers (both XGBoost and LR) decrease FPR and PR and increase FNR for the Caucasian group. At the same time, they increase FPR and PR and decrease FNR for the African American group. The accuracy and AUC increases only marginally for the African American group with a race-specific classifier, but stay almost the same for Caucasian. This suggests that the African American group may benefit in terms of fairness (error rate difference) when we include patterns in the data from the other group in a joint classifier. A separate classifier for the minority doesn't necessarily make the group better-off.

**Table 3: Performance of Joint Classifier (Trained on Data From All Racial Groups, Without Race as Predictor), Compared to Separate Classifiers (Trained on Race Level Data)**

| | LR | | XGBoost | |
|---|---|---|---|---|
| | Caucasian | African American | Caucasian | African American |
| *Joint Classifier* | | | | |
| Accuracy | 0.6560 | 0.6206 | 0.6648 | 0.6459 |
| AUC | 0.6825 | 0.6806 | 0.7044 | 0.7033 |
| FPR | 0.1479 | 0.1667 | 0.2159 | 0.2454 |
| FNR | 0.6334 | 0.6244 | 0.5113 | 0.4792 |
| PR | 0.2363 | 0.2638 | 0.3261 | 0.3734 |
| *Af. Am. Classifier* | | | | |
| Accuracy | 0.6494 | 0.6363 | 0.6486 | 0.6518 |
| AUC | 0.6741 | 0.6834 | 0.6855 | 0.7087 |
| FPR | 0.2472 | 0.2532 | 0.2818 | 0.2945 |
| FNR | 0.5043 | 0.4913 | 0.4549 | 0.4102 |
| PR | 0.3471 | 0.3718 | 0.3877 | 0.4315 |
| *Caucasian Classifier* | | | | |
| Accuracy | 0.6526 | 0.6126 | 0.6652 | 0.6352 |
| AUC | 0.6827 | 0.6778 | 0.7043 | 0.6921 |
| FPR | 0.1280 | 0.1453 | 0.1975 | 0.2284 |
| FNR | 0.6711 | 0.6676 | 0.5375 | 0.5228 |
| PR | 0.2091 | 0.2320 | 0.3046 | 0.3437 |

*2) Offense Type Specific Classifiers:* The results for offense-specific classifiers are shown in Figure 7. The results were obtained by applying the separate classifiers for a given offense type on the test data from the respective offense type. For comparison, we also include the performance of the joint classifier on each offense type. The first notable observation in Figure 7 is that, even for the joint classifier, most unfairness exists in the criminal traffic offense type.

[3]Ideally, group-specific classifiers would be applied on the respective group only. For completeness, we include performance metrics of the classifiers when applied on reserved test data from the other group as well. These numbers have been greyed out in the table to avoid confusion.

For felony and misdemeanour offense, unfairness is relatively small and is often in the reverse direction (i.e. African Americans are not the disadvantaged group). When we train separate classifiers by offense type, this trend continues. We also observe that for criminal-traffic classifiers, the unfairness is less than the unfairness in the joint classifier. The overall accuracy and AUC are not very different for joint and separate classifiers for any of the three offense types.

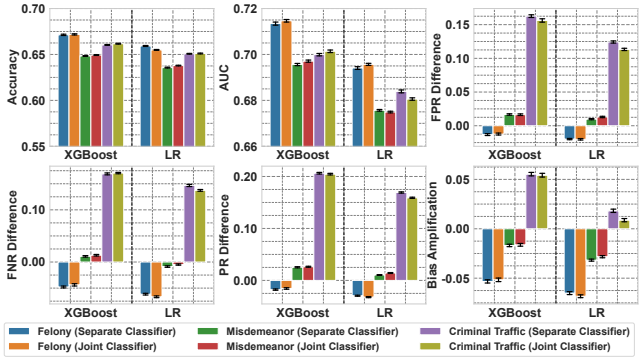


**Figure 7: Offense Type Separated Classifiers: A separate classifier for each offense type is trained on data from that offense type. The performance of the classifiers are then observed on respective offense types. For comparison, the performance of a joint classifier, that is trained on all the data and uses offense type as a predictor, is also shown by offense type. 95% confidence intervals are constructed by multiple train and test splits.**

Figure 8 shows the results when separate classifiers are applied on the reserve set (data from all offense types). We observe that the classifier trained only on criminal traffic data is more unfair overall compared to the joint classifier and other separate classifiers.

## 5 CONCLUSIONS AND FUTURE WORK

While it is well-known that data-centric factors influence algorithmic fairness in complex ways, research in algorithmic fairness is still very much algorithm-centric and depends on a few benchmark datasts for limited empirical evaluation. In this paper, we take a rigorous empirical approach towards understanding the effect of various data-centric factors on algorithmic fairness. We construct a new large scale dataset for recidivism prediction and show how it can be be used for the purposes of data-oriented analysis. Our results suggest that such data-centric factors should be explicitly taken into account while designing and evaluating algorithms. We also observe that the effect of data-centric factors on fairness metrics look very different from their effect on accuracy metrics in all the experiments, further suggesting that data-centric factors in fairness deserve independent attention.

In this paper, we studied data-centric factors like the training data size, group proportions and base rate difference, temporal aspects of data curation, the availability additional variables, data from difference offense types etc. A further theoretical analysis will be useful to understand the observations more. It will also be interesting to extend the idea further by studying other data-centric

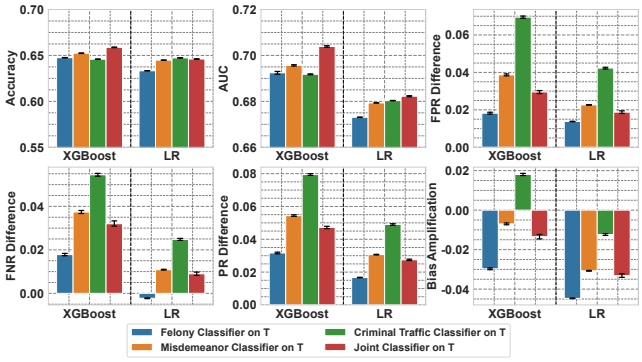


**Figure 8: Offense Type Separated Classifiers Applied to a Target Distribution with All Offense Types: Offense specific classifiers are trained with data from respective offense types only. The performance of the classifiers are then observed on reserved target distribution with all offense types. For comparison, the performance of a joint classifier, that is trained on all the data and uses offense type as a predictor, is also shown on the target distribution. 95% confidence intervals are constructed by multiple train and test splits.**

factors not studied in this paper. These may include, for example, the biases due to selective labeling, label noise, geographical factors, judge characteristics in past decisions etc. In the paper, we used two base classifiers (logistic regression and XGBoost). In future work, it may be interesting to study empirically how various data-centric factors interact with common algorithmic design choices, for example, other types of classifiers, whether the classifier is a deterministic or a randomized one, the type of loss function, type of fairness metric, fairness enforcing techniques etc.

Finally, while simulations are an excellent way to understand the fundamentals, the design of simulations should be further investigated in future work to align them even closer to real-world scenarios and improve the external validity of the observations. It would also be interesting to explore how biases in the original (bigger) dataset may also be considered while creating the auxiliary datasets from it and while interpreting the results.

We hope that the paper encourgaes further discussion on open questions such as: 1) what does good data quality mean in the context of algorithmic fairness; 2) what are the appropriate evauation and benchmarking methods for fairness and machine learning algorithms; and 3) what are the complementary theoretical explanations for the behavior of different algorithms under different data situations. A better understanding of these issues can lead to development of fair algorithms that are more robust and suitable for practical deployment.

## 6 ACKNOWLEDGMENTS

Naman Goel acknowledges the generous support provided by ETH Zurich; most of this work was done while Goel was at ETH Zurich. We thank members of the Center for Law and Economics at ETH Zurich for discussion and valuable feedback on this work.

## REFERENCES


[1] Nil-Jana Akpinar, Manish Nagireddy, Logan Stapleton, Hao-Fei Cheng, Haiyi Zhu, Steven Wu, and Hoda Heidari. 2022. A Sandbox Tool to Bias (Stress)-Test Fairness Algorithms. *arXiv preprint arXiv:2204.10233* (2022).

[2] David Arnold, Will Dobbie, and Crystal S Yang. 2018. Racial bias in bail decisions. *The Quarterly Journal of Economics* 133, 4 (2018), 1885–1932.

[3] David Arnold, Will S Dobbie, and Peter Hull. 2020. *Measuring racial discrimination in bail decisions.* Technical Report. National Bureau of Economic Research.

[4] Elliott Ash, Sam Asher, Aditi Bhowmick, Sandeep Bhupatiraju, Daniel Chen, Tanaya Devi, Christoph Goessmann, Paul Novosad, and Bilal Siddiqi. 2021. *In-group bias in the Indian judiciary: Evidence from 5 million criminal cases.* Technical Report. Working paper, August.

[5] Elliott Ash, Daniel L Chen, and Suresh Naidu. 2019. Ideas have consequences: The impact of law and economics on american justice. *Center for Law & Economics Working Paper Series* 4 (2019).

[6] Elliott Ash, Sergio Galletta, Dominik Hangartner, Yotam Margalit, and Matteo Pinna. 2020. The effect of Fox News on health behavior during COVID-19. *Available at SSRN 3636762* (2020).

[7] Carolyn Ashurst, Ryan Carey, Silvia Chiappa, and Tom Everitt. 2022. Why Fair Labels Can Yield Unfair Predictions: Graphical Conditions for Introduced Unfairness. *arXiv preprint arXiv:2202.10816* (2022).

[8] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2021. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. (2021).

[9] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review* 104 (2016), 671–732. https://doi.org/10.15779/Z38BG31

[10] Richard Berk. 2009. The role of race in forecasts of violent crime. *Race and social problems* 1, 4 (2009), 231–242.

[11] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.

[12] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 453–459.

[13] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. *FATML* (2017).

[14] Arpita Biswas, Marta Kolczynska, Saana Rantanen, and Polina Rozenshtein. 2020. The role of in-group bias and balanced data: A comparison of human and machine recidivism risk predictions. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*. 97–104.

[15] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability, and Transparency (FAT*)*.

[16] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. https://doi.org/10.1126/science.aal4230

[17] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *Advances in Neural Information Processing Systems* 31 (2018).

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

[19] Silvia Chiappa and William S Isaac. 2018. A causal Bayesian networks viewpoint on fairness. In *IFIP International Summer School on Privacy and Identity Management*. Springer, 3–20.

[20] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163. https://doi.org/10.1089/big.2016.0047

[21] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *International Conference on Knowledge Discovery and Data Mining (KDD)*. https://doi.org/10.1145/3097983.3098095

[22] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. 2021. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*. PMLR, 2144–2155.

[23] Bo Cowgill and Catherine E Tucker. 2020. Algorithmic fairness and economics. *Columbia Business School Research Paper* (2020).

[24] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 525–534.

[25] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[26] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems* 34 (2021).

[27] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic Fairness Datasets: the Story so Far. *arXiv preprint arXiv:2202.01711* (2022).

[28] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[29] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (2021), 136–143.

[30] Talia B Gillis. 2020. False dreams of algorithmic fairness: The case of credit pricing. *Available at SSRN 3571266* (2020).

[31] Naman Goel, Alfonso Amayuelas, Amit Deshpande, and Amit Sharma. 2021. The Importance of Modeling Data Missingness in Algorithmic Fairness: A Causal Perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[32] Naman Goel and Boi Faltings. 2019. Crowdsourcing with fairness, diversity and budget constraints. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 297–304.

[33] Moritz Hardt and Eric Price. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NIPS)*.

[34] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[35] Hans Hofmann. 1994. UCI Statlog (German Credit Data) Data Set. (1994).

[36] Kim Steven Hunt and Robert Dumville. 2016. *Recidivism among federal offenders: A comprehensive overview.* United States Sentencing Commission.

[37] Ranae Jabri. 2019. Predictive Power at What Cost? Economic and Racial Justice of Data-Driven Algorithms. *Economic and Racial Justice of Data-Driven Algorithms (July 1, 2019)* (2019).

[38] HV Jagadish, Francesco Bonchi, Tina Eliassi-Rad, Lise Getoor, Krishna Gummadi, and Julia Stoyanovich. 2019. The responsibility challenge for data. In *Proceedings of the 2019 International Conference on Management of Data*. 412–414.

[39] Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[40] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 306–316.

[41] Nathan Kallus and Angela Zhou. 2018. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*. PMLR, 2439–2448.

[42] Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *International Conference on Learning Representations*.

[43] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[44] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *Conference on Innovations in Theoretical Computer Science (ITCS)*.

[45] Ronny Kohavi and Barry Becker. 1996. UCI ADULT Data Set. (1996).

[46] Rodrigo L. Cardoso, Wagner Meira Jr, Virgilio Almeida, and Mohammed J. Zaki. 2019. A framework for benchmarking discrimination-aware models in machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 437–444.

[47] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 275–284.

[48] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. 2016. https://github.com/propublica/compas-analysis, 2016. (2016).

[49] Klas Leino, Matt Fredrikson, Emily Black, Shayak Sen, and Anupam Datta. 2019. Feature-Wise Bias Amplification. In *International Conference on Learning Representations (ICLR)*.

[50] Amanda Levendowski. 2018. How copyright law can fix artificial intelligence's implicit bias problem. *Wash. L. Rev.* 93 (2018), 579.

[51] Claire SH Lim, Bernardo S Silveira, and James M Snyder. 2016. Do judge characteristics matter? ethnicity, gender, and partisanship in texas state trial courts. *American Law and Economics Review* 18, 2 (2016), 302–357.

[52] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.

[53] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[54] Fatemeh Nargesian, Abolfazl Asudeh, and HV Jagadish. 2021. Tailoring data source distributions for fairness-aware data integration. *Proceedings of the VLDB*

*Endowment* 14, 11 (2021), 2519–2532.
[55] Andrew Ng. 2021. MLOps: From Model-Centric to Data-Centric AI. *DeepLearning AI* (2021).
[56] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336.
[57] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.
[58] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the Everything in the Whole Wide World Benchmark. (2021).
[59] Ashesh Rambachan and Jonathan Roth. 2020. Bias In, Bias Out? Evaluating the Folk Wisdom. In *1st Symposium on Foundations of Responsible Computing*.
[60] Anna Rogers. 2021. Changing the World by Changing the Data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2182–2194.
[61] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.
[62] Viktor Schlegel, Goran Nenadic, and Riza Batista-Navarro. 2020. Beyond leaderboards: A survey of methods for revealing weaknesses in natural language inference data and models. *arXiv preprint arXiv:2005.14709* (2020).
[63] David Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. 2018. Winner's curse? On pace, progress, and empirical rigor. (2018).
[64] Agnieszka Słowik and Léon Bottou. 2021. Algorithmic bias and data bias: Understanding the relation between distributionally robust optimization and data curation. *arXiv preprint arXiv:2106.09467* (2021).
[65] A TORRALBA. 2011. Unbiased Look at Dataset Bias. *Proc. of IEEE Computer Vision and Pattern Recognition, 2011* (2011), 1521–1528.
[66] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. 2020. From imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*. PMLR, 9625–9635.
[67] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5310–5319.
[68] Michael Wick, Jean-Baptiste Tristan, et al. 2019. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems* 32 (2019).
[69] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 547–558.
[70] Gal Yona, Amirata Ghorbani, and James Zou. 2021. Who's Responsible? Jointly Quantifying the Contribution of the Learning Algorithm and Data. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 1034–1041.
[71] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).

# A FAIRNESS AND PERFORMANCE METRICS

The four definitions of fairness metrics and the two performance measures used in our paper are described here more formally for completeness. Let $Y$ denote the true outcome, $\hat{Y}$ denote the predicted outcome and $A$ denote the sensitive attribute (i.e race). $\hat{Y} = 1$ means that the classifier labels a defendant as high risk for recidivism. $Y = 1$ means that a defendant actually recidivates. There are five race groups in our data. In our results, we focused mostly on Caucasian and African American groups. Therefore, we specify the groups in the below definitions for more clarity. But the definitions can be generalized for any pair of race groups depending on the context.

## Positive Rate(PR) Difference

$$\text{PR difference} = Pr(\hat{Y} = 1 | A = \text{African American}) - Pr(\hat{Y} = 1 | A = \text{Caucasian})$$

When PR difference is zero, *demographic parity* [28] is satisfied.

## False positive rate (FPR) Difference and False Negative Rate (FNR) Difference

$$\text{FPR Difference} = Pr(\hat{Y} = 1 | Y = 0, A = \text{African American}) - Pr(\hat{Y} = 1 | Y = 0, A = \text{Caucasian})$$

$$\text{FNR Difference} = Pr(\hat{Y} = 0 | Y = 1, A = \text{Caucasian}) - Pr(\hat{Y} = 0 | Y = 1, A = \text{African American})$$

When both FPR and FNR differences are zero, *equalized odds* is satisfied [34].

## Bias amplification

Bias amplification quantifies the difference in the PR difference (with sign) between two groups in the decisions of the classifier and the recidivism rate (or base rate) difference (with sign) of two groups in the data. It is often measured w.r.t base rate difference in the training data. However, generally, train and test data are assumed to come from same distribution. We don't make such assumption in our simulations and therefore, always measure bias amplification w.r.t. base rate difference in the test distribution. In the definition below, we assume base rate difference in the ground truth ($Y$).

$$\text{Bias amplification} = \text{PR Difference} - \text{Base Rate Difference in Ground Truth}$$

where

$$\text{Base Rate Difference in Ground Truth} = Pr(Y = 1 | A = \text{African American}) - Pr(Y = 1 | A = \text{Caucasian})$$

## A.1 Accuracy and AUC

The machine learning model used in our paper produce a probability risk score for the defendant to recidiviate. In our paper, we classify defendant above risk threshold 0.5 as $\hat{Y} = 1$ and others as $\hat{Y} = 0$. Accuracy is then defined as follows:

$$Accuracy = \frac{TP + TN}{N}$$

Where TP is the true positives (defendants with $\hat{Y} = 1$ and $Y = 1$), TN is the true negatives (defendants with $\hat{Y} = 0$ and $Y = 0$) and N is the total number of defendants.

AUC stands for "Area under the ROC curve", where the ROC curve is a plot illustrating the trade-off between True Positive Rate and False Positive Rate across all thresholds of the predicted probability risk scores. The X axis is the FPR and the y axis is the TPR in AUC curve, and the AUC is calculated as the area under the ROC curve. The higher the AUC, the better the model is at distinguishing positive and negative cases. AUC provides an overall performance measure of the model regardless of the threshold.

# B THE SECOND OUTCOME DEFINITION

As we discussed in section 3 (Target Variable), there may be multiple ways of defining outcome variable in recidivism prediction. It is interesting to study the role this plays in fairness.

Besides the 2 year follow-up period, one important assumption that we make when encoding missing outcome is the cut-off sentence length. In our main analysis, we used 180 days. Here we encode the missing outcome differently by changing the cut-off sentence length to 2 years.Since in this case a defendant may stayed in jail for most time within the 2 year after judgement and still labeled as non-missing, we decided to extend the 2-year follow-up period by the sentence length for every defendant to account for that. Note that this definition takes into account the sentence length given by the judge, which created even more bias in outcome encoding. But nevertheless it is worthwhile to see the effect of this alternative outcome.

## B.1 Descriptive Statistics of the Dataset with the Second Outcome Definition

We denote the outcome defined in the main text as $Y_{2y}$ and the new outcome defined above as $Y_{2y}$. Table 4 compares $Y_{2y}$ to $Y_{2y}$. Since we used a longer cutoff with $Y_{2y}$, there are fewer missing outcome. However, the recidivism rate among defendants with non-missing outcome does not change much. African American group has the highest rate of missing outcomes, which might suggest that they tend to get longer sentence length than other race groups.

## B.2 Recidivism Classification Performance with Alternate Outcome Definition

Similar to Table 2, Table 5 shows the performance of the LR classifiers and the XGBoost classifiers trained with second outcome definition $Y_{2y}$. The predictors remain the same as before, i.e. sex, type of offense, prior criminal count (for each type of offense), age at judgment and age at first offense.

**Table 4: Statistics of Two Outcome Definitions**

| Missing Outcome Rate | | | | | | |
|---|---|---|---|---|---|---|
| | Overall | Caucasian | African American | Hispanic | Native American | Asian |
| 180 days | 0.0807 | 0.0667 | 0.1286 | 0.0653 | 0.0686 | 0.0767 |
| 2 years | 0.0471 | 0.0367 | 0.0824 | 0.0365 | 0.0374 | 0.0513 |
| **Recidivism Rate (if observed)** | | | | | | |
| | Full sample | Caucasian | African American | Hispanic | Native American | Asian |
| 180 days | 0.4221 | 0.4034 | 0.4643 | 0.3876 | 0.5647 | 0.3780 |
| 2 years | 0.4168 | 0.3996 | 0.4534 | 0.3825 | 0.5593 | 0.3724 |

**Table 5: Recidivism Classifier Performance Metrics, Aggregate and by Group for $Y_{2y}$. With 95% confidence intervals.**

| | Overall | Caucasian | African American | Hispanic | Native American | Asian |
|---|---|---|---|---|---|---|
| *XGBoost* | | | | | | |
| Accuracy | 0.6589 ± 0.0016 | 0.6652 ± 0.0019 | 0.6460 ± 0.0038 | 0.6576 ± 0.0054 | 0.6273 ± 0.0064 | 0.6733 ± 0.0149 |
| AUC | 0.7018 ± 0.0018 | 0.7026 ± 0.0021 | 0.7005 ± 0.0034 | 0.6724 ± 0.0056 | 0.6852 ± 0.0066 | 0.7016 ± 0.0182 |
| FPR | 0.2177 ± 0.0029 | 0.2094 ± 0.0029 | 0.2353 ± 0.0046 | 0.1997 ± 0.0079 | 0.3223 ± 0.0123 | 0.2153 ± 0.0130 |
| FNR | 0.5136 ± 0.0043 | 0.5232 ± 0.0049 | 0.4970 ± 0.0064 | 0.5720 ± 0.0097 | 0.4125 ± 0.0105 | 0.5140 ± 0.0245 |
| PR | 0.3298 ± 0.0030 | 0.3163 ± 0.0032 | 0.3567 ± 0.0039 | 0.2872 ± 0.0076 | 0.4706 ± 0.0091 | 0.3163 ± 0.0116 |
| *Logistic Regression* | | | | | | |
| Accuracy | 0.6457 ± 0.0014 | 0.6555 ± 0.0016 | 0.6215 ± 0.0029 | 0.6540 ± 0.0054 | 0.5988 ± 0.0050 | 0.6728 ± 0.0152 |
| AUC | 0.6784 ± 0.0015 | 0.6794 ± 0.0019 | 0.6760 ± 0.0034 | 0.6533 ± 0.0061 | 0.6699 ± 0.0070 | 0.6878 ± 0.0176 |
| FPR | 0.1478 ± 0.0022 | 0.1430 ± 0.0023 | 0.1571 ± 0.0026 | 0.1293 ± 0.0058 | 0.2358 ± 0.0090 | 0.1339 ± 0.0113 |
| FNR | 0.6429 ± 0.0024 | 0.6469 ± 0.0030 | 0.6453 ± 0.0047 | 0.6948 ± 0.0087 | 0.5316 ± 0.0082 | 0.6521 ± 0.0222 |
| PR | 0.2351 ± 0.0018 | 0.2270 ± 0.0020 | 0.2467 ± 0.0028 | 0.1967 ± 0.0055 | 0.3658 ± 0.0069 | 0.2138 ± 0.0096 |

## C ADDITIONAL RESULTS FOR SECTION 4.3

Figure 9 shows the results including bias amplification and overall accuracy for Section 4.3 (The Role of Group Proportions and Base Rate Difference).

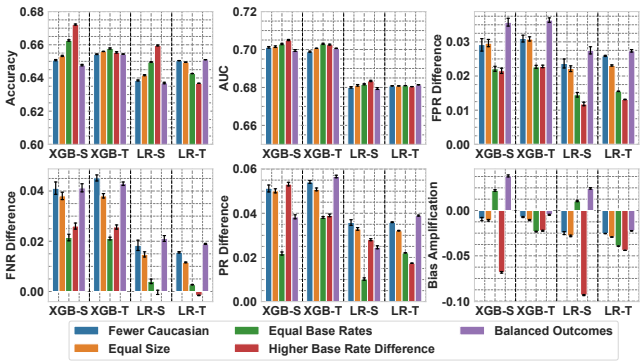


**Figure 9: The Role of Group Proportions and Base Rate Difference in Training Data: X axis shows the model and test distribution pair, where S denotes the same distribution and T denotes the target distribution. Y axis corresponds to the mean of each metric from 30 random samples (with 95% CI).**

## D ADDITIONAL RESULTS FOR SECTION 4.4

Figures 10 and 11 show the results including bias amplification and overall accuracy for Section 4.4 (The Role of Time). Figure 10 and 11 are the complete versions of figures 4 and figure 5 respectively.

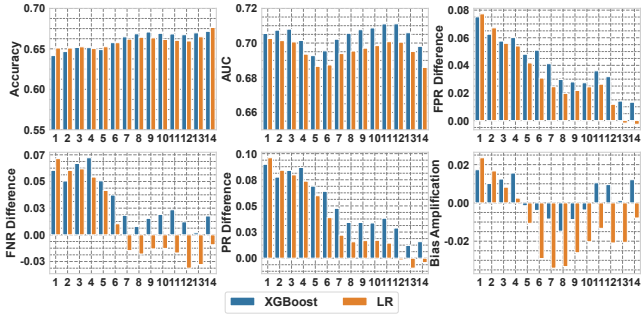


**Figure 10: The Role of Time: X axis corresponds to training datasets from two consecutive years between 2000 and 2018 (e.g., "1" on x axis denotes the training data from the years 2000 & 2001, "2" denotes the training data from the years 2001 & 2002 and so on. The test data comes from the next two years after a two year gap (e.g. if training data is from 2000 & 2001, test comes from 2003 & 2004.)**

**Table 6: Performance of Joint Classifier (Trained on Data From All Groups, Without Race as Predictor) Compared to Separate Classifiers (Trained on Group Level Data. With 95% confidence intervals**

| | LR | | XGBoost | |
|---|---|---|---|---|
| | Caucasian | African American | Caucasian | African American |
| *Joint Classifier* | | | | |
| Accuracy | 0.6560 ± 0.0020 | 0.6206 ± 0.0027 | 0.6648 ± 0.0020 | 0.6459 ± 0.0034 |
| AUC | 0.6825 ± 0.0022 | 0.6806 ± 0.0030 | 0.7044 ± 0.0023 | 0.7033 ± 0.0035 |
| FPR | 0.1479 ± 0.0026 | 0.1667 ± 0.0031 | 0.2159 ± 0.0042 | 0.2454 ± 0.0048 |
| FNR | 0.6334 ± 0.0035 | 0.6244 ± 0.0039 | 0.5113 ± 0.0045 | 0.4792 ± 0.0067 |
| PR | 0.2363 ± 0.0024 | 0.2638 ± 0.0023 | 0.3261 ± 0.0038 | 0.3734 ± 0.0041 |
| *Af. Am. Classifier* | | | | |
| Accuracy | 0.6494 ± 0.0034 | 0.6363 ± 0.0031 | 0.6486 ± 0.0062 | 0.6518 ± 0.0047 |
| AUC | 0.6741 ± 0.0023 | 0.6834 ± 0.0031 | 0.6855 ± 0.0096 | 0.7087 ± 0.0047 |
| FPR | 0.2472 ± 0.0116 | 0.2532 ± 0.0075 | 0.2818 ± 0.0090 | 0.2945 ± 0.0117 |
| FNR | 0.5043 ± 0.0096 | 0.4913 ± 0.0108 | 0.4549 ± 0.0133 | 0.4102 ± 0.0107 |
| PR | 0.3471 ± 0.0106 | 0.3718 ± 0.0085 | 0.3877 ± 0.0087 | 0.4315 ± 0.0099 |
| *Caucasian Classifier* | | | | |
| Accuracy | 0.6526 ± 0.0019 | 0.6126 ± 0.0003 | 0.6652 ± 0.0022 | 0.6352 ± 0.0038 |
| AUC | 0.6827 ± 0.0015 | 0.6778 ± 0.0004 | 0.7043 ± 0.0023 | 0.6921 ± 0.0051 |
| FPR | 0.1280 ± 0.0027 | 0.1453 ± 0.0019 | 0.1975 ± 0.0040 | 0.2284 ± 0.0052 |
| FNR | 0.6711 ± 0.0036 | 0.6676 ± 0.0029 | 0.5375 ± 0.0058 | 0.5228 ± 0.0075 |
| PR | 0.2091 ± 0.0025 | 0.2320 ± 0.0024 | 0.3046 ± 0.0039 | 0.3437 ± 0.0050 |

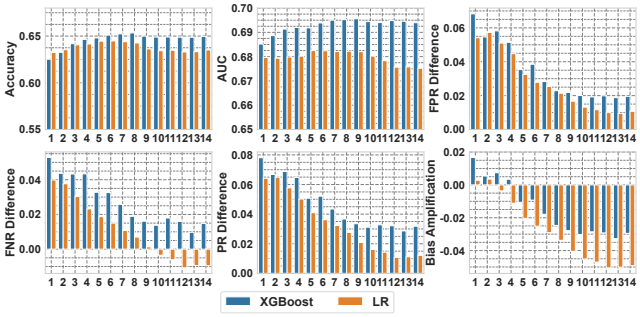


**Figure 11: The Role of Time: X axis corresponds to training datasets, as described in the caption of Figure 10 caption. Main difference is that the test data in this figure is the reserved data from all the years between 2000 and 2018.**

## E ADDITIONAL RESULTS FOR SECTION 4.6

Table 6 shows the same results as table 3, but with 95% confidence interval.