# Predictive Power at What Cost? Economic and Racial Justice of Data-driven Algorithms*

Ranae Jabri†

July 2019

**Abstract**

This paper studies how algorithms use variables to maximize predictive power at the cost of group equity. Group inequity arises if variables enlarge disparities in risk scores across groups. I develop a framework to examine a recidivism risk assessment tool using risk score and novel pretrial defendant case data from 2013-2016 in Broward County, Florida. I find that defendants' neighborhood data only negligibly improve predictive power, but substantially widen disparities in defendant risk scores and false positive rates across race and economic status. Higher risk scores may lead to longer pretrial incarceration and downstream consequences, by impacting labor market outcomes. These findings underscore that machine learning objectives tuned to maximize predictive power can be in conflict with racial and economic justice.

*Keywords:* big data, algorithms, risk assessment, discrimination, inequality, race, crime, recidivism, neighborhoods

*JEL codes:* K40, J15, K14, K42, H0, C53, C55, D8

*I am grateful to Patrick Bayer for his guidance, and Christopher Timmins, Robert Garlick, Seth Sanders, Erica Field and Rachel Kranton for their advice. Thank you also to Dhammika Dharmapala, Jeff Fagan, Brandon Garrett, Jo Hardin, Arnaud Maurel, John Rappaport, Juan Carlos Suarez Serrato, Curtis Taylor, Paul Dudenhefer, and seminar participants and discussants at Duke, 2018 Urban Economics Association Summer School, 2019 UChicago/FAIR Summer School on Socioeconomic Inequality, 2019 Conference on Empirical Legal Studies, 2019 Southern Economic Association Conference, 2020 AEA, UChicago Crime Lab, and 2020 Young Economist Symposium at UPenn for helpful comments and suggestions. I also thank Patrick Bayer and the Duke Economics department for support in acquiring access to data. Any and all errors are my own. First version: May 2018.

†Department of Economics, Duke University. Email: r.jabri@duke.edu.

# 1 Introduction

Data-driven algorithms increasingly inform decision-making across a wide variety of life-changing settings. Conventional data-driven algorithms leverage many input variables to maximize overall predictive power. In doing so, algorithms may include data variables that only marginally improve predictions, without considering whether using these variables introduces group inequity. Group inequity can result if input variables widen disparities in risk scores across groups. In this paper, I study how algorithms maximize predictive power at the cost of group equity, specifically in recidivism risk assessment tools. While I examine one commonly-used tool, such trade-offs can be found in algorithms that predict other outcomes in the criminal justice system and in society more broadly.

Economic and racial inequity in tools can present legal and ethical concerns, by introducing group inequity. Group inequity arises if variables enlarge disparities in risk scores across groups. Historically disadvantaged groups may disproportionately bear the large costs of only a marginal increase in predictive power. Predicting risk scores that are disproportionately higher for certain groups treats individuals differently by their group affiliation. Differential treatment by race is of particular concern, as race is a protected class. Many such algorithms do not use race or other protected classes explicitly because of concerns about discrimination.[1] Yet variables such as residential location which can proxy for race and socioeconomic status can be used in tools that are used in decision-making. In addition to the immediate consequences of disproportionately higher scores, algorithm outputs used in decision-making can affect downstream outcomes.[2]

While algorithms may be more efficient and make less errors than human decision makers (Kleinberg et al., 2017), the widespread use and non-transparent[3] nature of algorithms has raised concerns about algorithmic inequality (O'Neil, 2017; Eubanks, 2018). Discussion has focused on risk assessment tools used in the criminal justice system. Across the United States, judges are increasingly using risk-assessment tools that predict outcomes like recidivism,

[1]Explicitly using protected classes like race may infact be unlawful in many decision-making contexts. A prominent legal article on evidence-based sentencing and discrimination writes that "there appears to be a general consensus that using race would be unconstitutional" (Starr, 2014). Overall, the Supreme Court has "squarely rejected statistical discrimination — use of group tendencies as a proxy for individual characteristics — as a permissible justification for otherwise constitutionally forbidden discrimination" (Starr, 2014).

[2]For example, pretrial detention has been shown to causally decrease defendants' formal sector employment and take-up of government benefits after release (Dobbie et al., 2018).

[3]Algorithms can be considered non-transparent in many ways. Algorithms can be proprietary, and firms may not disclose the training data or algorithm used. Therefore "users" have no way to understand where their "risk score" came from. Machine learning techniques are increasingly used that can also be perceived as being less interpretable and explainable. Even when algorithms models or input variables are known, it may be unclear to users exactly how algorithms make implicit tradeoffs between predictive power and group equity.

whether an individual will, after being released from jail, commit another crime.[4] One of the most commonly used of those tools is the Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS. COMPAS collects a range of information about the defendant, with a questionnaire containing over one hundred questions. The information is then used to generate a recidivism risk score, which represents the defendant's risk of recidivism. Judges and other officials use these scores to make important decisions about a defendant, including the terms of their pretrial release and bail amount.[5]

In my paper, I show that tension between overall predictive power and group equity can lead to including variables that only marginally increase overall predictive power but substantially increase in group disparities in risk scores. This paper studies how using different variables affect this trade-off between overall predictive power and group equity. Residential location of defendants, which I examine, is one example of a variable that could proxy for race and economic status. The analysis can be extended to look at other similar variables. This is the first paper to study this tension between predictive power and group equity of input variables explicitly in a tool used in society. While there is evidence on how some input variables explain risk scores and contribute to predictive power, there is little evidence on how using group-level variables like neighborhood identifiers contribute to predictive power and group disparities in risk models.[6] In a review article, Byrne and Pattavina (2017) ask "what if the price of improved predictive accuracy [through community-level risk variables such as neighborhoods] is increased gender, race, or class-based disparity?" and write that "no study has been conducted to date that examines the accuracy of the neighborhood assessment data included in these risk models." Prior research finds mixed results of whether COMPAS tools are racially biased, depending on the definition of racial bias (Larson et al., 2016; Flores et al., 2016).[7] Subsequent papers have shown that different notions of fairness and accuracy

[4]"Dozens of jurisdictions and at least six entire states" have adopted pretrial risk assessment tools, and at least 28 full states use risk assessment tools at sentencing (Stevenson and Doleac, 2018).

[5]There is growing evidence on the impacts of using risk scores on pretrial judicial decision-making (Stevenson, 2018; Cowgill, 2018) and parole release decisions (Berk, 2017), and how judges use risk scores in sentencing (Garrett and Monahan, 2018). Kleinberg et al. (2017) develop a machine-learning algorithm that performs better than pretrial judges.

[6]There is evidence that COMPAS does not predict recidivism better than recidivism prediction using only a subset of COMPAS inputs (Dressel and Farid, 2018; Angelino et al., 2017). While I study the COMPAS general recidivism scale, Stevenson and Slobogin (2018) find that a subset of COMPAS variables explain 72% of variation in the COMPAS violent recidivism risk scale. These studies also use COMPAS data from Broward County, Florida; however, these studies include neither defendant neighborhood nor socioeconomic data. Moreover, I build on these studies by examining how these input variables explain and contribute to group equity in risk scores and outcomes.

[7]ProPublica (Larson et al., 2016) claim that COMPAS is racially biased by showing (1) the false-positive rate is higher for blacks than whites, and (2) the false-negative rate is higher for whites than blacks. Using the same data, Flores et al. (2016) find race to be statistically insignificant in a regression of recidivism on risk score levels, race, and its interaction terms. This paper also contributes to literature studying statistical

can be incompatible with each other (Chouldechova, 2017; Kleinberg et al., 2016; Berk et al., 2017),[8] In particular, Corbett-Davies et al. (2017) conceptualize algorithmic fairness as constrained optimization, and show that the decision rules under unconstrained optimization and constrained optimization differ. Using data from Broward County, Florida, they illustrate that satisfying fairness criteria reduces public safety. While we know the theoretical tradeoff exists between efficiency in prediction and equity, and this paper illustrates this tradeoff using different fairness criteria, my papers examines – in practice in a tool already deployed in society – how variables can be used that differentially affect groups in order to quantify exactly how specific input variables in an algorithm tradeoff overall predictive power and group equity. My paper shows how important it is for practice and policy to be explicitly aware of this tradeoff and how tools predict differentially across groups, as I show in practice that predicting slightly better overall can result in dramatically worse outcomes for Black and indigent defendants.

I consider this trade-off in COMPAS using a novel data set from Broward County, Florida, over 2013-2016 by combining data from the following sources: 1) new data that I acquired from the Broward County Clerk, 2) (Larson et al., 2016), and 3) U.S. Census and American Community Survey (ACS) data (Manson et al., 2017). This combined dataset makes this analysis possible. I find that including defendants' residential neighborhoods marginally increase predictive power, but differentially predict across groups and introduce dramatic group disparities. The same analysis could be performed for other input variables used in any data-driven algorithm.

Residential neighborhoods are an example of a controversial variable as defendants may have limited control over where they live. Yet, neighborhood would also act as a proxy for race (Berk, 2009) and socioeconomic status, due to persistent residential segregation by race and income in the US where minority and low-income individuals live in vastly different neighborhoods (Reardon et al., 2015). The main inputs of the COMPAS General Recidi-

discrimination (Fang and Moro, 2011), racial profiling (Durlauf, 2006), racial bias in policing (Anwar and Fang, 2006; Knowles et al., 2001) and the criminal justice system (Anwar et al., 2012; Sorensen et al., 2012; Anwar and Fang, 2015; Mechoulan and Sahuguet, 2015), and specifically in pretrial bail (Arnold et al., 2017). More broadly, this paper also relates to research on discrimination in other algorithmic decision-making and allocation settings (Datta et al., 2015; Sweeney, 2013; Kay et al., 2015; Lum and Isaac, 2016).

[8]Chouldechova (2017) prove that if an algorithm is "test-fair" (all people with the same risk score have the same likelihood of recidivism across all groups) and groups have different rates of recidivism, then the rate of false-positives and false-negatives will not be equal across groups. Similarly, Kleinberg et al. (2016) define three notions of fairness and prove they cannot all be satisfied simultaneously except in special cases. See (Berk et al., 2017) for a summary of trade-offs between different definitions of fairness, and trade-offs between fairness and accuracy. Kleinberg and Mullainathan (2019) show there is a trade-off between simplicity and fairness, and Kleinberg et al. (2018) show that excluding race from prediction can lead to less efficient and less fair outcomes.

vism Scale are as follows: (1) prior criminal history, (2) peer criminal networks[9], (3) drug involvement, and (4) indicators of juvenile delinquency (Northpointe, 2012). Information about a defendant's peer criminal networks is likely correlated with neighborhoods where defendants live. Black and Hispanic youth live in different neighborhoods, and have shown to be "almost twice as likely to report that significant numbers of their peers belong to gangs" compared to white counterparts (Graham, 2018).[10] The questionnaire used to assess defendants also contains questions about characteristics of neighborhoods where defendants live. In particular, Brennan et al. (2009) reports that COMPAS uses a "high crime neighborhood" scale, which contains information on neighborhood crime, gang activity, and drug activity.[11] Official clerk records also contain defendant addresses.

First, I decompose defendants' risk scores into their criminal information (current charge, criminal history, and juvenile history) and a residual. Next, I use novel defendant-level data on where defendants live to decompose the residual using census tract-level variables and census tract fixed effects to quantify unobservable factors about a defendant's peer networks and neighborhood characteristics. Second, I analyze how input variables contribute to explaining group differences in risk scores, to assess how including variables contribute to average gaps across groups in risk scores. Third, I examine how input variables contribute to overall predictive power. I compare how input variables explain COMPAS scores and predict recidivism overall to how they explain group differences in COMPAS scores. Input variables contribute to group disparities in scores if the variable itself explains more of the average gap across groups in risk scores than it explains of the overall variance in risk scores. Finally, I assess how variables contribute to predictive power across groups (subgroup validity).[12] That is, does including defendants' census tract-level information improve out-of-sample performance equally across subgroups or widen differences in correct prediction and false positive rates across subgroups? Taken together, I assess how models tradeoff predictive power and group equity in prediction.

Overall, I find that including defendants' neighborhood factors (fixed effects or neighborhood demographic and poverty variables) negligibly improves predictive power. However,

[9]Whether a defendant associates with "highly delinquent friends" involved in drugs, crime or gangs.

[10]There may be criminal behavior peer effects, as there is evidence that juvenile offenders have peer effects on recidivism of other offenders in the same facility (Bayer et al., 2009).

[11]Brennan et al. (2009) write that "living in a high-crime neighborhood is an established correlate of both delinquency and adult crime." There are empirical foundations for the relationship between recidivism risk, race, and neighborhoods (Berk, 2009; Stahler et al., 2013), and also emerging evidence that moving to a new neighborhood after release from prison decreases recidivism (Kirk, 2019).

[12]Ayres (2002) writes about "subgroup validity" in the context of outcome tests: "when a particular observable characteristic is only a valid proxy of desert for some races, then a decisionmaker's unwillingness to engage in disparate racial treatment may induce just the racial disparities in outcomes that are generally a concern."

these variables unequally predict and substantially widen risk score disparities across race and indigent status. Differences in defendants' neighborhoods census tract fixed effects and neighborhood-level variables account for 3.7 to 6.7 percent of the overall variance in risk scores explained by linear and nonlinear models. I find that neighborhood-level variables add negligible predictive power overall in predicting recidivism (-0.03 to 0.04 percentage points in correct prediction). Yet, these variables widen large group disparities in risk scores. Census tract variables account for 19.2 to 23.9 percent of the black-white gap in risk scores and for 9.2 to 18.6 percent of the differences in risk scores between 'indigent" defendants, who use a public defender, and "non-indigent" defendants who do not. Differences in score gaps across race and indigent status would be substantially smaller if these variables were not used. Finally, I find that census tract-level variables differentially predict across race and indigent status of defendants: including census tract-level variables increases over-prediction for black and indigent defendants but decreases over-prediction for white and non-indigent defendants.

With the broad deployment of data-driven algorithms, there may be consequences of disparities in algorithm predictions, in particular for protected groups if increasing predictive power comes at the expense of economic and racial equity. If systems are designed to maximize predictive power without consideration of group disparities, tools such as COMPAS will implicitly overlook these group disparities, and treat individuals differently across groups. Including neighborhood-level variables introduces disproportionately large group disparities and higher scores that can further impact defendants' economic outcomes. Ultimately, the design of systems can carefully account for this trade-off by quantifying and controlling exactly how to compromise between predictive power and group disparities, whether through objective functions that account for a social welfare function, or through preprocessing of input variables.

The paper is organized as follows. Section 2 introduces the context, and Section 3 describes the data that I use in my analysis and presents descriptive statistics. Section 4 details the empirical framework. Section 5 presents the results. Section 6 discusses policy implications of this paper and Section 7 concludes.

# 2 Context

Risk assessment tools are increasingly used as inputs to decision-making across many settings, including in the criminal justice system. These tools are used for a number of purposes; among the most important is to predict the recidivism risk of a criminal defendant. One of the most commonly used risk assessment tools is COMPAS. A publication on COMPAS

(Brennan et al., 2009) reveals that COMPAS uses "more advanced statistical methods for predictive modeling and classification" (Brennan et al., 2009). In Broward County, nearly everyone arrested (all except those with murder charges and other capital crimes that are not eligible for pretrial release) are assessed using COMPAS (Angwin, 2016).

Defendants are screened using COMPAS after their arrest. The COMPAS screener records information about the defendant[13], and poses questions to defendants (Larson et al., 2016). The topics of these questions include: family criminality, peers (criminality, drug use, and gang membership), substance abuse, residence/stability, social environment of their neighborhood, education, work, leisure/recreation, and value statements on social isolation, criminal personality, anger and criminal attitudes (Larson et al., 2016). The exact proprietary model and inputs used in the COMPAS recidivism scale are not explicitly known.[14] The COMPAS model outputs a raw score that is mapped to a decile score from 1 to 10 (highest risk), and a "text score" of low (1-4), medium (5-7) and high (8-10) (Larson et al., 2016). COMPAS produces recidivism, failure to appear and violent recidivism risk scores. In particular, I study the COMPAS recidivism risk score, a score that purportedly represents the likelihood that a defendant will be rearrested within two years of release. As I cannot work directly with the algorithm, I examine the raw risk scores, which are the direct outputs of the COMPAS recidivism risk tool and the richest information about how COMPAS ranks defendants.

While I do not observe all of the input variables to COMPAS, I explain 64-65 percent of the total variation in the COMPAS raw risk scores (R-squared) using linear and nonlinear polynomial forms. These answers to these questions are self-reported, and could be subject to reporting bias.[15]

[13]Topics include current charges, official criminal history, disciplinary infractions while incarcerated, and non-compliance.

[14]I have found no recent document of Northpointe/Equivant within the last 5-7 years that explains the exact inputs to the General Recidivism scale. I also contacted Northpointe/Equivant who would not comment on the inputs to the scale. In a recent article in the New York Times, Northpointe/Equivant claims that the algorithm is proprietary and a "trade-secret" (Liptak, 2017).

[15]In their online FAQ, Northpointe/Equivant acknowledges that "many factors may distort the data and introduce errors in either the self-report data or the official criminal records."(http://www.northpointeinc.com/files/downloads/FAQ_Document.pdf). They claim to have a lie test/defensiveness test, a random responding test (to detect inconsistent answers), and an "inconsistency" (of social history with risk factors) test, to detect and alert for potential misreporting and errors. Defendants have an incentive to strategically misreport if they understand that certain characteristics may give them a higher score, and they believe that a higher score will make them worse off. This may be a concern if strategic misreporting is differential across race.

# 3 Data and Descriptive Statistics

I conduct my analysis by combining novel data that I collected and assembled from the Broward County Clerk with data from several sources: data that ProPublica obtained for their investigation into racial bias in COMPAS scores (Larson et al., 2016); Census 2010 tract-level demographic and racial composition data; and American Community Survey (ACS) 2012-2016 tract-level socioeconomic data (Manson et al., 2017). I collect criminal records data from the Broward County Clerk website using the Commercial Data API. Appendix A describes in detail the process by which I collected the raw criminal records clerk data, processed and coded the raw data to create a data set for analysis, cleaned and validated the data set, merged the data set with the other sources, and validated the merging.

ProPublica obtained COMPAS score data on all COMPAS-screened defendants between January 1, 2013, and December 27, 2014, in Broward County, Florida, through a Freedom of Information Act request. ProPublica also gathered data on demographic characteristics (age, marital status, race), initial arrests, subsequent recidivism arrests, and criminal record and juvenile history.

Using the Broward County Clerk data, I match criminal cases with COMPAS screening data by first and last name[16], date of birth (with 3 days flexibility), and whether a COMPAS screening is within 30 days of any arrest associated with a case (following Larson et al. (2016)). I also validate the cases that I match. I merge records with ProPublica data, by first name, last name and date of birth, and case record number. I merge defendant-level data with Census 2010 tract-level demographic and racial composition data, and American Community Survey (ACS) 2012-2016 tract-level socioeconomic data, by defendant's census tract (Manson et al., 2017). In my analysis, I use the sample of black, Hispanic and white male pretrial defendants who have addresses that successfully geocode, and match with census tracts and census tract-level data. I analyze the sample of men, as COMPAS uses separate models for men and women for prediction and testing (Brennan et al., 2008). Table 1 contains descriptive statistics for this sample. Comparing defendant outcomes by race, black men are younger than white men, and also more likely to have a criminal history than white and Hispanic men. Blacks are more likely to have a felony charge than white men, and have higher recidivism rates on average. Blacks have higher COMPAS scores on average. For my analysis, I study the COMPAS raw score which I standardize. The COMPAS raw score is the raw output from the COMPAS algorithm that is mapped to a decile (1-10) and low-medium-high score. I analyze the COMPAS raw score which contains more information on exactly how

[16]I use a flexible wild card match that allows the first portions of names to match with longer names in records. For example, "Em" matches with "Emma."

COMPAS ranks defendants. The recidivism outcome includes all criminal offenses resulting in a jail booking, which is consistent with both ProPublica and Northpointe/Equivant's definition of recidivism. The recidivism outcome that I use for most of the analysis is recidivism within 2 years of release, conditional on defendants having at least 2 years at risk. I also perform robustness checks using the hazard of recidivism to account for selection and censoring in this outcome.

**Table 1:** Average defendant characteristics and outcomes

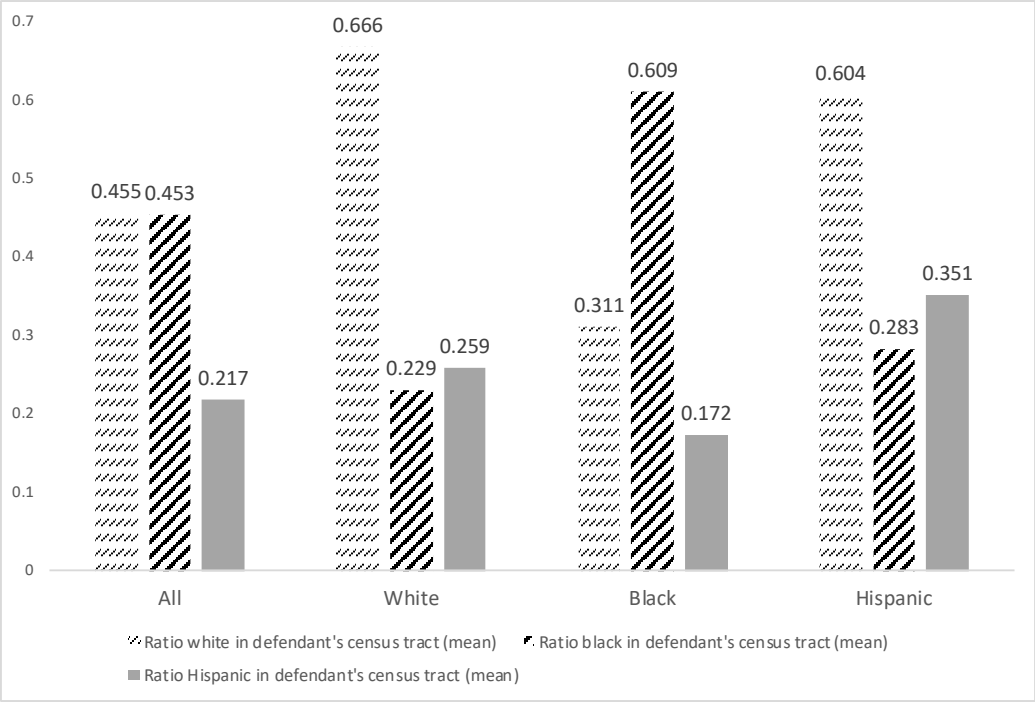
| | All N = 4,775 | White N = 1,592 | African-American N = 2,758 | Hispanic N=425 |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| *Demographic and criminal history* | | | | |
| Age | 34.646 | 37.727 | 32.771 | 35.320 |
| Indigent proxy | 0.528 | 0.455 | 0.589 | 0.405 |
| Priors count | 3.420 | 2.220 | 4.360 | 1.817 |
| Juvenile felony count | 0.077 | 0.041 | 0.105 | 0.027 |
| Juvenile misdemeanor count | 0.094 | 0.044 | 0.130 | 0.046 |
| Juvenile other count | 0.120 | 0.104 | 0.138 | 0.057 |
| *Criminal outcomes* | | | | |
| Felony charge | 0.672 | 0.599 | 0.726 | 0.602 |
| COMPAS Raw Score (Z score) | 0.018 | -0.372 | 0.318 | -0.463 |
| Recidivism (2 year) | 0.364 | 0.299 | 0.414 | 0.262 |

*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* Sample of black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data. Indigent proxy is 1 if the defendant qualifies for a public defender. Recidivism (2 year) is recidivism outcome within 2 years conditional on the defendant having at least 2 years at risk.

Located directly north of Miami-Dade County, Broward County is one of the largest counties in Florida, and includes the city of Fort Lauderdale. Broward County ranks high in terms of black-white segregation, and census tracts where individuals live contain information about an individual's race. In 2010, the black-white segregation (dissimilarity) index of the combined Miami-Fort Lauderdale - Pompano Beach metropolitan statistical area was 64.8, the 23rd highest of the 102 largest metropolitan statistical areas in the US (Frey and Myers, 2005), and can be described as a "highly segregated city."[17] Comparing census tracts where defendants live, I find that defendant race is correlated with the racial composition of census tracts where defendants live (Figure 1). Black and white defendants are more likely to live in tracts where 50 percent or more of the population is their race.

[17]Dissimilarity indices greater than sixty can be considered highly segregated (Graham, 2018).

**Figure 1:** Mean defendant census tract racial composition by defendant race




*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* Sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data.

# 4 Empirical Framework

I outline a framework to assess the extent that risk scores implicitly balance the trade-off between predictive power and group equity – specifically economic and racial equity. First, I measure how much individual components contribute to explaining outcomes in order to determine the extent that COMPAS scores can be attributed to individual score components. Second, I measure how factors contribute to overall recidivism predictive power to understand the marginal contribution of each factor to predicting recidivism. Third, I quantify how much group differences in risk scores are explained by level differences in components across groups using a decomposition framework. By measuring how components contribute to group differences in risk scores, I quantify how large group differences would be with out individual components. Finally, I assess how individual score components contribute to overall prediction and prediction across groups. Taken together, I compare how much components explain overall variation in risk scores and group differences in risk scores, and how components predict overall, across race, and socioeconomic status. Specifically, I quantify this trade-off for covariates like census tract fixed effects and census tract-level variables.

## 4.1 How much of overall variance in risk scores is explained by input variables?

To decompose risk scores into the weight score components, I regress continuous COMPAS raw risk scores on score components:

$$COMPAS_i = \alpha_0 + \alpha_1 C_i + \varepsilon_i$$

where $C_i$ is a vector containing individual-level characteristics. In the context of COMPAS, the individual level variables that I observe and include are: number of priors, juvenile history (number of juvenile felonies, number of juvenile misdemeanors, number of juvenile "other" category), charge-severity fixed effects, a proxy of whether defendant is indigent and uses a public defender, and age controls (age and age-squared). I run linear specifications of all individual characteristics and second order polynomial of all individual characteristics excluding charge-severity fixed effects.

A concern is that there may be factors that I do not observe that could be used to predict risk scores. As a first step I do not include group fixed effects to quantify unobservable characteristics, and consider how factors that I do not observe can be quantified using the regression residuals $\varepsilon_i$. Next, I can further decompose the regression residual using group fixed effects $\xi_j$ or group-level variables $N_j$:

$$COMPAS_i = \alpha_0 + \alpha_1 C_i + \xi_j + \varepsilon_i \quad (1)$$

$$COMPAS_i = \alpha_0 + \alpha_1 C_i + \alpha_2 N_j + \varepsilon_i \quad (2)$$

In the context of COMPAS, I include neighborhood (census tract) fixed effects $\xi_j$ to capture unobservable characteristics, and census tract-level demographic variables $N_j$ which is a vector containing the census tract-level shares of population that are black, Hispanic, with income below the poverty level, with a bachelors or associate's degree or higher, and labor force participation. I include census tract fixed effects to account for unobservable characteristics of locations where defendants live that can be important in predicting recidivism. For example, the explanatory power of census tract variables is $\frac{R^2_{CT} - R^2_{noCT}}{R^2_{CT}}$ where $R^2_{CT}$ is the adjusted R-squared of models including census tract variables and $R^2_{noCT}$ is the adjusted R-squared of models excluding census tract variables. I use the adjusted R-squared which accounts for the number of input variables, which is especially important given the large number of census tract fixed effects being estimated.

### 4.1.1 Census Tract Fixed Effects

To understand the information contained in census tract fixed effects, I decompose the estimated census tract fixed effects on observed neighborhood characteristics:

$$\begin{aligned}\xi_j = \phi_0 + \phi_1 RatioBlack_j + \phi_2 RatioHispanic_j + \phi_3 RatioPoverty_j \\ + \phi_4 RatioEducation_j + \phi_5 LFP_j + \eta_j\end{aligned}$$

where $\xi_j$ is estimated from Equation 1 and 2.[18] $RatioBlack_j$ and $RatioHispanic_j$ are the census tract ratios of population that are black and Hispanic, respectively, $RatioPoverty_j$ is the census tract level proportion with income below the poverty level, $RatioEducation_j$ is the census tract proportion with a bachelors or associate's degree or higher, and $LFP_j$ is the census tract labor force participation.

## 4.2 How are group differences in risk scores explained by differences in input variables across groups?

I decompose differences in mean risk scores across two groups into what is due to group differences in the levels of each score input in order to address whether certain components contribute disproportionately more to explaining the group difference in risk scores than explaining overall outcomes. Using the decomposition from Equation 1, the mean group difference in risk scores $COMPAS$ can be decomposed using the estimates of each component and the average group difference in levels of components:

$$\begin{aligned}E[COMPAS|G_i = 1] - E[COMPAS|G_i = 0] = \alpha_1(E[C|G_i = 1] - E[C|G_i = 0]) \\ + E[\xi|G_i = 1] - E[\xi|G_i = 0] \\ + E[\varepsilon|G_i = 1] - E[\varepsilon|G_i = 0]\end{aligned}$$

where $G_i$ denotes group membership.

Therefore, from the data:

$$\begin{aligned}\overline{\widehat{COMPAS}}_{g1} - \overline{\widehat{COMPAS}}_{g0} = \hat{\alpha}_1(\overline{C}_{g1} - \overline{C}_{g0}) \\ + (\overline{\hat{\xi}}_{g1} - \overline{\hat{\xi}}_{g0}) \\ + (\overline{\hat{\varepsilon}}_{g1} - \overline{\hat{\varepsilon}}_{g0})\end{aligned}$$

[18] Allowing for $C_i$ in Equation 1 to be a is a vector containing individual-level characteristics (linear specification) or a vector with second order polynomials of all individual characteristics excluding charge-severity fixed effects (nonlinear specification).

where the mean of component $C$ is $\overline{C}_{g1} = \dfrac{\sum_{i=1}^{N} C_i \cdot G_i}{\sum_{i=1}^{N} G_i}$, and $\overline{\hat{\xi}}_{g1} = \dfrac{\sum_{i=1}^{N} \hat{\xi}_i \cdot G_i}{\sum_{i=1}^{N} G_i}$.

Using the decomposition, the contribution of group differences in neighborhood fixed effects (because of where defendants live) to explaining the group difference in COMPAS scores is:

$$\frac{\overline{\hat{\xi}}_{g1} - \overline{\hat{\xi}}_{g0}}{\overline{\widehat{COMPAS}}_{g1} - \overline{\widehat{COMPAS}}_{g0}} \tag{3}$$

Using the census tract-level variables to perform the decomposition, the contribution of group differences in neighborhood-level variables (because of where defendants live) to explaining the group difference in COMPAS scores is:

$$\frac{\hat{\alpha_2}(\overline{N}_{g1} - \overline{N}_{g0})}{\overline{\widehat{COMPAS}}_{g1} - \overline{\widehat{COMPAS}}_{g0}} \tag{4}$$

In my analysis, I decompose average group differences by race and indigent status of defendants.[19]

## 4.3 How do input variables predict recidivism outcomes out-of-sample?

How much of overall variance in outcomes is predicted by components? To understand the total increase in predictive power from including group variables, I use logistic regression of score components to predict recidivism (binary outcome) out-of-sample.

$$Recidivism_i = 1\left[\beta_0 + \beta_1 C_i + \beta_2 N_j - \rho_i \geq 0\right]$$

where $\rho_i$ is standard logistically distributed. I also allow for $C_i$ to contain second order polynomial interactions of all individual characteristics except charge-severity fixed effects. I assess the out-of-sample performance of models including and excluding variables using

[19]Race cannot explicitly legally enter risk scores, and COMPAS claims to not use race in their models. There are also ethical issues with using the indigent status of a defendant, since their socioeconomic status is not part of their criminal record and not part of their behavior that they can control. I also find qualitatively similar results in analysis that includes the indigent proxy to decompose the average indigent-non-indigent risk score gap. Differences across economic status in COMPAS scores are slightly disproportionately attributed to neighborhood differences across economic status, compared to how much neighborhood variables explain the overall variance in risk scores.

five- and tenfold cross-validation, which are recommended as "good compromise" between bias and variance of the true prediction error (Hastie et al., 2009).

To assess the predictive power of including census tract variables, I compare the out-of-sample performance of models including and excluding census tract variables. The predictive power from census tract variables is $R^2_{CT} - R^2_{noCT}$ where $R^2_{CT}$ is the pseudo R-squared of models including census tract variables and $R^2_{noCT}$ is the pseudo R-squared of models excluding census tract variables. I compare the predictive power of census tract variables to the predictive power of other variables.

### 4.3.1 Robustness using hazard models

Recidivism outcomes are two year rearrest after release which is the same measure that COMPAS predicts. Because of the nature of pretrial release and right censoring in the rearrest observation period, all COMPAS screened defendants are not necessarily released from jail for two years and "at risk" of being rearrested for two years. There could be several cases of selection bias. First, a defendant's first arrest could be less than two years prior to the end of the rearrest observation period; then it would be impossible for them to be at risk for two years. Second, a defendant who is arrested more than two years before the end of the rearrest observation period could be detained or re-incarcerated (for a previous crime) so that they are not released for at least two years over the observed period (right censoring issues). Therefore, directly using the probability of recidivism two years after release would estimate based on a subpopulation that could be "selected" non-randomly from the study population.[20] Appendix Table A1 compares the "selected sample" of defendants who have at least two years at risk (out of jail) to those who do not have at least two years out of jail.[21] To address selection and right-censoring issues in the data, I use the fact that most defendants are released before the end of the observation period to model the hazard of recidivism.[22] The hazard rate of recidivism for the $i$th defendant is:

$$h(t|\boldsymbol{x_i}) = h_0(t) \exp\left\{\beta_0 + \beta_1 C_i + \beta_2 N_j + \delta J_i + \rho_i\right\},$$

Defendants may be in jail during the observed time period, during which time they are

[20] Northpointe/Equivant will have the same selection issue due to the right censoring in arrest data, and there may be individuals who do not have two years out of jail. Another concern is that data for recidivism is from Broward County only, which may lead to the sample of recidivism outcomes being further selected. Tool makers may face similar issues depending on how long rearrest is observed.

[21] Appendix Table A1 finds that the "selected" sample and the rest of defendants are statistically different on whether the defendant qualifies for a public defender, the racial composition of the defendant's census tract, and the count of juvenile misdemeanors.

[22] Appendix Figure A1 shows that few outliers are in jail for the majority of the observed period, and most are in jail for less than 10 days.

not at risk, and cannot be rearrested. I remove any time that defendants are in jail as time not at risk, and split the duration that defendants are at risk into segments separated by jail stays if they exist. To assess the change in predictive power of the model, I examine how the pseudo R-squared from k-fold cross validation changes between the model with and without census tract-level variables.

An additional concern is that recidivism outcomes may be affected by the COMPAS risk scores which judges use in their decision-making. Judges use COMPAS risk scores to decide the terms of pretrial release, which can affect the length of a defendant's pretrial incarceration, which can in turn affect their recidivism outcomes. This is part of a larger concern if pretrial treatment and recidivism are jointly determined (Bushway and Smith, 2007). To account for this, I include time-varying covariates for defendant treatment in the criminal justice system (length of jail stays), $J_i$ is a vector of the lengths of jail stays (in months). Lengths of jail stays during the observed period are included as time varying regressors. The time that they spend in jail may change their rate of rearrest. This addresses the interaction between defendant COMPAS risk scores and recidivism outcomes if recidivism outcomes are only affected through the number and length of pretrial jail stays. It is possible that there are other unobservable characteristics that may interact with the way that COMPAS risk scores affect recidivism outcomes that I cannot address. Controlling for jail stays could be "over-controlling" if jail stays are a function of COMPAS risk scores. I also assume that rearrest conditional on committing a crime does not differ by race or group.

In Appendix C, I also implement a multi-stage framework to first estimate how components explain COMPAS and then see how the weighted components contribute to predicting recidivism. I find qualitatively similar results. First, I estimate the weight of each potential input variable component in predicting outcomes. I decompose defendants' risk scores into an ex ante probability of recidivism (predicted using a second order polynomial of current charge, criminal and juvenile history) and a residual. In turn, I decompose the residual using defendant neighborhoods effects. Next, I analyze how input variables contribute to predicting recidivism and how they introduce group disparities into model outputs. A variable introduces "group disparities" or contributes to group differences "disproportionately" if the variable itself explains more of the average gap across groups in risk scores than the average gap across groups in predicted outcomes.

## 4.4 How do input variables contribute to predictive power across groups?

I assess how input variables contribute to recidivism prediction overall and across groups, by comparing out-of-sample predictive performance of COMPAS risk scores to COMPAS residualized of input variables. To understand how defendant census tract data contributes to predictive power, I compare correct prediction, false positive and false negative rates of COMPAS risk scores to COMPAS risk scores residualized of defendants' census tract variables.

Using half of the data, I decompose COMPAS raw risk scores on individual score components as in Equations 1 and 2 in Section 4.1:

$$COMPAS_i = \alpha_0 + \alpha_1 C_i + \xi_j + \varepsilon_i$$
$$COMPAS_i = \alpha_0 + \alpha_1 C_i + \alpha_2 N_j + \varepsilon_i$$

$C_i$ contains individual characteristics. I also use a nonlinear specification that includes up to second order polynomial interactions of all individual characteristics excluding charge-severity fixed effects.

Using the second half of the data, I residualize COMPAS raw risk scores of the weight of individual input variables. For example, for census tract-level variables and census tract fixed effects[23]:

$$ResidScore_{i,-\xi} = COMPAS_i - \hat{\xi}_j \tag{5}$$
$$ResidScore_{i,-N} = COMPAS_i - \hat{\alpha_2} N_j \tag{6}$$

### 4.4.1 Recidivism Classification:

**COMPAS Classification:** Previous studies classify defendants as high risk of recidivating if they receive a COMPAS score of Medium or High (Larson et al., 2016; Dressel and Farid, 2018). In this prediction exercise, I classify defendants as being high risk to recidivate using the same criteria. Following previous literature, I focus on the threshold between Low and Medium scores, as Medium and High scores "garner more interest from supervision agencies than low scores" (Larson et al., 2016; Northpointe, 2012). The Low-Medium-High and decile

[23]When predicting census tract fixed effects out-of-sample, I end up dropping 136 defendants from 118 census tracts that are represented in the half of the data used to "train" or decompose the COMPAS score.

score are seen by judges, while the raw risk score is not.

I calculate rates of correct prediction (true positive and true negatives) as the fraction of defendants whose classification matches whether they actually recidivate within 2 years. The false positive rate is the proportion of false positives cases of all defendants who do not recidivate, and the false negative rate is the proportion of false negative cases of all defendants who do recidivate.

**Residualized Score Classification:** Using the residualized COMPAS scores, I classify the same number of people as high risk to recidivate as under the COMPAS score.[24] I calculate the rates of correct prediction, false positives, and false negatives of the COMPAS score residualized of census tract-level variables $ResidScore_{-N}$ and census tract fixed effects $ResidScore_{-\xi}$.

**Prediction across groups:** I calculate the rates of correct prediction, false positives, and false negatives of the COMPAS score residualized of census tract-level variables $ResidScore_{-N}$ and census tract fixed effects $ResidScore_{-\xi}$ by race and indigent status.

**Contribution of individual variables:** The contribution of census tract-level variables and fixed effects to predictive power is given by the difference between the prediction rate from COMPAS and the prediction rate and the score residualized of census tract-level variables. I use the overall and group differences in correct prediction, false positive rate, and false negative rate to assess how census tract-level variables contribute to recidivism prediction overall and across groups, respectively.

This exercise assumes that rearrest conditional on committing a crime does not differ by race or indigent status, and that there are no selection issues with the outcome, recidivism within two years.[25] I also assume that variables that COMPAS uses that I do not observe are orthogonal to census tract-level variables. If unobservable variables are not orthogonal to census tract-level variables, census tract-level variables may capture the contribution of these

[24]The ranking cutoff differs across exercises slightly because the census tract fixed effects exercise drops some defendants in census tracts not represented in the training sample. With census tract fixed effects, 46.5 percent of defendants are predicted to recidivate with high risk with the COMPAS risk score and with the score residualized of census tract fixed effects. Even restricting to the sample of defendants at risk for at least 2 years to assess predictive power, with COMPAS 47.3% of the sample is "predicted" to recidivate, and with the score residualized of census tract fixed effects 47.1% of the sample is "predicted" to recidivate. Without dropping any individuals (in the exercise using census tract-level variables), even after restricting to the sample who have at least two years at risk , under the COMPAS ranking 47.7% of individuals are "predicted" to recidivate, and under the score residualized of census tract-level variables 47.4% of individuals are "predicted" to recidivate.

[25]This is the outcome that COMPAS is predicting. Due to right censoring in arrest data and data limitations, the sample of defendants who are at risk for at least 2 years during the observed may be a selected sample. Another limitation of this exercise is that defendants outcomes could be affected by the scores they receive and their pretrial treatment. Section 4.3 addresses these issues by using a Cox Proportional hazard model and controlling for pretrial jail stays (assumption is that defendants outcomes are only affected through the length and number of jail stays).

omitted variables in prediction.[26] This would change the interpretation of the exercise, but it is still possible to compare the overall change in prediction and the changes in prediction across groups from including these variables.

# 5 Results

## 5.1 Overall Explained Variance in Risk Scores

This section decomposes COMPAS risk scores to measure the weight of each component in COMPAS. Table 2 shows the results from a linear decomposition of the standardized COMPAS risk score into input components. Column (1) includes all variables in the linear model. Columns (2) to (6) include all variables excluding one set of variables at a time, neighborhood-level variables, count of priors, juvenile record variables, age controls, charge-severity fixed effect. Comparing Column (1) with Columns (2) to (6) yields the marginal increase in explanatory power that each set of variables contributes to the model.

Column (1) shows that input variables explain 55.4 percent of the total variation in the COMPAS raw risk score.[27] The explanatory power of the model without including neighborhood variables is 0.517 in Column (6). Including neighborhood fixed effects increases overall explanatory power 7.2%. Neighborhood fixed effects accounts for 6.7% of the total variation explained by the linear model. Number of criminal priors accounts for 30.1% of the total variation explained by the linear model, and age controls which accounts for 45.1% of the total variation explained by the linear model. From these estimates, moving from the 25th percentile to the 75th percentile of neighborhood unobservable characteristics (fixed effect) carries the same weight as nearly 4 priors.

[26] Omitted variable bias direction depends on the sign of the correlation between the omitted variable and census tract-level variables.

[27] I use the Adjusted R-squared due to the large number of fixed effects included in the model.

**Table 2:** Linear decomposition of standardized COMPAS raw risk score using census tract fixed effects

| | (1) All | (2) All - Priors | (3) All - Juvenile | (4) All - Age controls | (5) All - Charges | (6) All - Census Tract |
|---|---|---|---|---|---|---|
| Priors count | 0.094 | | 0.096 | 0.061 | 0.097 | 0.099 |
| | (0.003) | | (0.003) | (0.003) | (0.003) | (0.002) |
| Juvenile felony count | 0.094 | 0.239 | | 0.165 | 0.095 | 0.104 |
| | (0.037) | (0.049) | | (0.040) | (0.037) | (0.032) |
| Juvenile misdemeanor count | -0.021 | 0.186 | | 0.103 | -0.025 | -0.026 |
| | (0.028) | (0.038) | | (0.044) | (0.028) | (0.025) |
| Juvenile other count | 0.059 | 0.120 | | 0.224 | 0.061 | 0.044 |
| | (0.024) | (0.033) | | (0.038) | (0.024) | (0.020) |
| Age | -0.095 | -0.039 | -0.097 | | -0.099 | -0.086 |
| | (0.006) | (0.007) | (0.006) | | (0.006) | (0.006) |
| Age squared | 0.001 | 0.000 | 0.001 | | 0.001 | 0.001 |
| | (0.000) | (0.000) | (0.000) | | (0.000) | (0.000) |
| Census tract fixed effects | Y | Y | Y | Y | Y | |
| Charge degree fixed effects | Y | Y | Y | Y | | Y |
| Census tract fixed effect 25th Percentile | -0.189 | | | | | |
| Census tract fixed effect 50th Percentile | 0.018 | | | | | |
| Census tract fixed effect 75th Percentile | 0.168 | | | | | |
| RMSE | 0.671 | 0.787 | 0.673 | 0.838 | 0.677 | 0.699 |
| R-squared | 0.633 | 0.495 | 0.630 | 0.427 | 0.625 | 0.518 |
| Adjusted R-squared | 0.554 | 0.387 | 0.551 | 0.304 | 0.546 | 0.517 |
| Observations | 4775 | 4775 | 4775 | 4775 | 4775 | 4775 |

*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* Column (1) shows results from the linear decomposition of COMPAS: the coefficients from regressing the standardized COMPAS raw score on all variables including census tract fixed effects. Successive columns (2) - (6) show the coefficients from the linear decomposition, removing a set of variables. Sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data (including indicator variable for census tracts with missing census tract-level data).

Table 3 shows the results from a linear decomposition of the standardized COMPAS risk score using census tract-level variables (ratio black, Hispanic, with income below poverty level, with a bachelors or associates degree, and labor force participation). Comparing Column (1) and (6) shows that including neighborhood-level variables increases overall explanatory power from 0.517 to 0.537; 3.7 percent of the overall explained variance in COMPAS risk scores can be attributed to neighborhood-level variables. Moving from a neighborhood with 14% less people with incomes below the poverty level has the same weight in the COMPAS score as having 1 less prior.

**Table 3:** Linear decomposition of standardized COMPAS raw risk score using census tract-level group variables

| | (1) All | (2) All - Priors | (3) All - Juvenile | (4) All - Age controls | (5) All - Charges | (6) All - Census Tract |
|---|---|---|---|---|---|---|
| Priors count | 0.092 | | 0.094 | 0.060 | 0.095 | 0.099 |
| | (0.002) | | (0.002) | (0.003) | (0.002) | (0.002) |
| Juvenile felony count | 0.095 | 0.252 | | 0.176 | 0.095 | 0.104 |
| | (0.031) | (0.044) | | (0.035) | (0.031) | (0.032) |
| Juvenile misdemeanor count | -0.020 | 0.191 | | 0.110 | -0.026 | -0.026 |
| | (0.025) | (0.035) | | (0.040) | (0.025) | (0.025) |
| Juvenile other count | 0.049 | 0.119 | | 0.212 | 0.047 | 0.044 |
| | (0.021) | (0.028) | | (0.033) | (0.021) | (0.020) |
| Age | -0.083 | -0.035 | -0.085 | | -0.088 | -0.086 |
| | (0.005) | (0.006) | (0.005) | | (0.006) | (0.006) |
| Age squared | 0.000 | -0.000 | 0.001 | | 0.001 | 0.001 |
| | (0.000) | (0.000) | (0.000) | | (0.000) | (0.000) |
| Census tract: Ratio black | 0.120 | 0.301 | 0.115 | 0.389 | 0.141 | |
| | (0.057) | (0.064) | (0.057) | (0.070) | (0.058) | |
| Census tract: Ratio hispanic | 0.059 | -0.126 | 0.056 | 0.162 | 0.084 | |
| | (0.085) | (0.093) | (0.085) | (0.103) | (0.087) | |
| Census tract: Ratio income below poverty level | 0.726 | 0.680 | 0.732 | 0.456 | 0.775 | |
| | (0.130) | (0.149) | (0.130) | (0.156) | (0.131) | |
| Census tract: Ratio bachelors/associate's degree or higher (25 and over) | -0.318 | -0.638 | -0.326 | -0.493 | -0.322 | |
| | (0.109) | (0.127) | (0.109) | (0.138) | (0.111) | |
| Census tract: Labor force participation (16 and over) | -0.289 | -0.298 | -0.295 | -0.034 | -0.294 | |
| | (0.077) | (0.089) | (0.077) | (0.098) | (0.078) | |
| Charge degree fixed effects | Y | Y | Y | Y | | Y |
| RMSE | 0.684 | 0.790 | 0.685 | 0.848 | 0.692 | 0.699 |
| R-squared | 0.539 | 0.384 | 0.536 | 0.290 | 0.526 | 0.518 |
| Adjusted R-squared | 0.537 | 0.382 | 0.535 | 0.287 | 0.525 | 0.517 |
| Observations | 4775 | 4775 | 4775 | 4775 | 4775 | 4775 |

*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* Column (1) shows results from the linear decomposition of COMPAS: the coefficients from regressing the standardized COMPAS raw score on all variables including census tract variables. Successive columns (2) - (6) show the coefficients from the linear decomposition, removing a set of variables. Sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data (including indicator variable for census tracts with missing census tract-level data).

It makes sense that census tract fixed effects explain more of the overall variation in COMPAS risk scores because they will pick up COMPAS score unobservables at the neighborhood and individual level that are correlated with census tracts like peer networks. Yet, when I include a limited set of census tract-level variables, they still explain part of the COMPAS score and increase the explanatory power of the model.

Table 4 presents results from decomposing COMPAS risk scores using a second-order polynomial specification that includes up to second order polynomial interactions in all variables except with census tract-level variables and charge-severity fixed effects. Overall, I find similar results in the nonlinear decompositions as in the linear decompositions. Including census tract fixed effects increases the adjusted R-squared from 0.525 in Column (2) to 0.563 in Column (1). Neighborhood fixed effects accounts for 6.7 percent of the total variation explained by the nonlinear model. Including neighborhood-level variables in the model in Column (2) increases the adjusted explanatory power to 0.543, accounting for 3.3 percent of the total variation explained by the nonlinear model.

**Table 4:** Nonlinear decomposition of standardized COMPAS raw risk score using census tract group variables (2nd order polynomial of priors, juvenile history, age, indigent proxy)

| | (1) All (FE) | (2) All (Variables) | (3) All - Census Tract |
|---|---|---|---|
| Census tract: Ratio black | | 0.082 | |
| | | (0.057) | |
| Census tract: Ratio hispanic | | 0.045 | |
| | | (0.085) | |
| Census tract: Ratio income below poverty level | | 0.671 | |
| | | (0.129) | |
| Census tract: Ratio bachelors/associate's degree or higher (25 and over) | | -0.364 | |
| | | (0.108) | |
| Census tract: Labor force participation (16 and over) | | -0.320 | |
| | | (0.076) | |
| Census tract fixed effects | Y | | |
| Charge degree fixed effects | Y | Y | Y |
| Census tract fixed effect 25th Percentile | -0.198 | | |
| Census tract fixed effect 50th Percentile | 0.026 | | |
| Census tract fixed effect 75th Percentile | 0.186 | | |
| RMSE | 0.665 | 0.679 | 0.693 |
| R-squared | 0.641 | 0.546 | 0.527 |
| Adjusted R-squared | 0.563 | 0.543 | 0.525 |
| Observations | 4775 | 4775 | 4775 |

*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* Column (1) shows results from the non linear decomposition of COMPAS: the coefficients from regressing the standardized COMPAS raw score on all variables (charge severity fixed effects, census tract-level fixed effects, second order polynomial of prior count, juvenile history (juvenile felony count, juvenile misdemeanor count, juvenile other count), age controls (age, age squared). Column (2) shows results from the non linear decomposition of COMPAS: the coefficients from regressing the standardized COMPAS raw score on all variables (charge severity fixed effects, census tract-level variables, second order polynomial of prior count, juvenile history (juvenile felony count, juvenile misdemeanor count, juvenile other count), age controls (age, age squared). Column (3) shows results from the non linear decomposition of COMPAS on all variables excluding census tract-level variables. Sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data (including indicator variable for census tracts with missing census tract-level data).

To understand the information contained in neighborhood fixed effects, I decompose the estimated census tract fixed effects on observed neighborhood characteristics in Appendix Table B1. I find that estimated neighborhood fixed effects are partly explained by neighborhood racial composition, poverty status, education levels, and labor force participation.

## 5.2 Detailed Decomposition across Groups

Figure 2 shows how input variables explain racial and economic differences in COMPAS scores. I find that neighborhood fixed effects explain 23.9% of the actual black-white gap in COMPAS scores. Using neighborhood-level variables in place of fixed effects, I find that neighborhood-level variables explain 21.0% of the actual black-white gap in COMPAS scores. I find similar results decomposing the estimated black-white gap rather than the actual black-white gap. Using a nonlinear model of COMPAS risk scores to decompose the black-white

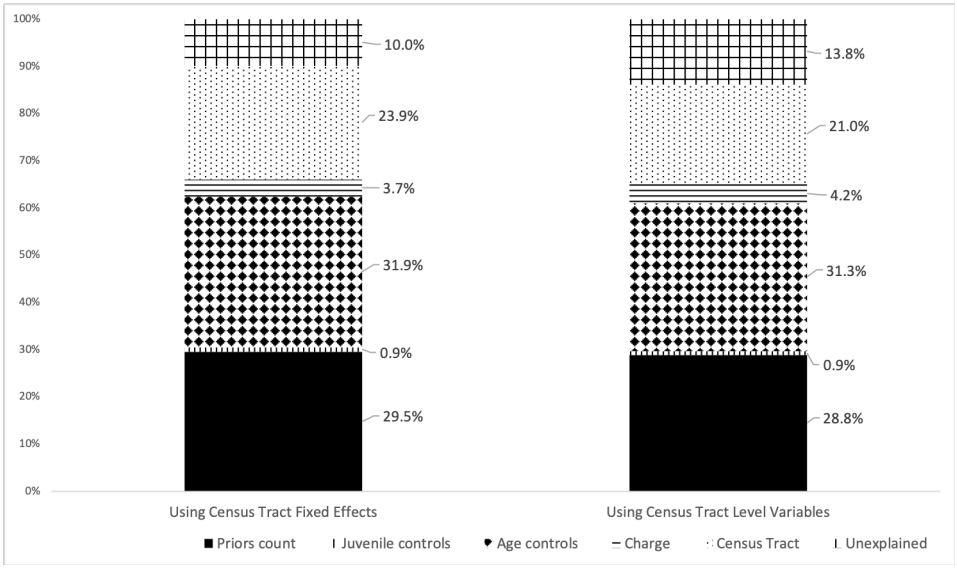
gap yields similar results: neighborhood fixed effects explain 21.4% of the actual black-white gap in COMPAS risk scores.[28] Neighborhood-level variables explain 19.2% of the actual black-white gap in COMPAS risk scores.[29]

[28] $(0.063 - -0.084)/(0.318 - -0.369)$

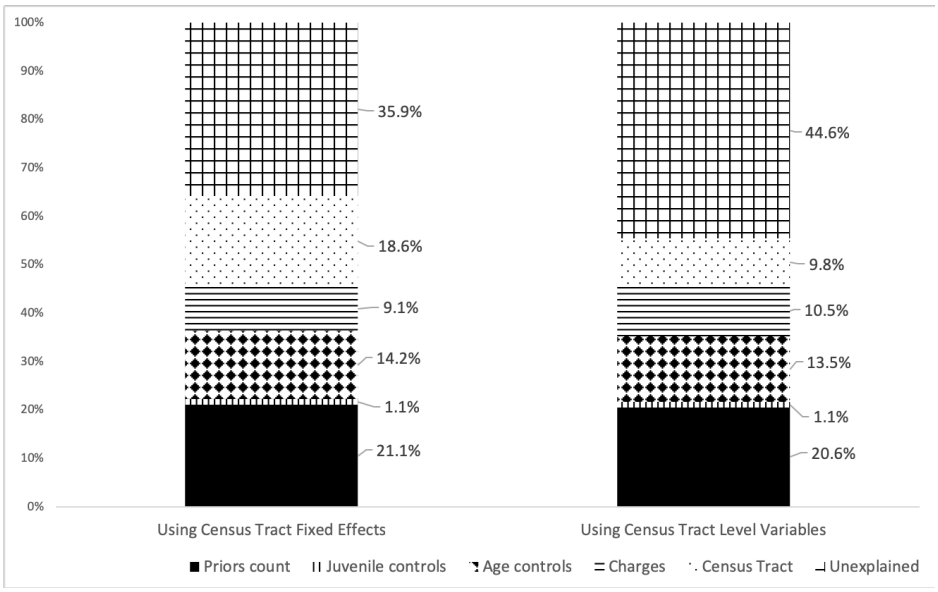
[29] $(0.082 * (0.610 - 0.228) + 0.045 * (0.172 - 0.259) + 0.671 * (0.248 - 0.158) + -0.364 * (0.262 - 0.379) + -0.320(0.342 - 0.347))/(0.318 - -0.369)$

**Figure 2:** Decomposing the actual black-white and indigent-non-indigent COMPAS risk score gaps

Black-white COMPAS risk score gap




Indigent-non-indigent COMPAS risk score gap




*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year.
*Notes:* Sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data. The actual black-white gap in standardized COMPAS scores is 0.687. The estimated black-white gap is 0.626 using the model with census tract fixed effects and 0.600 using the model with census tract-level variables. The actual indigent-non-indigent gap in standardized COMPAS scores is 0.514. The estimated indigent-non-indigent gap is 0.515 using the model with census tract fixed effects and 0.515 using the model with census tract-level variables.

Neighborhood fixed effects explain 18.6% of the actual indigent-non-indigent gap in COMPAS scores. These neighborhood-level variables explain 9.8% of the actual indigent-non-indigent gap in COMPAS scores. I find similar results decomposing the estimated indigent-non-indigent COMPAS score gap. Using a nonlinear framework, I find that neighborhood fixed effects explain 15.6% of the actual indigent -non-indigent gap in COMPAS scores.[30] I also find that neighborhood-level variables explain 9.2% of the actual indigent -non-indigent gap in COMPAS scores.[31] I find qualitatively similar results using a multi-stage decomposition framework in Appendix C. I also find qualitatively similar results using the indigent proxy to decompose black-white gaps COMPAS score gaps (Appendix Figure B3 and Appendix Figure B4) and indigent-non-indigent COMPAS score gaps (Appendix Figure B5 and Appendix Figure B6).

While census tract-level variables contribute only marginally to predictive power, they explain substantial proportions of the average black-white and indigent-non-indigent gaps in COMPAS scores. Neighborhood differences across race and indigent status disproportionately contribute to explaining differences in risk scores across race, compared to how neighborhood variables explain overall variance in risk scores. That is, black and indigent defendants have disproportionately higher risk scores, and disproportionately bear the cost of marginally increasing predictive power.

## 5.3 Overall Predictive Power

This section measures how individual factors contribute to overall predictive power out-of-sample using cross-validation. Table 5 shows the results from linear and nonlinear logistic prediction of recidivism. Column (1) "All" shows the predictive performance of the model including all variables (and interactions for nonlinear models). Each successive column (2) to (6) shows the predictive performance of models excluding a set of variables. The marginal increase in predictive power of a set of variables is the change in the pseudo R-squared of the model with all variables with the pseudo R-squared of the model from all variables excluding one set of variables. For example, the marginal increase in predictive power of census tract-level variables is Column (1) - Column (2). The analysis focuses on the predictive power of census tract-level variables, as there are many census tract fixed effects to estimate with a comparatively small dataset.[32]

[30] $(0.040 - -0.040)/(0.263 - -0.251)$

[31] $(0.082 * (0.498 - 0.404) + 0.045 * (0.206 - 0.229) + 0.671 * (0.229 - 0.194) + -0.364 * (0.290 - 0.326) + -0.320(0.339 - 0.352))/(0.263 - -0.251)$

[32] I find that census tract fixed effects decrease the Pseudo R-squared in out-of-sample cross validation (Appendix Table B2). The decrease in the Pseudo R-squared could because the model is estimating many fixed effects with a relatively small dataset.

**Table 5:** Logistic regression prediction of recidivism (2y): Cross-Validation

| | (1) All | (2) - Priors | (3) - Juvenile | (4) - Age controls | (5) -Charges | (6) -Census Tract |
|---|---|---|---|---|---|---|
| *Panel A. Linear specification* | | | | | | |
| 5-fold Cross-Validation: Pseudo R-squared | 0.097 | 0.054 | 0.093 | 0.060 | 0.098 | 0.094 |
| 10-fold Cross-Validation: Pseudo R-squared | 0.095 | 0.059 | 0.096 | 0.061 | 0.098 | 0.095 |
| *Panel B. Second-order polynomial specification* | | | | | | |
| 5-fold Cross-Validation: Pseudo R-squared | 0.083 | - | - | - | - | 0.082 |
| 10-fold Cross-Validation: Pseudo R-squared | 0.085 | - | - | - | - | 0.081 |

*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* Sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data (including indicator variable for census tracts with missing census tract-level data). Outcome variable is recidivism within 2 years conditional on at least 2 years at risk. For the linear functional form, Column (1) "All" controls for charge severity fixed effects, priors count, juvenile history (juvenile felony count, juvenile misdemeanor count, juvenile other count), age, age squared, census tract-level variables (ratio black, ratio Hispanic, ratio income below poverty level, ratio bachelors/associate's degree or higher, labor force participation). For linear functional form, each successive column removes a set of variables. For the nonlinear functional form, Column (1) "All" model allows for charge severity fixed effects, census tract-level variables, and second order polynomial of priors count, juvenile history (juvenile felony count, juvenile misdemeanor count, juvenile other count), age controls (age, age squared). For nonlinear functional form Column (6), the model is all interactions excluding census tract-level variables.

Comparing Column (1) and Column (6) in Table Table 5 reveals there is a negligible increase in predictive power from marginally including neighborhood-level variables. The findings are similar using a nonlinear polynomial functional form of individual characteristics in the logistic regression in Panel B of Table 5. The predictive power of census tract-level variables in predicting recidivism is similar to that of the set of juvenile record variables, charge-severity fixed effects and the indigent proxy, which all seem to have only negligible predictive power (Panel A of Table 5). Including age controls (age and age-squared) and the priors count increases the Pseudo R-squared by 0.034 to 0.037 and 0.036 to 0.043.

## 5.4 Out-of-sample Prediction by Group

In this section, I assess out-of-sample predictive power overall, and across race and indigent status of models including and excluding census tract-level variables. Table 6 shows the correct prediction, false positive, and false negative rates Of COMPAS risk scores and the COMPAS risk scores residualized of census tract-level variables (from Equation 6). I find that overall predictive power marginally increases by 0.3 percentage points. Looking separately by racial group, overall predictive power decreases for blacks by 1.1 percentage points and increases for whites by 2.1 percentage points. Specifically, including census tract variables leads to relatively worse outcomes for blacks (higher over-prediction and lower under-prediction), and relatively better outcomes for whites (lower over-prediction and higher under-prediction). For blacks, including census tract-level variables increases the false positive rate by 3.2 percentage points and only slightly decreases the false negative rate by

1.8 percentage points. On the other hand, for whites, the false positive rate decreases by 5.2 percentage points and increases the false negative rate by 5.2 percentage points. Results are qualitatively the same using a nonlinear model to decompose COMPAS and for prediction in Table 7.

**Table 6:** Out-of-sample Prediction Rates of COMPAS and $ResidScore_{-N}$ (COMPAS residualized of census tract-level variables)

| | Correct | False Positive | False Negative |
|---|---|---|---|
| | (1) | (2) | (3) |
| *Overall* | | | |
| COMPAS | 0.628 | 0.378 | 0.361 |
| $ResidScore_{-N}$ | 0.625 | 0.382 | 0.361 |
| | | | |
| *By race* | | | |
| COMPAS: Black | 0.590 | 0.507 | 0.281 |
| $ResidScore_{-N}$: Black | 0.600 | 0.475 | 0.299 |
| COMPAS: White | 0.685 | 0.219 | 0.542 |
| $ResidScore_{-N}$: White | 0.664 | 0.271 | 0.490 |
| | | | |
| *By economic status* | | | |
| COMPAS: Indigent | 0.607 | 0.481 | 0.258 |
| $ResidScore_{-N}$: Indigent | 0.609 | 0.476 | 0.258 |
| COMPAS: Non-indigent | 0.649 | 0.286 | 0.474 |
| $ResidScore_{-N}$: Non-indigent | 0.641 | 0.298 | 0.474 |

*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* $ResidScore_{-N}$ is the COMPAS score residualized of census tract-level variables. The weight given to census tract-level variables is estimated from Equation 2 in Section 4.1 using a random 50% of the sample. The other 50% sample is used to predict risk scores and assess out-of sample performance of COMPAS and $ResidScore_{-N}$. The total sample is all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data. Correct prediction is the (# of true positives cases + # of true negatives cases)/(# of defendants). False positive rate is the (# of false positive cases)/(# of defendants who do not recidivate). False negative rate is the (# of false negative cases)/(# of defendants who recidivate). Indigent status proxy is 1 if defendant qualifies for a public defender. Recidivism (2 year) outcome is recidivism within 2 years conditional on the defendant having at least 2 years at risk.

**Table 7:** Out-of-sample Prediction Rates of COMPAS and $ResidScore_{-N}$ (COMPAS residualized of census tract-level variables using a second-order polynomial specification)

| | Correct | False Positive | False Negative |
|---|---|---|---|
| | (1) | (2) | (3) |
| *Overall* | | | |
| COMPAS | 0.628 | 0.378 | 0.361 |
| $ResidScore_{-N}$ | 0.624 | 0.384 | 0.361 |
| | | | |
| *By race* | | | |
| COMPAS: Black | 0.590 | 0.507 | 0.281 |
| $ResidScore_{-N}$: Black | 0.598 | 0.478 | 0.299 |
| COMPAS: White | 0.685 | 0.219 | 0.542 |
| $ResidScore_{-N}$: White | 0.664 | 0.271 | 0.490 |
| | | | |
| *By economic status* | | | |
| COMPAS: Indigent | 0.607 | 0.481 | 0.258 |
| $ResidScore_{-N}$: Indigent | 0.608 | 0.478 | 0.258 |
| COMPAS: Non-indigent | 0.649 | 0.286 | 0.474 |
| $ResidScore_{-N}$: Non-indigent | 0.640 | 0.300 | 0.474 |

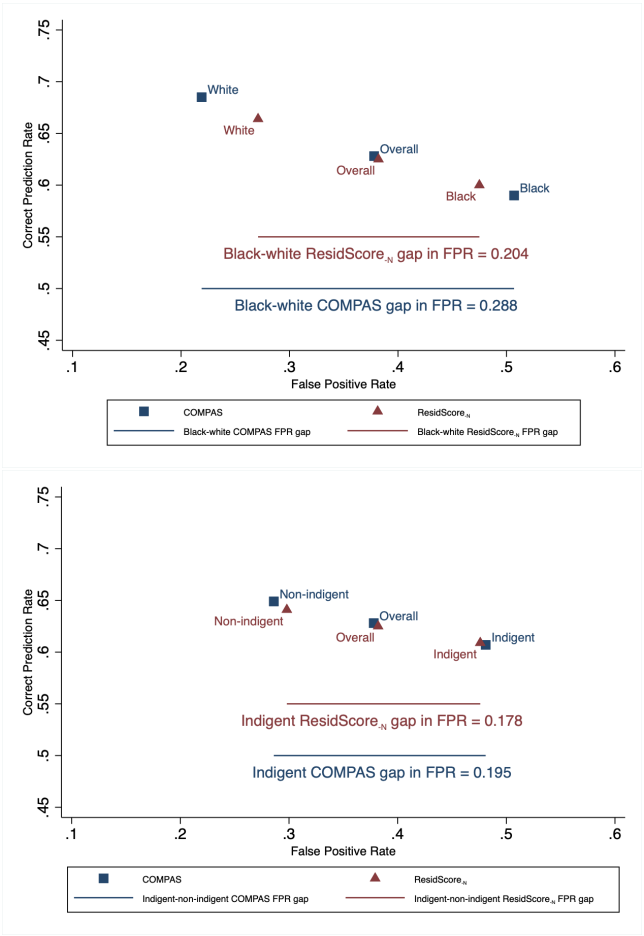
*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* $ResidScore_{-N}$ is the COMPAS score residualized of census tract-level variables. The weight given to census tract-level variables is estimated from Equation 2 in Section 4.1 using a random 50% of the sample. I use a nonlinear specification that includes up to second order polynomial interactions of all individual characteristics excluding charge-severity fixed effects. The other 50% sample is used to predict risk scores and assess out-of sample performance of COMPAS and $ResidScore_{-N}$. The total sample is all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data. Correct prediction is the (# of true positives cases + # of true negatives cases)/(# of defendants). False positive rate is the (# of false positive cases)/(# of defendants who do not recidivate). False negative rate is the (# of false negative cases)/(# of defendants who recidivate). Indigent status proxy is 1 if defendant qualifies for a public defender. Recidivism (2 year) outcome is recidivism within 2 years conditional on the defendant having at least 2 years at risk.

By indigent status, including census tract variables slightly decreases the correct prediction rate by 0.2 percentage points for indigent defendants but increases the correct prediction by 0.8 percentage points for non-indigent defendants. Indigent defendants have slightly

higher false positives rates and lower true negatives rates, and non-indigent defendants have less false positives, with COMPAS compared to the COMPAS score residualized of defendants' census tract variables. Results from including census tract fixed effects are qualitatively similar comparing prediction across groups of COMPAS risk scores to (1) COMPAS risk scores residualized of census tract fixed effects (using linear case of Equation 5) in Table B5. Results are also qualitatively similar using a second-order specification to residualize COMPAS risk scores residualized of census tract fixed effects (using second-order specification of Equation 5) in Table B6. However, when using census tract fixed effects instead of census tract-level variable, overall predictive power marginally decreases by 0.3-0.5 percentage points. This finding is consistent with results from logistic regression prediction of recidivism using census tract fixed effects in Section 5.3. This is likely because of the large number of fixed effects being estimated with a small data set.

Figure 3 shows the correct prediction rates and false positive rate overall, and by race and indigent status. Defendants' neighborhood data over-predicts for black defendants who do not recidivate, and substantially widens black-white differences in false positives rates. Including defendants' neighborhood data also slightly widens economic differences in false positives rates.

**Figure 3:** Comparing out-of-sample prediction rates of COMPAS and $ResidScore_{-N}$ (COMPAS score residualized of census tract-level variables) by race and indigent status




*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* Correct prediction is the (number of true positives + true negatives)/(number of defendants). False positive rate is the (number of false positives)/(number of defendants do not recidivate).

## 5.5 Discussion

Overall, census tract-level variables contribute marginally to predictive power overall, but predict unequally across race and socioeconomic status. The former finding is consistent across Section 5.4 which compares the overall predictive power of models including and excluding census tract-level variables in recidivism prediction, and Section 5.3 which compares out-of-sample prediction of COMPAS and COMPAS residualized of census tract-level vari-

ables.[33] Section 5.3 finds that including these variables in scores leads to more over-prediction for black and indigent defendants, and less over-prediction for white and non-indigent defendants. Additionally, there is also slightly less under-prediction for black defendants, and slightly more under-prediction for white defendants. Decision-making by judges is determined by the decile score or the Low-Medium-High score, which all depend on a defendant's position within the distribution of predicted recidivism. Therefore, variables that unequally predict across race and socioeconomic status make blacks and indigent relatively worse off in the distribution of defendant scores.

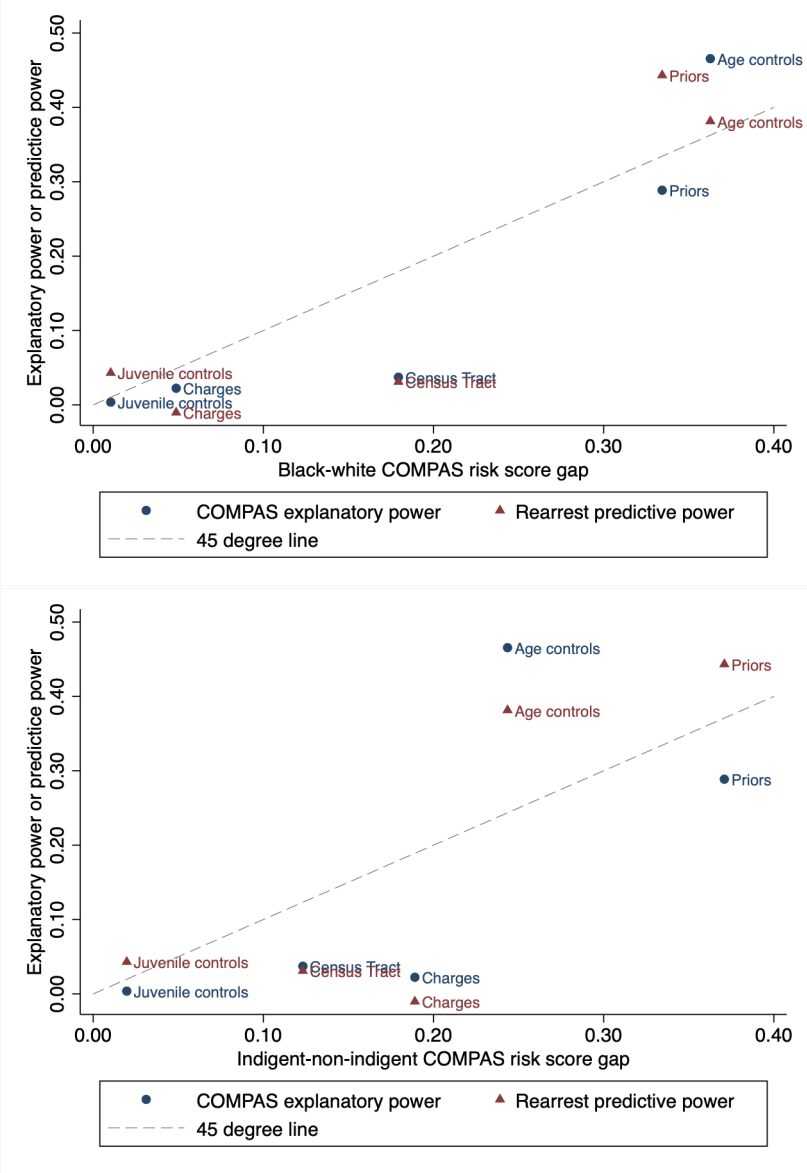
Census tract-level variables marginally contribute to overall predictive power but greatly contribute to black-white and indigent-non-indigent COMPAS score gaps. Figure 4 shows this imbalance of census tract-level variables (from the 45 degree line) compared to other variables (juvenile controls, charges, prior counts and age controls) with the black-white COMPAS score gap. A similar story emerges with the indigent-non-indigent COMPAS gap. Charge severity fixed effects are another component that can be considered unbalanced in how it contributes to overall predictive power and the indigent-non-indigent risk score gap. Moreover, these variables also contribute disproportionately compared to how they explain the overall COMPAS score.

Taken together, census tract-level variables contribute differential predictive power across race and indigent status by relatively increasing the false positive rate for black and indigent defendants (given higher scores) and decreasing the false negative rate for white and indigent defendants (given lower scores).[34] Not using census tract-level variables would slightly reduce overall predictive power but greatly reduce black-white and indigent-non-indigent risk score differences.

[33] A key difference between Section 5.4 and Section 5.3 is how census tract-level variables are treated. Section 5.4 excludes census tract-level variables from the model and other variables coefficients can "re-weight" once census-tract-level variables are excluded from the model. Section 5.3 decomposes COMPAS using all variables including census tract-level variables (other variables will not "re-weight"), and residualizes COMPAS scores of census tract-level variables by subtracting their contribution from COMPAS scores. Findings are consistent across both treatments.

[34] This is also consistent with the findings of the Appendix C that the census tract-level variables disproportionately contribute to average COMPAS risk score gaps compared to average recidivism gaps. Neighborhood variables explain more of the black-white and indigent-non-indigent COMPAS score gaps than the recidivism black-white and indigent-non-indigent gaps. Using census tract-level variables in recidivism prediction leads to more over-prediction of black and indigent defendants scores, who then have higher scores when they do not recidivate. Whereas, whites have less over-prediction and more under-prediction, which results in lower scores for those who recidivate relatively more.

**Figure 4:** Contribution of components to overall explanatory and predictive power vs. black-white risk score gap and indigent-non-indigent risk score gap (using census track-level variables)




*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* Contribution of a set of variables to overall explanatory power is the fraction of the adjusted R-squared that is attributed to marginally including a set of variables in the linear decomposition from Table 2. Specifically, it is (adjusted R-squared with component - adjusted R-squared without component) as a fraction of the adjusted R-squared with the component. Contribution of a set of variables to overall explanatory power is the fraction of the adjusted R-squared that is attributed to marginally including a set of variables in the logistic regression cross validation exercise from Table 5. Specifically, it is (pseudo R-squared with component - pseudo R-squared component) as a fraction of the pseudo R-squared with the component. The contribution of a set of variables to the group difference of the risk score gap is the fraction of the group difference in risk score gap attributed to a set of variables, from Equation 4.

# 6 Policy Implications

In the context of COMPAS, my analysis shows that not using variables like defendants' residential neighborhood can substantially reduce group inequities while marginally decreasing how COMPAS predicts recidivism overall. Therefore, a policy implication is to carefully consider exactly how each variable, and the overall system, balances this trade-off by carefully considering all input variables. Algorithm design may not explicitly account for how to balance this trade-off between predictive power and group equity. Data inputs that slightly improve recidivism prediction may predict unequally and introduce disproportionately large disparities across groups in scores.

While I explicitly study this trade-off in COMPAS, this trade-off between individual and group prediction can exist in other tools used in recidivism prediction, in the criminal justice system, and society more broadly. At least three other popular risk assessment tools – Level of Service Inventory-Revised (LSI-R), Ohio Risk Assessment System (ORAS) and Wisconsin Risk-Needs Assessment –include companions and associates of assessed individuals as dynamic risk factors; ORAS also includes neighborhoods as a dynamic risk factor (Taxman and Pattavina, 2013; Byrne and Pattavina, 2017). Defendant residential neighborhoods are just one example of a group variable that is contentious – defendants can be made worse off on the basis of where they live, rather than the actions that they have committed. The controversial nature of using defendant neighborhoods is compounded by the fact that defendant neighborhoods only marginally increase predictive power and induce substantial disparities in defendants' scores across race and economic class groups.

Another implication arises as the tools are presented as race neutral, and used in contexts where policy makers are concerned about discrimination. Proprietary tools, such as COMPAS, are not transparent about model inputs and specifications. With transparency about models and input variables, users can be more aware of this trade-off that tools are implicitly making. However, users still not be able to ascertain exactly how tools implicitly trade-off predictive power and other concerns about group disparities, and in particular racial disparities.[35] Policy makers can account for this trade-off explicitly and reconsider machine learning system optimization priorities to maximize predictive power. Pope and Sydnor (2011) propose a method to harness the predictive power of proxy variables outside of the predictive power of protected classes. They argue this is a "reasonable compromise that helps to satisfy demands for fair treatment without drastically reducing the predictive power of statistical models." It is also possible to preprocess contentious variables by resid-

[35]Stevenson and Slobogin (2018) make an analogous argument for transparency in risk assessment: judges will not understand the weight of age which is problematic given the "double-edged sword" of age and the role that it plays in risk assessment.

ualizing them of protected groups like race, and include other variables that are part of an individual's record in the criminal justice system.

The COMPAS recidivism score predicts rearrest, which may be a racially biased outcome if there are disparities in policing or intensity of policing by neighborhood. Pretrial decision-making is mandated to minimize any negative effects of *future crime* by released defendants on society, and terms of pretrial release are to be decided in consideration of public safety (Karnow, 2008). Ideally, we would be predicting whether a crime was *actually* committed rather than re-arrest when considering the recidivism risk that a defendant poses. This concern about the outcome being predicted may apply to other settings where tools are used in the criminal justice system, as well as outside of the criminal justice system.

As COMPAS tools are used in pretrial decision-making, they have the potential to impact an individual's subsequent outcomes. It is unclear how exactly group disparities in risk scores may interact with judge decision-making. Holding judge decision-making constant, this paper's findings may be compounded if having a higher risk score results in longer pretrial detention length.[36] Using data from Broward County, Cowgill (2018) finds that being at the margin of COMPAS scores has an impact on pretrial detention length, and also slightly increases recidivism. More generally, studies have shown the causal effect of bail decisions on case decisions including conviction (Stevenson, 2017; Gupta et al., 2016; Leslie and Pope, 2017; Dobbie et al., 2018). Across different contexts, these studies find that pretrial detention increases the likelihood of conviction by 6-13 percentage points, and also increases the sentence length (Didwania, 2018). Taken together with my findings that black and indigent defendants have disproportionately higher risk scores, risk scores may further compound defendants' subsequent case outcomes in the criminal justice system.

Moreover, racial and economic disparities in risk scores also have the potential to amplify pre-existing inequality in other related settings. For example, if risk scores have an impact on pretrial detention length, there could also be an impact on labor market outcomes. Dobbie et al. (2018) find that pretrial detention in excess of three days does not increase recidivism two years after release, and that pretrial detention causally decreases defendant formal employment and take up of government benefits post release.[37] Having higher risk scores can impact their labor market outcomes if defendants are detained rather than released during pretrial proceedings. Therefore, the costs of score disparities can also include higher costs

[36]Judge decision-making could interact with risk scores differently depending on the information that judges have about risk score input variables and models. However, COMPAS, and other proprietary tools are not transparent about model inputs and methods.

[37]Dobbie et al. (2018) show that defendants who are at the margin of being released within three days of pretrial detention are 9.4 percentage points more likely to be employed in the formal sector, earning $948 per year, and 10.7 percentage points more likely to have any income, 3-4 years after bail.

of pretrial detention, and subsequent labor market costs and incarceration costs from higher conviction rates.[38] Beyond pretrial decisions, there is further evidence on the causal impact of incarceration on crime and employment outcomes, suggesting that any group disparities in algorithms used in decision-making in the criminal justice system can also have negative downstream consequences.[39] Due to the link between incarceration and labor market outcomes, disparities in risk scores can exacerbate labor market outcomes and subsequent recidivism, creating a vicious cycle. Tools are used in a variety of different contexts related to economic outcomes and economic inequality, from loan allocation to hiring skill assessment tools. Concerns about group disparities in risk scores are particularly important given the widespread use of data-driven tools.

# 7 Conclusions

In this paper, I study an implicit trade-off in data-driven prediction models between predictive power and equity across groups. I investigate this trade-off in the COMPAS recidivism risk tool, using risk score and defendant characteristic data over 2013-2016 from Broward County, Florida. I examine defendant neighborhoods as a proxy for race and socioeconomic status, and find that defendant neighborhood variables explain disproportionately more of average racial and economic COMPAS gaps compared to how they explain overall variation in a defendant's risk scores. While defendant census tract-level variables add negligible predictive power, they predict unequally across race and socioeconomic status, and substantially widen differences in COMPAS scores across race and economic status and false positive rates across race.

These findings have policy implications, as substantial group disparities in scores could amplify existing inequality through the downstream consequences of having a higher COMPAS score. When is it warranted to use factors correlated with race to increase the predictive power of a data-driven algorithm? More broadly, there are many policy implications to understanding the impacts of recently deployed decision systems in society. These research questions contribute to understanding the larger issue of if and how data-driven algorithms, e.g. artificial intelligence technologies, reinforce and worsen, or ameliorate preexisting inequality in society. As data-driven algorithms are utilized to make decisions with increasingly far-reaching societal implications, the ability to audit their predictions and conclusions

[38]The cost of risk score disparities can be quantified in part by labor market income. There can be other costs due to longer pretrial detention, which can in turn lead to higher rates of conviction, and longer subsequent incarceration as a result.

[39]Mueller-Smith (2015) finds each year of incarceration increases future crime and decreases post-release employment by 3.6 percentage points, while Kling (2006) finds no effects.

while understanding fairness across groups is vital (Hardt et al., 2016; Jagtap and Sane, 2014).

# References


Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. (2017). Learning certifiably optimal rule lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 35–44. ACM.

Angwin, J. (2016). The Hidden Discrimination In Criminal Risk-Assessment Scores. Interview by Kelly McEvers and Audie Cornish, NPR. Source: http://www.npr.org/2016/05/24/479349654/the-hidden-discrimination-in-criminal-risk-assessment-scores. Accessed on May 6, 2017.

Anwar, S., Bayer, P., and Hjalmarsson, R. (2012). The Impact of Jury Race in Criminal Trials. *Quarterly Journal of Economics*, 127(2):1017–1055.

Anwar, S. and Fang, H. (2006). An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *American Economic Review*, 96(1):127–151.

Anwar, S. and Fang, H. (2015). Testing for racial prejudice in the parole board release process: Theory and evidence. *The Journal of Legal Studies*, 44(1):1–37.

Arnold, D., Dobbie, W., and Yang, C. S. (2017). Racial Bias in Bail Decisions. Mimeo. Source: https://www.princeton.edu/ wdobbie/files/racialbias.pdf. Accessed on May 6, 2017.

Ayres, I. (2002). Outcome tests of racial disparities in police practices. *Justice research and Policy*, 4(1-2):131–142.

Bayer, P., Hjalmarsson, R., and Pozen, D. (2009). Building criminal capital behind bars: Peer effects in juvenile corrections. *The Quarterly Journal of Economics*, 124(1):105–147.

Berk, R. (2009). The role of race in forecasts of violent crime. *Race and social problems*, 1(4):231.

Berk, R. (2017). An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology*, 13(2):193–216.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2017). Fairness in criminal justice risk assessments: the state of the art. *arXiv preprint arXiv:1703.09207*.

Brennan, T., Breitenbach, M., and Dieterich, W. (2008). Towards an explanatory taxonomy of adolescent delinquents: Identifying several social-psychological profiles. *Journal of Quantitative Criminology*, 24(2):179–203.

Brennan, T., Dieterich, W., and Ehret, B. (2009). Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36(1):21–40.

Bushway, S. and Smith, J. (2007). Sentencing using statistical treatment rules: what we don't know can hurt us. *Journal of Quantitative Criminology*, 23(4):377–387.

Byrne, J. and Pattavina, A. (2017). Next generation assessment technology: The potential and pitfalls of integrating individual and community risk assessment. *Probation Journal*, 64(3):242–255.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. pages 1–17.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM.

Cowgill, B. (2018). The impact of algorithms on judicial discretion: Evidence from regression discontinuities.

Datta, A., Tschantz, M. C., and Datta, A. (2015). Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies*, 1(92–112).

Didwania, S. H. (2018). The immediate consequences of pretrial detention: Evidence from federal criminal cases.

Dobbie, W., Goldin, J., and Yang, C. S. (2018). The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, 108(2):201–40.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580.

Durlauf, S. N. (2006). Assessing racial profiling. *The Economic Journal*, 116(515):F402–F426.

Eubanks, V. (2018). Automating inequality: how high-tech tools profile, police, and punish the poor.

Fang, H. and Moro, A. (2011). Theories of statistical discrimination and affirmative action: A survey. In *Handbook of social economics*, volume 1, pages 133–200. Elsevier.

Flores, A. W., Lowenkamp, C. T., and Bechtel, K. (2016). 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.". *Federal Probation*, 80(2).

Frey, W. H. and Myers, D. (2005). Racial segregation in us metropolitan areas and cities, 1990–2000: Patterns, trends, and explanations. *Population studies center research report*, (05-573).

Garrett, B. L. and Monahan, J. (2018). Judging risk. *Virginia Public Law and Legal Theory Research Paper No. 2018-44*.

Graham, B. S. (2018). Identifying and estimating neighborhood effects. *Journal of Economic Literature*, 56(2):450–500.

Gupta, A., Hansman, C., and Frenchman, E. (2016). The heavy costs of high bail: Evidence from judge randomization. *The Journal of Legal Studies*, 45(2):471–505.

Hardt, M., Price, E., Srebro, N., et al. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction, springer series in statistics.

Jagtap, D. and Sane, S. S. (2014). Direct discrimination aware data mining. *International Journal of Computer Applications*, 95(25).

Karnow, C. E. (2008). Setting bail for public safety. *Berkeley J. Crim. L.*, 13:1.

Kay, M., Matuszek, C., and Munson, S. A. (2015). Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828. ACM.

Kirk, D. S. (2019). The effects of neighborhood context and residential mobility on criminal persistence and desistance. In *The Oxford handbook of developmental and life-course criminology*.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. (2018). Algorithmic fairness. In *AEA Papers and Proceedings*, volume 108, pages 22–27.

Kleinberg, J. and Mullainathan, S. (2019). Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability. Technical report, National Bureau of Economic Research.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. Mimeo. Source: https://arxiv.org/abs/1609.05807. Accessed on May 6, 2017.

Kling, J. R. (2006). Incarceration length, employment, and earnings. *American Economic Review*, 96(3):863–876.

Knowles, J., Persico, N., and Todd, P. (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109(1):203–229.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm. ProPublica.

Leslie, E. and Pope, N. G. (2017). The unintended impact of pretrial detention on case outcomes: Evidence from new york city arraignments. *The Journal of Law and Economics*, 60(3):529–557.

Liptak, A. (2017). Sent to prison by a software program's secret algorithms. *The New York Times*.

Lum, K. and Isaac, W. (2016). To predict and serve? *Significance*, 13(5):14–19.

Manson, S., Schroeder, J., Van Riper, D., and Ruggles, S. (2017). Ipums national historical geographic information system: Version 12.0 [database]. Technical report, University of Minnesota.

Mechoulan, S. and Sahuguet, N. (2015). Assessing racial disparities in parole release. *The Journal of Legal Studies*, 44(1):39–74.

Mueller-Smith, M. (2015). The criminal and labor market impacts of incarceration. *Unpublished Working Paper*.

Northpointe (2012). Practitioners' Guide to COMPAS. Source: http://www.northpointeinc.com/files/technical_documents/FieldGuide2_081412.pdf.

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pope, D. G. and Sydnor, J. R. (2011). Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal: Economic Policy*, 3(3):206–31.

Reardon, S. F., Fox, L., and Townsend, J. (2015). Neighborhood income composition by household race and income, 1990–2009. *The Annals of the American Academy of Political and Social Science*, 660(1):78–97.

Sorensen, T., Sarnikar, S., and Oaxaca, R. L. (2012). Race and gender differences under federal sentencing guidelines. *American Economic Review*, 102(3):256–60.

Stahler, G. J., Mennis, J., Belenko, S., Welsh, W. N., Hiller, M. L., and Zajac, G. (2013). Predicting Recidivism for Released State Prison Offenders: Examining the Influence of Individual and Neighborhood Characteristics and Spatial Contagion on the Likelihood of Reincarceration. *Criminal Justice Behavior*, 40(6):690–711.

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stan. L. Rev.*, 66:803.

Stevenson, M. (2017). Distortion of justice: How the inability to pay bail affects case outcomes.

Stevenson, M. (2018). Assessing risk assessment in action. *Minn. L. Rev.*, 103:303.

Stevenson, M. T. and Doleac, J. L. (November 2018). Roadblock to reform.

Stevenson, M. T. and Slobogin, C. (2018). Algorithmic risk assessments and the double-edged sword of youth. *Behavioral sciences & the law*, 36(5):638–656.

Sweeney, L. (2013). Discrimination in Online Ad Delivery. *Communications ACM. New York, NY, USA: ACM*, 5(56):44–54.

Taxman, F. S. and Pattavina, A. (2013). *Simulation Strategies to Reduce Recidivism*. Springer.

# A Appendix: Data

### Data construction

I use the Broward County Clerk Commercial Data application programming interface (API) to gather criminal case details (addresses and information about public defenders) of all 13,083 COMPAS pretrial screenings who were COMPAS assessed in Broward County over 2013-2014 (from a FOIA dataset from ProPublica).[40] I match criminal cases with 9,717 COMPAS screenings using first, last name (wild card), date of birth (with 3 days flexibility), and whether COMPAS screening is within 30 days of any arrest associated with a case (following Larson et al. (2016)).

Of the 9,717 COMPAS screenings matched with criminal cases, 9,335 have addresses in their criminal cases.[41] For the 9,335 cases with addresses, I use the Census Batch Geocoder and Google Sheets Geocode Cells Utility and Two-Way Geocoding tool to geocode 8,386 cases to census tracts.[42] I exclude 3 defendants who have addresses outside the US.

After obtaining the Census state, county and tract codes, I merge the cases with tract-level Census 2010 data (2010 Census: SF 1a - P& H Tables [Blocks & Larger Areas]) and 2012-2016 (5-year) ACS data obtained using IPUMS. There are currently 1,324 cases that are mapped to census tracts that do not match with the IPUMS census data.[43] Cases excluded from analysis because census tract data are not matched are composed of more White defendants than the whole sample. Overall, Black, Hispanic and White defendants live in 1,176 census tracts. Nine hundred and sixty-six census tracts have census data and are included in the analysis. Only 740 census tracts are used in the analysis, 352 of which are singletons where only one defendant is living.

I merge those data with data provided by ProPublica on demographic characteristics (age, marital status, race), initial arrests and charges, subsequent recidivism arrests, and

[40] The FOIA data include 20,281 total COMPAS screenings. I drop observations that correspond to people with multiple COMPAS scores on the same screening day (329 observations). I filter for pretrial screenings (agency text is pretrial, which drops 6,541 observations), for intake as the assessment reason and for the "Risk and Prescreen" scale set (drops 328 observations).

[41] Of the 385 cases without addresses: 185 cases have defendants "at large", 12 defendants have addresses that are listed as unknown, 4 are homeless, 1 refused to answer, and 2 are in custody, and 179 have empty address fields.

[42] I use the Census Batch Geocoder to map addresses that are in a consistent format (9,219), and map 8,030 to census tracts. I map 1) incomplete or inconsistently formatted addresses (113 cases) and 2) addresses that do not match using Census Geocoder as extracted from the API (1,189) to GPS coordinates using the "Geocode Cells" utility in Google Sheets. Using the "Two-way Geocoding" tool in Google Sheets, I map them back to addresses so the addresses are in a more consistent format, and then feed them into the Census Geocoder.

[43] I double checked these cases using the Census Geocoder.

criminal record and juvenile history for 11,757 individuals.[44] This yields a dataset of 8,419 defendants. Of the 8,419 defendants, I filter for the 6,246 men who are Black, Hispanic and White. Nearly 6000 (5,805, to be exact) have addresses that are successfully geocoded and have census tract codes, and 4,839 have census tract-level Census 2010 and ACS 2012-2016 data. 441 defendants do not have census tract codes because 155 have no address in the Broward County Clerk website that the Commercial Data API can retrieve and 286 have addresses that are not geocoded. Of the 155 defendants who do not have an address, 126 defendants are "at large", 6 defendants have addresses that are listed as unknown, 3 are homeless, 1 refused to answer, 1 is in custody, and 18 have empty address fields.[45]

## A.1 Differences in recidivism definitions

ProPublica defines recidivism as "a criminal offense that resulted in a jail booking and took place after the crime for which the person was COMPAS scored" (which they claim is consistent with Northpointe's definition of recidivism). While I consider the same recidivism events as ProPublica in defining recidivism outcomes, ProPublica and Flores et al. (2016) define their two-year recidivism outcome as whether or not the pretrial defendant recidivates within two years or is out of jail for two years. Using this outcome variable, ProPublica creates a sample of defendants who either recidivate within two years or are at risk for two years (two years out of jail). On the other hand, I define two-year recidivism outcomes for *all people who are at risk for the same amount of time (2 years).* I use this definition to create samples of people who are at risk for two years, and use these selected samples in score prediction and outcome tests with my alternative score prediction. This will still be a selected sample (all at risk for two years) of pretrial defendants and will exclude people who are not released for two years during the observed four-year period.

While pretrial defendant COMPAS screening data are from January 2013 to December 2014, the arrest data are from January 2013 to April 2016. Therefore, defendants initially arrested from April 2014 to December 2014 cannot possibly have been out of jail for two years, and are at risk for a shorter time. A concern is that of the defendants initially arrested between April 2014 and December 2014, only those who are rearrested by April 2016 will

[44]This file is entitled "people", a file in the COMPAS sqlite3 database file. ProPublica describes this data base as "containing criminal history, jail and prison time, demographics, and COMPAS risk scores for defendants from Broward County", but does not explicitly say these are the full records for all defendants who were COMPAS scored in 2013-2014.

[45]Dropping further observations: I also drop people for whom the COMPAS decile scores are negative, because the decile score should be between 1 and 10. The raw score maps to a decile score and to a text score. I verify that there are strict cutoffs that map the raw score to the text score, with the exception of two male defendants. I drop these two defendant's observations as they may be due to data errors. I drop (52) observations who are arrested for recidivism while in jail/custody, as they may be data errors.

appear in the data set. However, part of the sample (defendants initially arrested between April 2014-December 2014) are a *further* selected group because of right truncation.

Table A1 compares the "selected" sample to those who do not have at least two years out of jail to recidivate ("rest" sample) out of the full sample of male defendants. I find the selected sample is statistically different on some observables, including whether the defendant qualifies for a public defender ("indigent proxy"), the racial composition of the defendant's census tract, and the count of juvenile misdemeanors.

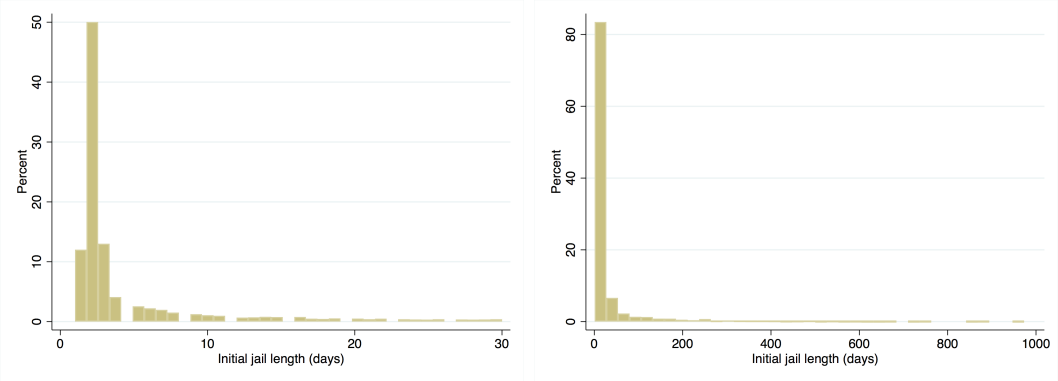
**Table A1:** Descriptive statistics: Selected sample of defendants with 2 years compared to rest of defendants

| | Selected sample N = 3,166 | Rest N = 1,673 | Difference p-value |
|---|---|---|---|
| | Mean. (1) | Mean (2) | *t*-test (3) |
| Age | 34.670 | 34.601 | 0.849 |
| Indigent proxy | 0.494 | 0.592 | 0.000 |
| Ratio White of total population | 0.448 | 0.469 | 0.021 |
| Defendant's census tract: Ratio black | 0.462 | 0.438 | 0.014 |
| Defendant's census tract: Ratio hispanic | 0.214 | 0.223 | 0.069 |
| Defendant's census tract: Ratio income below poverty level | 0.214 | 0.209 | 0.144 |
| Defendant's census tract: Ratio bachelors/associate's degree or higher (25 and o | 0.305 | 0.311 | 0.147 |
| Defendant's census tract: Labor force participation (16 and over) | 0.347 | 0.342 | 0.138 |
| Priors count | 3.474 | 3.317 | 0.295 |
| Juvenile felony count | 0.083 | 0.065 | 0.208 |
| Juvenile misdemeanor count | 0.108 | 0.066 | 0.005 |
| Juvenile other count | 0.117 | 0.124 | 0.647 |
| Felony charge | 0.666 | 0.685 | 0.177 |
| COMPAS Raw Score (Z score) | 0.031 | -0.005 | 0.246 |

*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* The column (1) contains the mean of the subsample of defendants who have atleast two years out of jail to recidivate out of the full sample. The column (2) contains the mean of the subsample of defendants who do not have atleast two years out of jail to recidivate out of the full sample. The column (3) tests the difference between the samples using a two sample t test for equal means and reports the two-sided p-value. Indigent proxy is 1 if the defendant qualifies for a public defender. Ratio Black, ratio Hispanic, ratio income below poverty levels variables are for defendant's census tract.

Figure A1 contains the distribution of initial jail stay length in days. There are a few outliers who are in jail for the majority of the observed period, but most are in jail for less than 10 days.

**Figure A1:** Initial Jail Length (days) Distribution: Men



# B Appendix: Figures and Tables

**Table B1:** Explaining estimated census tract fixed effects with census tract-level characteristics

| | (1) Linear | (2) Nonlinear |
|---|---|---|
| Census tract: Ratio black | 0.1184 | 0.0725 |
| | (0.0327) | (0.0329) |
| Census tract: Ratio hispanic | 0.0688 | 0.0526 |
| | (0.0605) | (0.0601) |
| Census tract: Ratio income below poverty level | 0.7416 | 0.6863 |
| | (0.0605) | (0.0614) |
| Census tract: Ratio bachelors/associate's degree or higher (25 and over) | -0.3047 | -0.3465 |
| | (0.0607) | (0.0599) |
| Census tract: Labor force participation (16 and over) | -0.2900 | -0.3270 |
| | (0.0402) | (0.0399) |
| Constant | -0.0323 | 0.0293 |
| | (0.0452) | (0.0454) |
| R-squared | 0.1907 | 0.1685 |
| Adjusted R-squared | 0.1898 | 0.1676 |
| Observations | 4775 | 4775 |

*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* Sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data (including indicator variable for census tracts with missing census tract-level data). Each specification regresses estimated census tract fixed effects on census tract-level characteristics. In Column (1), census tract fixed effects are estimated from regressing the standardized COMPAS raw score on defendant criminal history (count of priors), juvenile history (juvenile felony count, juvenile misdemeanor count, juvenile other count), age controls (age and age squared), charge degree severity fixed effects, and census tract fixed effects. Column (2) estimates census tract fixed effects from regressing the standardized COMPAS raw score on a charge degree severity fixed effects, census tract fixed effects, and up to all second order polynomial interactions of priors, juvenile history, and age controls (age and age squared). Robust standard errors in parentheses.

**Table B2:** Logistic regression prediction of recidivism (2y): Cross-Validation Exercise

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Priors count | 0.112 | | 0.114 | 0.071 | 0.112 | 0.112 |
| | (0.011) | | (0.011) | (0.010) | (0.011) | (0.010) |
| Juvenile felony count | -0.015 | 0.145 | | 0.061 | -0.019 | 0.004 |
| | (0.076) | (0.075) | | (0.079) | (0.076) | (0.071) |
| Juvenile misdemeanor count | -0.060 | 0.186 | | 0.102 | -0.070 | -0.003 |
| | (0.078) | (0.086) | | (0.096) | (0.076) | (0.078) |
| Juvenile other count | 0.278 | 0.403 | | 0.503 | 0.283 | 0.256 |
| | (0.119) | (0.132) | | (0.132) | (0.118) | (0.101) |
| Age | -0.114 | -0.034 | -0.124 | | -0.112 | -0.114 |
| | (0.027) | (0.026) | (0.027) | | (0.027) | (0.022) |
| Age squared | 0.001 | 0.000 | 0.001 | | 0.001 | 0.001 |
| | (0.000) | (0.000) | (0.000) | | (0.000) | (0.000) |
| Census tract fixed effects | Y | Y | Y | Y | Y | |
| Charge degree fixed effects | Y | Y | Y | Y | | Y |
| 5-fold Cross-Validation: Pseudo R-squared | 0.045 | 0.020 | 0.040 | 0.016 | 0.055 | 0.093 |
| 10-fold Cross-Validation: Pseudo R-squared | 0.057 | 0.022 | 0.052 | 0.024 | 0.055 | 0.096 |
| Observations | 2571 | 2571 | 2571 | 2571 | 2573 | 3100 |

*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* Sample is a sample of black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data (including indicator variable for census tracts with missing census tract-level data).

## B.1 Robustness: Cox proportional hazard models

Using Cox proportional hazard models, I continue to find that priors and age controls have predictive power; however, census tract-level variables have neglible predictive power. Using a linear functional form, including census tract-level variables slightly improves recidivism prediction in Panel A of Table B3. Panel B of Table B3 shows that census tract-level variables actually marginally decrease out of sample predictive power of the hazard model with a nonlinear functional form (Table B3).

**Table B3:** Prediction of recidivism (2y) using Cox proportional hazard model: Cross-Validation

| | (1) All | (2) - Priors | (3) - Juvenile | (4) - Age controls | (5) -Charges | (6) -Census Tract |
|---|---|---|---|---|---|---|
| *Panel A. Linear specification* | | | | | | |
| 5-fold Cross-Validation: Pseudo R-squared | 0.051 | 0.025 | 0.050 | 0.029 | 0.050 | 0.044 |
| 10-fold Cross-Validation: Pseudo R-squared | 0.050 | 0.025 | 0.051 | 0.028 | 0.049 | 0.048 |
| *Panel B. Second-order polynomial specification* | | | | | | |
| 5-fold Cross-Validation: Pseudo R-squared | 0.038 | - | - | - | - | 0.050 |
| 10-fold Cross-Validation: Pseudo R-squared | 0.053 | - | - | - | - | 0.058 |

*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* Sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data (including indicator variable for census tracts with missing census tract-level data). Outcome variable is recidivism within 2 years conditional on at least 2 years at risk. For the linear functional form, Column (1) "All" controls for charge severity fixed effects, priors count, juvenile history (juvenile felony count, juvenile misdemeanor count, juvenile other count), age, age squared, census tract-level variables (ratio black, ratio Hispanic, ratio income below poverty level, ratio bachelors/associate's degree or higher, labor force participation). For linear functional form, each successive column removes a set of variables. For the nonlinear functional form, Column (1) "All" model allows for charge severity fixed effects, census tract-level variables, and second order polynomial of priors count, juvenile history (juvenile felony count, juvenile misdemeanor count, juvenile other count), age controls (age, age squared). For nonlinear functional form Column (6), the model is all interactions excluding census tract-level variables.

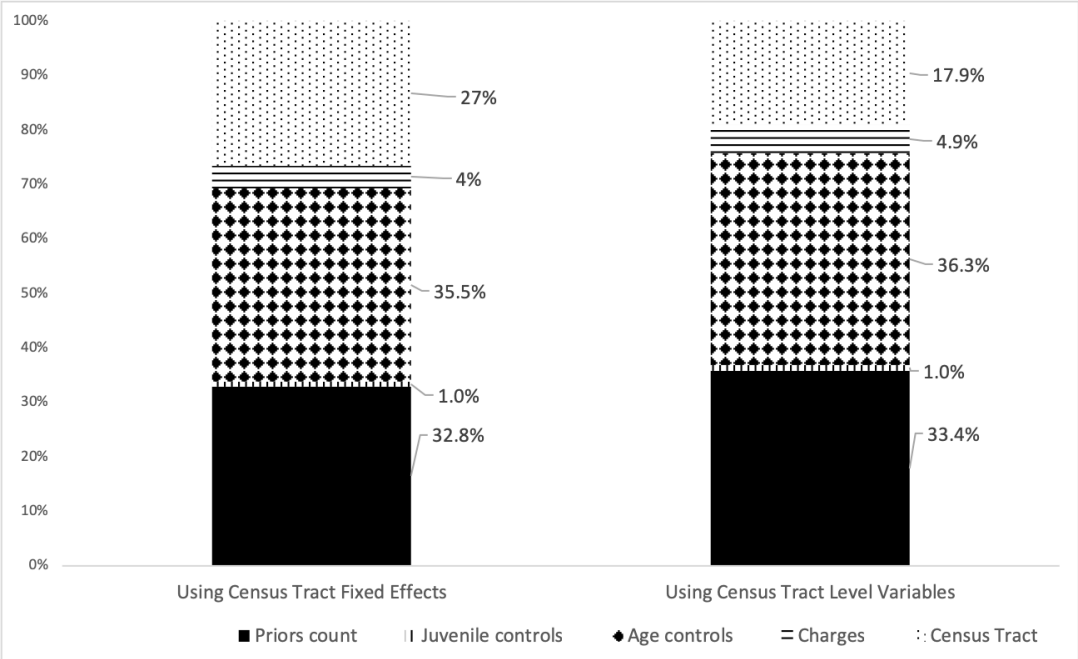
Results are similarly inconclusive after including jail stay variables to account for the possibility that recidivism outcomes are affected by pretrial treatment (either COMPAS score or judges) through the length of pretrial jail stays in Table B4. The count of priors and age controls still do contribute predictive power. However, the predictive power of census tract variables varies from negative to no predictive power across the number of folds used in k-fold cross validation, and between the assumed functional form.

**Table B4:** Prediction of recidivism (2y) using Cox proportional hazard model including jail stay length controls: Cross-Validation

| | (1) All | (2) - Priors | (3) - Juvenile | (4) - Age controls | (5) -Charges | (6) -Census Tract |
|---|---|---|---|---|---|---|
| *Panel A. Linear specification with jail stay lengths* | | | | | | |
| 5-fold Cross-Validation: Pseudo R-squared | 0.052 | 0.029 | 0.056 | 0.029 | 0.054 | 0.054 |
| 10-fold Cross-Validation: Pseudo R-squared | 0.055 | 0.028 | 0.057 | 0.031 | 0.056 | 0.055 |
| *Panel B. Second-order specification with jail stay lengths* | | | | | | |
| 5-fold Cross-Validation: Pseudo R-squared | 0.042 | - | - | - | - | 0.049 |
| 10-fold Cross-Validation: Pseudo R-squared | 0.054 | - | - | - | - | 0.056 |

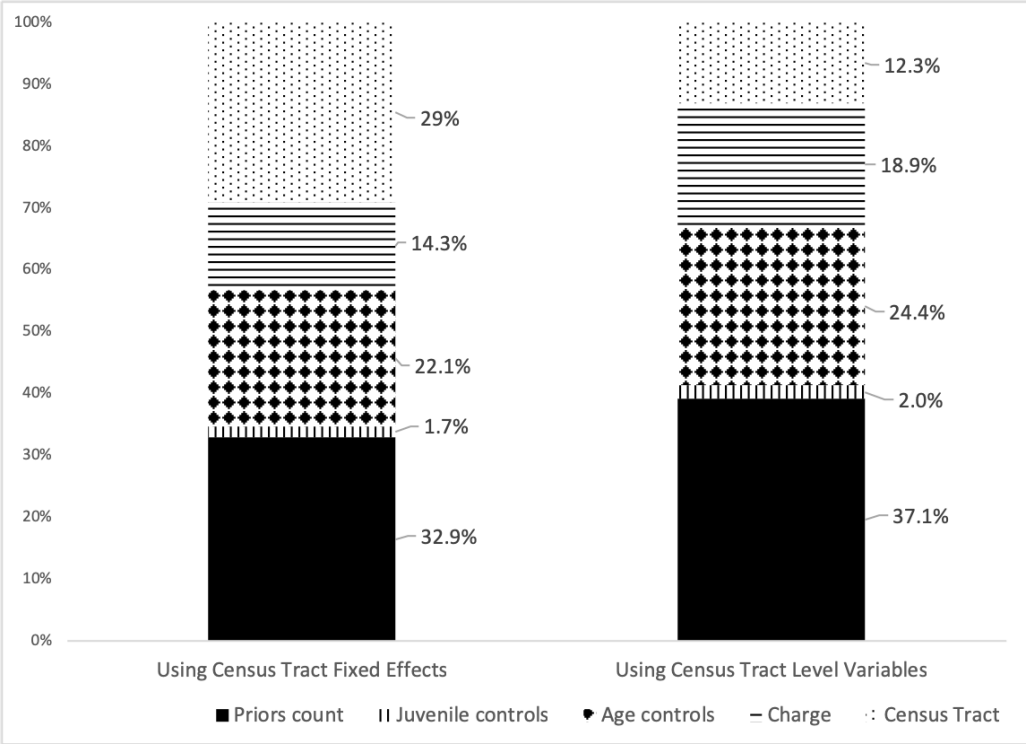
*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* Sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data (including indicator variable for census tracts with missing census tract-level data). Outcome variable is recidivism within 2 years conditional on at least 2 years at risk. For the linear functional form, Column (1) "All" controls for charge severity fixed effects, priors count, juvenile history (juvenile felony count, juvenile misdemeanor count, juvenile other count), age, age squared, census tract-level variables (ratio black, ratio Hispanic, ratio income below poverty level, ratio bachelors/associate's degree or higher, labor force participation), and jail stay length variables as time varying covariates. For linear functional form, each successive column removes a set of variables. For the nonlinear functional form, Column (1) "All" model allows for charge severity fixed effects, census tract-level variables, second order polynomial of priors count, juvenile history (juvenile felony count, juvenile misdemeanor count, juvenile other count), age controls (age, age squared), and jail stay length variables as time varying covariates. For nonlinear functional form Column (6), the model is all interactions excluding census tract-level variables.

**Figure B1:** Decomposing the estimated black-white COMPAS score gap




*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year.
*Notes:* The actual black-white gap in standardized COMPAS scores is 0.687. The estimated black-white gap is 0.626 using the model with census tract fixed effects and 0.600 using the model with census tract-level variables. Sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data.

**Figure B2:** Decomposing the estimated indigent COMPAS score gap




*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* The actual black-white gap in standardized COMPAS scores is 0.514. The estimated black-white gap is 0.515 using the model with census tract fixed effects and 0.515 using the model with census tract-level variables. Sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data.

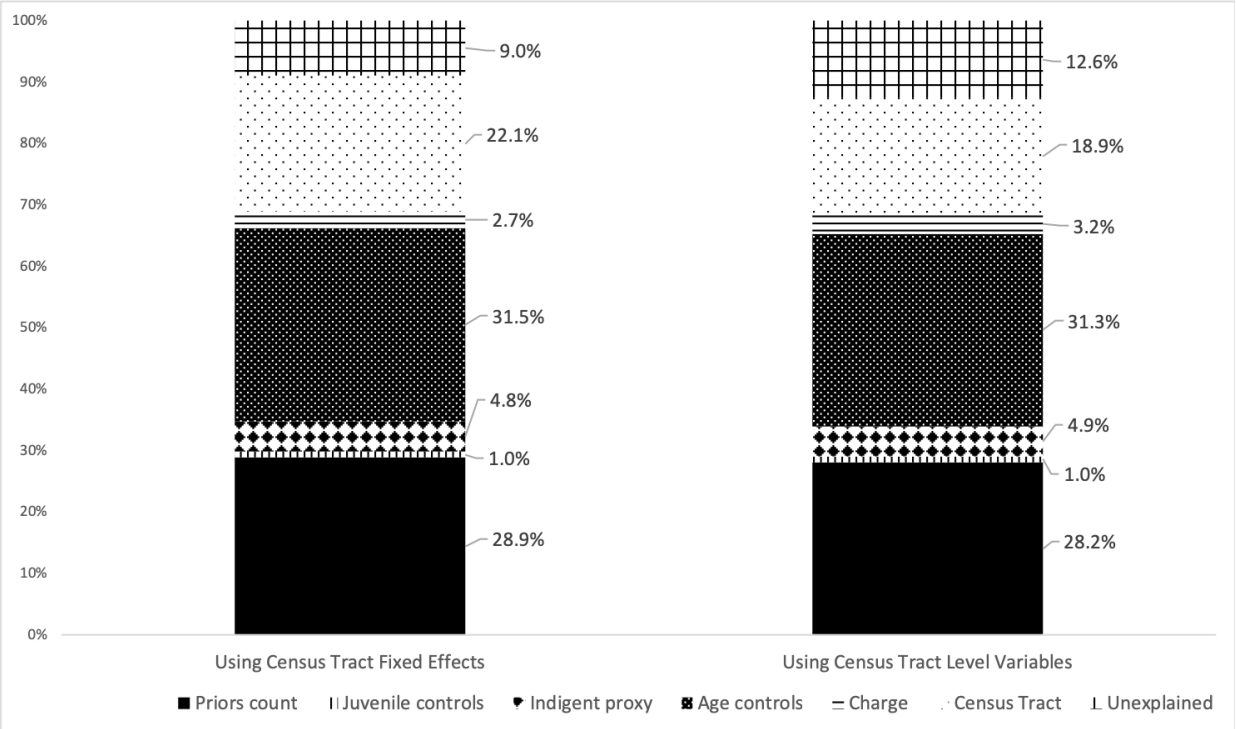
## B.2 Including Indigent Proxy in Risk Score Decomposition by Group

In Figure B3, I find that neighborhood fixed effects explain 22.1% of the actual black-white gap in COMPAS scores. Using neighborhood-level variables in place of fixed effects, I find that neighborhood-level variables explain 18.9% of the actual black-white gap in COMPAS scores. I find similar results when using the estimated black-white gap rather than the actual black-white gap (Figure B4). Using a nonlinear model of COMPAS risk scores to decompose the black-white gap yields similar results: neighborhood fixed effects explain 19.5% of the actual black-white gap in COMPAS risk scores.[46] Neighborhood-level variables explain 17.1% of the actual black-white gap in COMPAS risk scores.[47]

[46] $(0.058 - -0.077)/(0.318 - -0.369)$

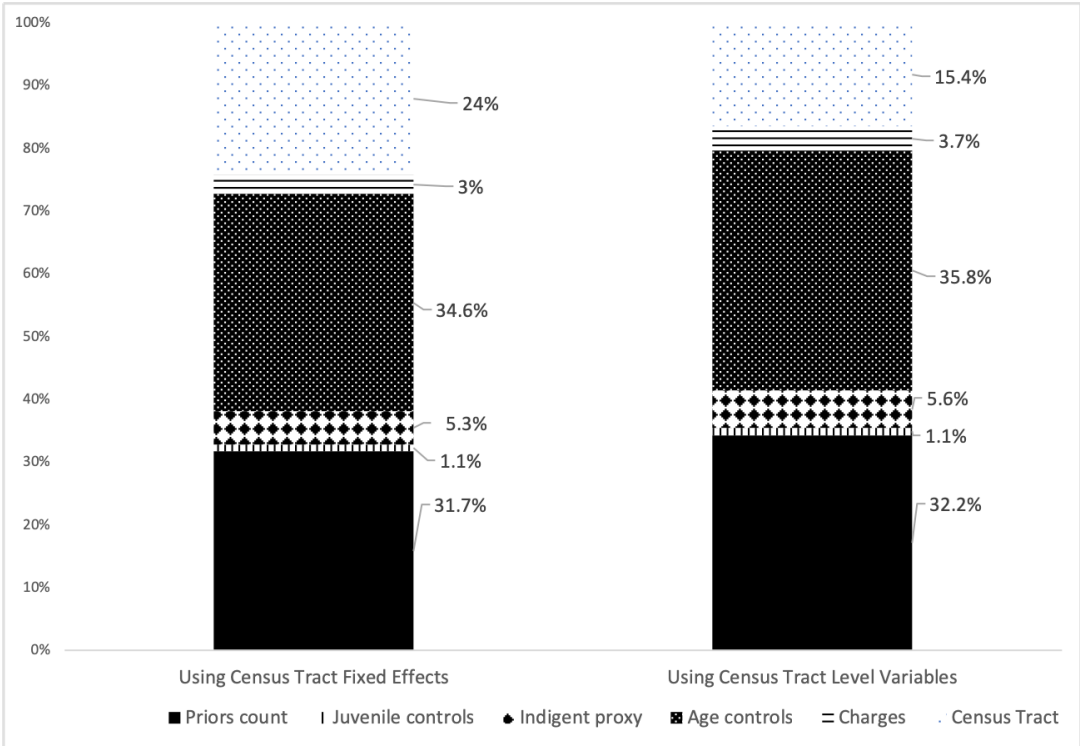
[47] $(0.066 * (0.610 - 0.228) + 0.037 * (0.172 - 0.259) + 0.592 * (0.248 - 0.158) + -0.355 * (0.262 - 0.379) + -0.277(0.342 - 0.347))/(0.318 - -0.369)$

**Figure B3:** Decomposing the actual black-white COMPAS score gap




*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* The actual black-white gap in standardized COMPAS scores is 0.687. The estimated black-white gap is 0.626 using the model with census tract fixed effects and 0.600 using the model with census tract-level variables. Sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data.

**Figure B4:** Decomposing the estimated black-white COMPAS score gap



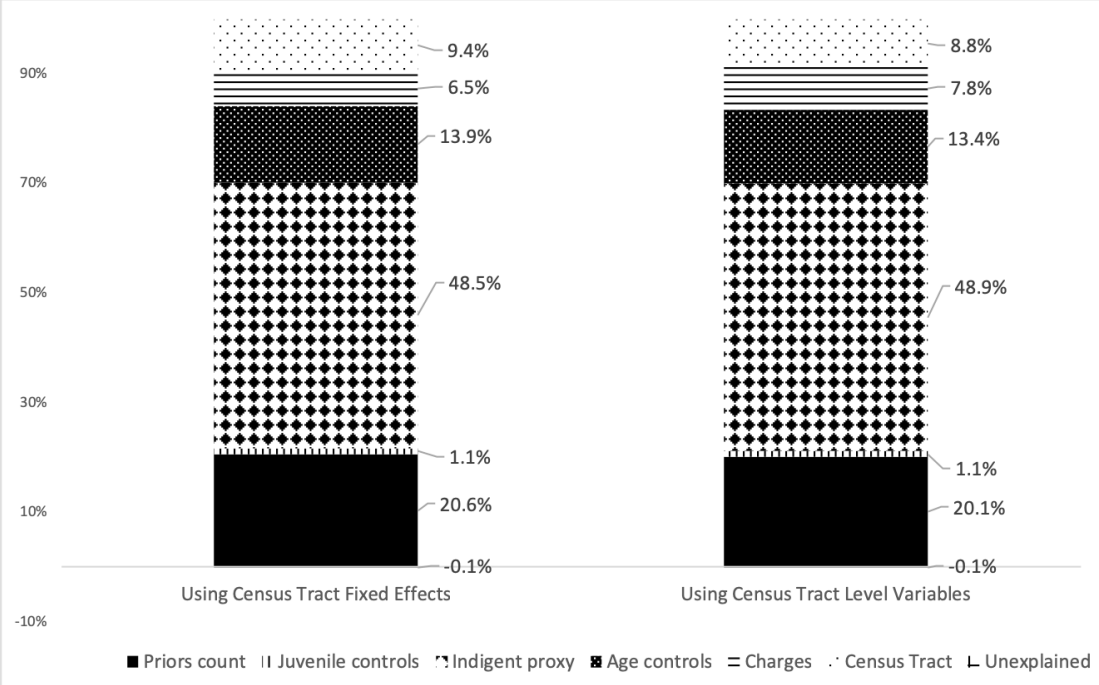

*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* The actual black-white gap in standardized COMPAS scores is 0.687. The estimated black-white gap is 0.626 using the model with census tract fixed effects and 0.600 using the model with census tract-level variables. Sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data.

In Figure B5, I find that neighborhood fixed effects explain 9.4% of the actual indigent-non-indigent gap in COMPAS scores. These neighborhood-level variables explain 8.8% of the actual indigent-non-indigent gap in COMPAS scores. I find similar results using the estimated indigent-non-indigent COMPAS score gap (Figure B5). Using a nonlinear framework, I find that neighborhood fixed effects explain 7.6% of the actual indigent -non-indigent gap in COMPAS scores.[48] I also find that neighborhood-level variables explain 8.2% of the actual indigent -non-indigent gap in COMPAS scores.[49]

[48] $(0.01843 - -0.02060)/(0.26346 - -0.25095)$

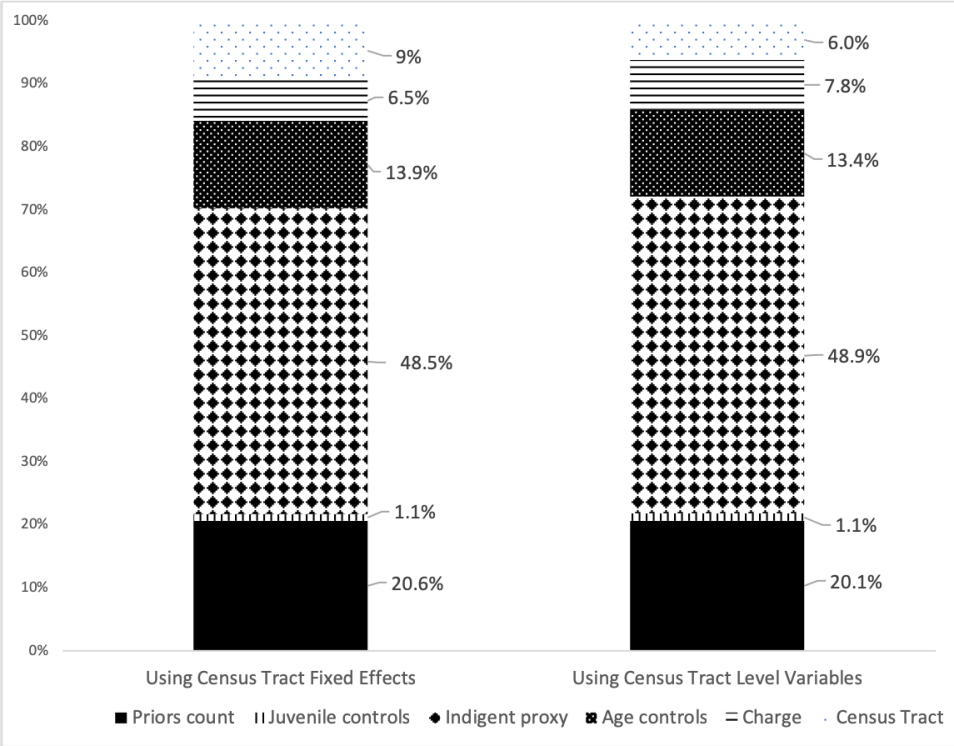
[49] $(0.066 * (0.498 - 0.404) + 0.037 * (0.206 - 0.229) + 0.592 * (0.229 - 0.194) + -0.355 * (0.290 - 0.326) + -0.277(0.339 - 0.352))/(0.263 - -0.251)$

**Figure B5:** Decomposing the actual indigent-non-indigent COMPAS score gap




*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year.
*Notes:* The actual black-white gap in standardized COMPAS scores is 0.514. The estimated black-white gap is 0.515 using the model with census tract fixed effects and 0.515 using the model with census tract-level variables. Sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data.

**Figure B6:** Decomposing the estimated indigent COMPAS score gap




*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year.
*Notes:* The actual black-white gap in standardized COMPAS scores is 0.514. The estimated black-white gap is 0.515 using the model with census tract fixed effects and 0.515 using the model with census tract-level variables. Sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data.

**Table B5:** Out-of-sample Prediction Rates of COMPAS and $ResidScore_{-\xi}$ (COMPAS residualized of census tract fixed effects)

| | Correct | False Positive | False Negative |
|---|---|---|---|
| | (1) | (2) | (3) |
| *Overall* | | | |
| COMPAS | 0.615 | 0.385 | 0.385 |
| $ResidScore_{-\xi}$ | 0.619 | 0.380 | 0.383 |
| | | | |
| *By race* | | | |
| COMPAS: Black | 0.579 | 0.498 | 0.317 |
| $ResidScore_{-\xi}$: Black | 0.593 | 0.465 | 0.330 |
| COMPAS: White | 0.675 | 0.233 | 0.533 |
| $ResidScore_{-\xi}$: White | 0.661 | 0.265 | 0.504 |
| | | | |
| *By economic status* | | | |
| COMPAS: Indigent | 0.611 | 0.478 | 0.263 |
| $ResidScore_{-\xi}$: Indigent | 0.631 | 0.451 | 0.253 |
| COMPAS: Non-indigent | 0.620 | 0.306 | 0.515 |
| $ResidScore_{-\xi}$: Non-indigent | 0.608 | 0.320 | 0.522 |

*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* $ResidScore_{-\xi}$ is the COMPAS score residualized of census tract fixed effects. The weight given to census tract-level variables is estimated from Equation 1 in Section 4.1 using a random 50% of the sample. The other 50% sample is used to predict risk scores and assess out-of sample performance of COMPAS and $ResidScore_{-\xi}$. The total sample is all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data. Correct prediction is the (# of true positives cases + # of true negatives cases)/(# of defendants). False positive rate is the (# of false positive cases)/(# of defendants who do not recidivate). False negative rate is the (# of false negative cases)/(# of defendants who recidivate). Indigent status proxy is 1 if defendant qualifies for a public defender. Recidivism (2 year) outcome is recidivism within 2 years conditional on the defendant having at least 2 years at risk.

**Table B6:** Out-of-sample Prediction Rates of COMPAS and COMPAS - Census Tract Fixed Effects (COMPAS residualized of census tract fixed effects using second-order polynomial specification)

| | Correct | False Positive | False Negative |
|---|---|---|---|
| | (1) | (2) | (3) |
| *Overall* | | | |
| COMPAS | 0.616 | 0.384 | 0.385 |
| $ResidScore_{-\xi}$ | 0.620 | 0.380 | 0.379 |
| | | | |
| *By race* | | | |
| COMPAS: Black | 0.579 | 0.498 | 0.317 |
| $ResidScore_{-\xi}$: Black | 0.594 | 0.463 | 0.330 |
| COMPAS: White | 0.676 | 0.232 | 0.533 |
| $ResidScore_{-\xi}$: White | 0.662 | 0.271 | 0.489 |
| | | | |
| *By race* | | | |
| COMPAS: Indigent | 0.611 | 0.478 | 0.263 |
| $ResidScore_{-\xi}$: Indigent | 0.635 | 0.449 | 0.246 |
| COMPAS: Non-indigent | 0.620 | 0.305 | 0.515 |
| $ResidScore_{-\xi}$: Non-indigent | 0.607 | 0.322 | 0.522 |

*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* $ResidScore_{-\xi}$ is the COMPAS score residualized of census tract fixed effects using a second-order polynomial specification. The weight given to census tract-level variables is estimated from Equation 1 in Section 4.1 using a random 50% of the sample. I use a nonlinear specification that includes up to second order polynomial interactions of all individual characteristics excluding charge-severity fixed effects. The other 50% sample is used to predict risk scores and assess out-of sample performance of COMPAS and $ResidScore_{-\xi}$. The total sample is all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data. Correct prediction is the (# of true positives cases + # of true negatives cases)/(# of defendants). False positive rate is the (# of false positive cases)/(# of defendants who do not recidivate). False negative rate is the (# of false negative cases)/(# of defendants who recidivate). Indigent status proxy is 1 if defendant qualifies for a public defender. Recidivism (2 year) outcome is recidivism within 2 years conditional on the defendant having at least 2 years at risk.

# C Appendix: Multi-stage Framework

This empirical framework assesses how models implicitly balance the trade-off between predictive power and group disparities. In much of my analysis, I compare how a model with "other inputs" performs compared to a model that considers only the criminal record and alleged actions of defendants, by using criminal history, juvenile history and charges-severity fixed effects to predict an "alternative score". The alternative score is trained using criminal history and charge-severity characteristics $\boldsymbol{C_i}$ on half the sample:

$$Recidivism_i = f(\boldsymbol{C_i})$$

I use a third-order polynomial of criminal history (count of priors), juvenile history (juvenile misdemeanor count, felony count and "other" count) and current charge-severity characteristics to predict recidivism (2 years).[50] The alternative score prediction $\widehat{Score_i}$ is predicted for the other half of the sample:

$$\widehat{Score_i} = \hat{f}(\boldsymbol{C_i})$$

I am also considering alternative functional forms of input variables. These control for input variables individually rather than using an alternative score directly in the analysis.

## C.1 Trade-off between predictive power and equity across groups

I propose a two-stage framework to assess the extent of the trade-off between predictive power and group disparities. I am also considering an alternative framework that does not rely on multi-stage regressions. Using the standardized alternative score prediction and standardized COMPAS Recidivism Raw Score, I regress the COMPAS score $COMPAS_i$ on the alternative score prediction and controls $\boldsymbol{X_i}$:

$$COMPAS_i = \alpha\widehat{Score_i} + \boldsymbol{X_i}'\gamma + \varepsilon_i$$

where $\varepsilon_i$ captures factors in COMPAS on top of criminal and juvenile history, charge-severity fixed effects, and age. In the second stage, I model recidivism using each component's

[50]This specification best predicts recidivism; however, I also predict the score using OLS and logit specifications as robustness checks. I also find that results are robust to using a linear specification of criminal history (count of priors), juvenile history (juvenile misdemeanor count, felony count and "other" count) and current charge characteristics. To ensure the results are not driven by outliers, I assign the 1st and 100th percentile of scores the lowest score in the 2nd percentile and the highest score the 99th percentile, respectively (96% winsorization). I also show that results are robust to using the full sample.

contribution to COMPAS estimated in the first stage:

$$Recidivism_i = \beta_1(\hat{\alpha}\widehat{Score_i}) + (\boldsymbol{X_i}'\hat{\boldsymbol{\gamma}})\boldsymbol{\phi} + \beta_2(\hat{\varepsilon}_i) + \eta_i$$

Here, I make the assumption that rearrest conditional on committing a crime does not differ by race. I bootstrap standard errors over the two stages to account for the alternative score and the neighborhood fixed effects being estimates.

The subset of people who are observed for two years after release (the outcome predicted by COMPAS) are not necessarily a random sample of COMPAS-screened defendants because of selection and right censoring in the data. As a robustness check, I use a Cox proportional hazards model to estimate the hazard of rearrest of defendants. This robustness check also addresses the issue that the recidivism outcomes of defendants can be impacted by judges using COMPAS to determine the terms of defendants' pretrial release. The coefficient estimates measure the contribution of each (COMPAS weighted) component to predicting the hazard rate of recidivism. $\beta_1$ captures how the alternative score contributes to recidivism prediction, and $\beta_2$ the contribution of other factors in COMPAS on top of age and the alternative score.

### C.1.1 Decomposing COMPAS scores

The mean difference in COMPAS scores between Blacks and Whites can be decomposed using the estimates of each component and the average group difference in levels of components:

$$\begin{aligned}\overline{\widehat{COMPAS}}_{Black} - \overline{\widehat{COMPAS}}_{White} = \hat{\alpha}(&\overline{\widehat{Score}}_{Black} - \overline{\widehat{Score}}_{White}) \\ &+ \hat{\gamma}_{1,age}(\overline{X}_{Black} - \overline{X}_{White}) \\ &+ \hat{\gamma}_{2,age}(\overline{X}^2_{Black} - \overline{X}^2_{White}) \\ &+ (\overline{\hat{\varepsilon}}_{Black} - \overline{\hat{\varepsilon}}_{White})\end{aligned}$$

where $\overline{\widehat{Score}}_{Race} = \dfrac{\sum_{i=1}^{N} \widehat{Score_i} 1\{i = Race\}}{\sum_{i=1}^{N} 1\{i = Race\}}$, age terms are defined similarly,

$$\text{and } \overline{\hat{\varepsilon}}_{Race} = \frac{\sum_{i=1}^{N} \hat{\varepsilon}_i \cdot 1\{i = Race\}}{\sum_{i=1}^{N} 1\{i = Race\}}.$$

### C.1.2 Decomposing Recidivism

For the probability of recidivism:

$$\overline{\widehat{Recidivism}}_{Black} - \overline{\widehat{Recidivism}}_{White} = \hat{\beta}_1(\hat{\alpha}(\overline{\widehat{Score}}_{Black} - \overline{\widehat{Score}}_{White})) + \hat{\phi}_1(\gamma_{1,age}(\overline{X}_{Black} - \overline{X}_{White})) + \hat{\phi}_2(\gamma_{2,age}(\overline{X}^2_{Black} - \overline{X}^2_{White})) + \hat{\beta}_2(\overline{\hat{\varepsilon}}_{Black} - \overline{\hat{\varepsilon}}_{White})$$

I compare how mean the Black-White level difference in components contributes to the mean Black-White difference in COMPAS scores versus recidivism in equation 7. For short hand, I will refer to "other factors" as input variables in COMPAS other than criminal history, juvenile history, charge-severity fixed effects, and age (represented by $\varepsilon$). The left-hand side of equation 7 represents how "other factors" contribute to the Black -White difference in recidivism. The right-hand side of equation 7 represents how "other factors" contribute to the Black-White difference in COMPAS scores.

$$\frac{\hat{\beta}_2(\overline{\hat{\varepsilon}}_{Black} - \overline{\hat{\varepsilon}}_{White})}{\overline{\widehat{Recidivism}}_{Black} - \overline{\widehat{Recidivism}}_{White}} < \frac{\overline{\hat{\varepsilon}}_{Black} - \overline{\hat{\varepsilon}}_{White}}{\overline{\widehat{COMPAS}}_{Black} - \overline{\widehat{COMPAS}}_{White}} \tag{7}$$

If equation 7 holds, then "other factors" explain disproportionately more of the Black-White COMPAS gap than the Black-White recidivism gap.

### C.1.3 How do neighborhood fixed effects account for the black-white difference in COMPAS scores compared to the black-white difference in recidivism?

To assess defendants' neighborhoods as an example of a potential input variable, I account for unobserved characteristics of defendant neighborhoods (census tracts) using defendant neighborhood fixed effects:

$$COMPAS_i = \alpha\widehat{Score}_i + \xi_j + \boldsymbol{X_i}'\gamma + \nu_i$$
$$Recidivism_i = \beta_1(\hat{\alpha}\widehat{Score}_i) + (\boldsymbol{X_i}'\hat{\boldsymbol{\gamma}})\boldsymbol{\phi} + \beta_2 \cdot \hat{\nu}_i + \beta_3 \cdot \hat{\xi}_j + \eta_i$$

$\beta_3$ is interpreted as how defendant neighborhood characteristics contribute to predicting recidivism. I use the above regressions to decompose the Black-White difference in COMPAS scores and recidivism into differences in levels of the alternative score, age controls, and neighborhood fixed effects. For COMPAS scores:

$$\begin{aligned}\overline{\widehat{COMPAS}}_{Black} - \overline{\widehat{COMPAS}}_{White} = \hat{\alpha}(&\overline{\widehat{Score}}_{Black} - \overline{\widehat{Score}}_{White}) \\ &+ \hat{\gamma}_{1,age}(\overline{X}_{Black} - \overline{X}_{White}) \\ &+ \hat{\gamma}_{2,age}(\overline{X^2}_{Black} - \overline{X^2}_{White}) \\ &+ (\overline{\hat{\xi}}_{Black} - \overline{\hat{\xi}}_{White}) \\ &+ (\overline{\hat{\nu}}_{Black} - \overline{\hat{\nu}}_{White})\end{aligned}$$

where $\overline{\widehat{Score}}_{Race} = \dfrac{\sum_{i=1}^{N} \widehat{Score}_i 1\{i = Race\}}{\sum_{i=1}^{N} 1\{i = Race\}}$, age terms are defined similarly, and $\overline{\hat{\xi}}_{Race} = \dfrac{\sum_{i=1}^{N} \xi_j(i) 1\{i = Race\}}{\sum_{i=1}^{N} 1\{i = Race\}}$ with $\hat{\xi}_j(i)$ defined as defendant $i$'s fixed effect for the census tract $j$ where they live. To understand whether neighborhood fixed effects introduce group disparities, I compare whether neighborhood fixed effects contribute disproportionately more to predicting the Black-White gap in COMPAS than the Black-White gap in recidivism.

## C.2 Results

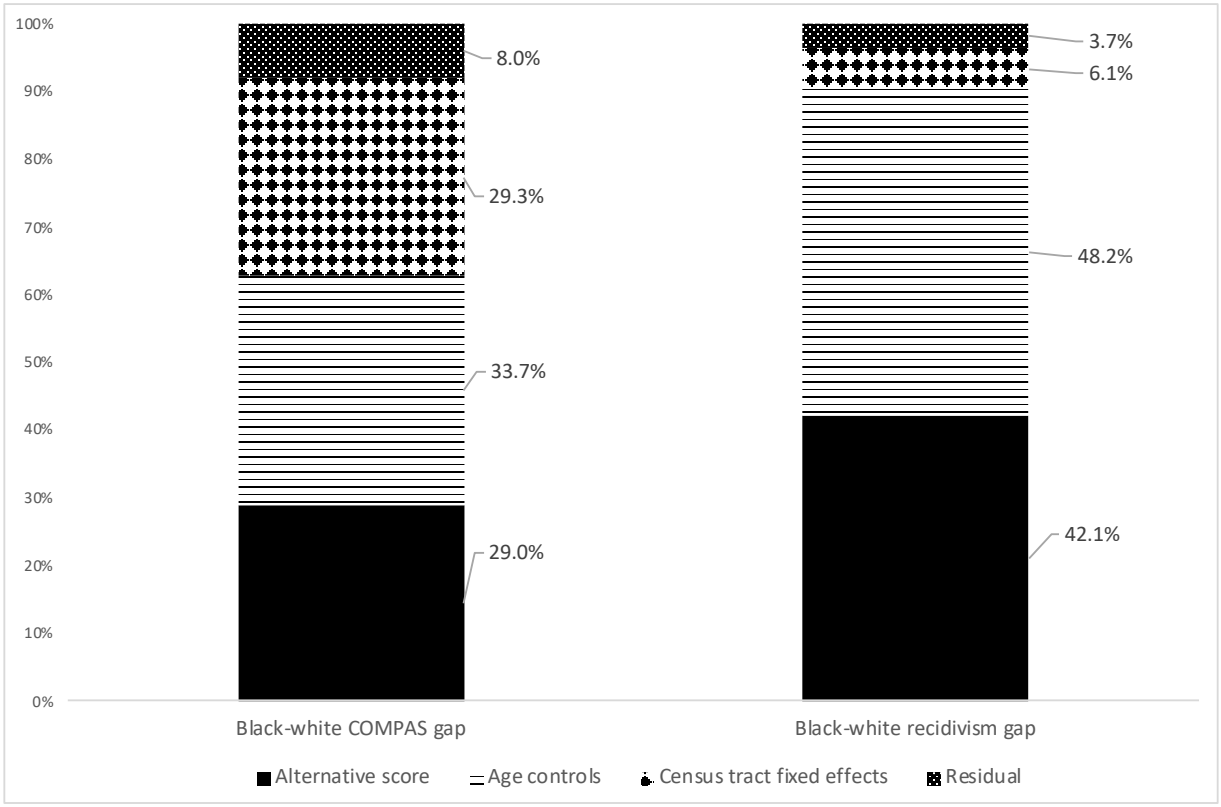
Table C1 reports the results. I decompose the factors other than what enters the alternative score into neighborhood fixed effects and a further residual ($\nu$). Column (1) shows the first stage decomposing COMPAS into the weighted contributions of each component. After including neighborhood fixed effects, the model explains 55.6% of the variation in COMPAS scores. I find that neighborhood fixed effects in column (3) are statistically insignificant, and marginally increase the adjusted R-squared compared to column (2). That is, neighborhood fixed effects have a negligible effect on predicting recidivism.

**Table C1:** Decomposition of COMPAS and prediction of recidivism: Census tract fixed effects

| | COMPAS (1) | Recidivism (2) | Recidivism (3) |
|---|---|---|---|
| $\widehat{\text{Score}}_{\text{sieves,2y}}$ (Z score) | 0.483 | | |
| | (0.018) | | |
| Age | -0.090 | | |
| | (0.011) | | |
| Age squared | 0.001 | | |
| | (0.000) | | |
| $\hat{\alpha} \cdot \widehat{\text{Score}}_{\text{sieves,2y}}$ (Z score) | | 0.222 | 0.220 |
| | | (0.025) | (0.024) |
| $\hat{\gamma_1}\cdot$ Age | | 0.311 | 0.313 |
| | | (0.069) | (0.067) |
| $\hat{\gamma_2}\cdot$ Age Squared | | 0.444 | 0.450 |
| | | (0.239) | (0.234) |
| Residual ($\hat{\nu_i}$) | | 0.069 | 0.069 |
| | | (0.022) | (0.022) |
| Census tract fixed effect ($\hat{\xi_j}$) | | | 0.036 |
| | | | (0.023) |
| Constant | 2.402 | 0.997 | 0.999 |
| | (0.208) | (0.094) | (0.095) |
| Census tract fixed effects | Y | | |
| Adjusted R-squared | 0.555 | 0.089 | 0.090 |
| Observations | 1885 | 1885 | 1885 |

*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* Column (1) shows the decomposition of COMPAS: the coefficients from regressing the standardized COMPAS raw score on the regressors. Columns (2) - (3) show the prediction of recidivism: coefficients from regressing recidivism on the weighted contributions of each regressor, where the weights are from the column (1) specification. Column (2) and column (3) shows the specification without and with the residual estimated from the column (1) specification as a linear regressor, respectively. Sample is a random 50% sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data (including indicator variable for census tracts with missing census tract-level data). Bootstrap standard errors from 200 repetitions in parentheses in column (1) over the first stage COMPAS regression, and columns (2)-(3) over the two stages. Boostrap procedures draw samples of the size of the sample size (1885).

**Figure C1:** Decomposing the black-white COMPAS and recidivism gap into the alternative score, age controls, census tract fixed effects, and a residual




*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* I show how the black-white level difference in the alternative score, age controls (age and age squared), census tract fixed effects, and a residual contribute to explaining the black-white gap in the standardized COMPAS raw score (Z score) and the black-white gap in recidivism. Numbers besides the bar plot show the percentage of total estimated black-white gap that is explained by each component. The total estimated black-white gap in the standardized COMPAS raw score is 0.706, and the total estimated black-white gap in recidivism is 0.111. Sample is a random 50% sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data.

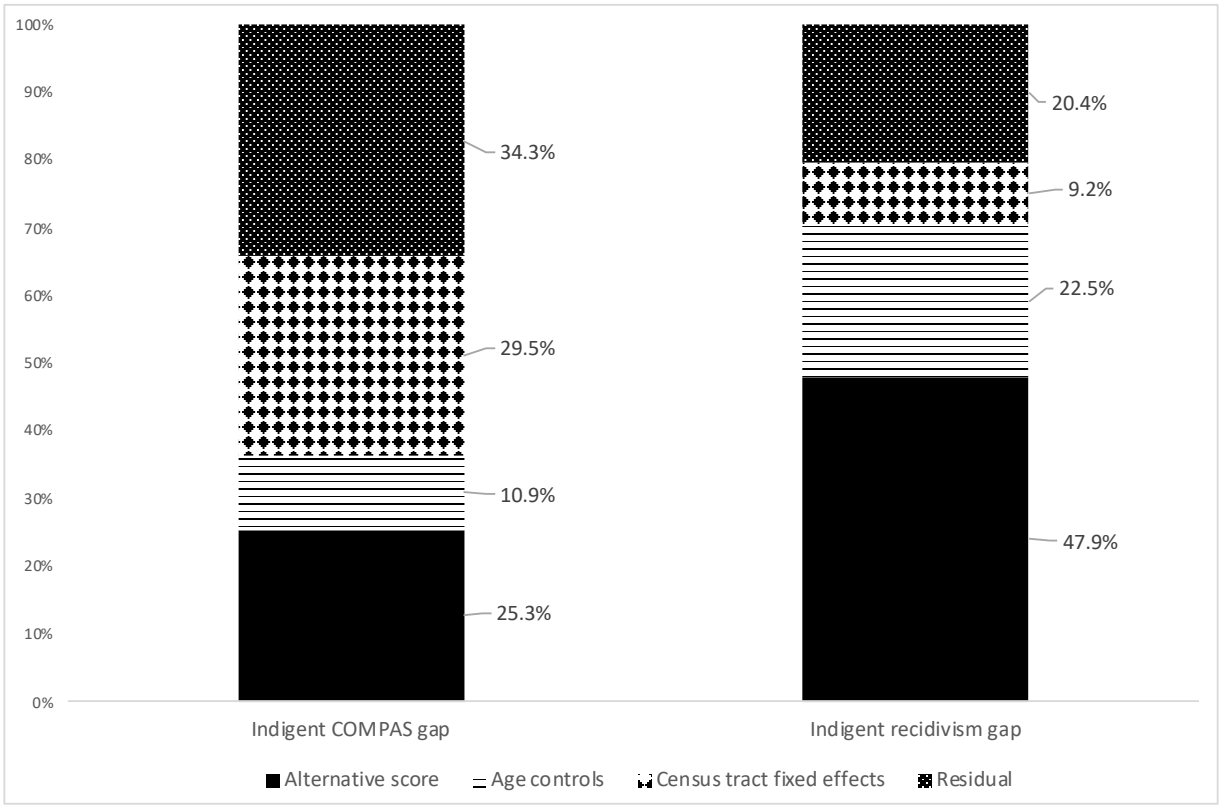
Next, I consider how these components contribute to racial disparities in COMPAS scores. In Figure C1, I find that defendants' neighborhoods explain 29.4% of the average black-white COMPAS gap,[51] but only explain 3.7% of the average black-white gap in recidivism.[52] Additional other factors that are in COMPAS explain 8.0% of the average black-white COMPAS gap,[53], but only 2.2% of the average black-white gap in recidivism[54]. I also find similar results using the Cox Proportional hazards model to decompose the contributions of the

[51]This corresponds to 29.3% of the black-white COMPAS gap that is predicted using the model.
[52]This corresponds to 6.1% of the predicted black-white gap in recidivism.
[53]This corresponds to 8.0% of the predicted black-white COMPAS gap.
[54]3.7% of the predicted black-white gap in recidivism

components to the log black-white relative hazard "risk score".

In Figure C2, I find that defendants' neighborhoods explain 29.5% of the predicted COMPAS gap between indigent defendants who use a public defender and non-indigent defendants who do not. Yet, defendants' neighborhoods only explain 9.2% of the predicted gap in recidivism between indigent and non-indigent defendants.

**Figure C2:** Decomposing the indigent COMPAS and recidivism gap into the alternative score, age controls, census tract fixed effects, and a residual




*Source:* ProPublica COMPAS data set, Broward County Clerk Data, 2010 US Census, 2012-2016 ACS 5-year. *Notes:* I show how the indigent level difference in the alternative score, age controls (age and age squared), and a residual contribute to explaining the indigent gap in the standardized COMPAS raw score (Z score) and the indigent gap in recidivism. Numbers besides the bar plot show the percentage of total estimated indigent gap that is explained by each component. The total estimated indigent gap in the standardized COMPAS raw score is 0.496, and the total estimated indigent gap in recidivism is 0.058. Sample is a random 50% sample of all black, Hispanic, white male pretrial defendants who have addresses that successfully geocode and match with census tract data.