Columbia University

Graduate School of Arts & Sciences

Human Rights Studies Master of Arts Program

Actuarial Injustice: Discrimination in Crime Prediction Software

Julie Ciccolini

Thesis adviser: Bernard Harcourt

Submitted in partial fulfillment of the

requirements for the degree of

Master of Arts

May 2018

Table of Contents

Abstract

The actuarial justice movement has propagated an unprecedented increase in the use of crime prediction software in the criminal justice system. Specifically, two forms of crime prediction software - predictive policing and risk assessment instruments – are now informing high-stakes police and judicial decisions that have direct consequences on individual's civil rights. While advocates claim that the software can alleviate human biases in the system, critics believe it may actually exacerbate them. Due to the conflicting definitions of fairness across legal, technical, and statistical disciplines, there has been no consensus on the software's potential for discrimination.

In order to demonstrate how discrimination can manifest in crime prediction software, I examined a risk assessment instrument designed to predict pretrial felony rearrest for racial discrimination. The instrument is currently used in New York City and to date, has never been independently reviewed. I found that while the instrument demonstrates acceptable predictive validity for all racial subgroups, black defendants receive significantly higher scores on average than white defendants. Although there were small effect sizes, these differences may transcend into discrimination via disparate impact. Most noteworthy, I discovered that only one of the eight predictor variables in the model - whether or not a defendant had any prior arrests - was significantly predictive of future re-arrest. In fact, a redesigned model that predicts rearrest based solely on a defendant's number of prior arrests performed just as well as the original model.

These findings indicate that crime prediction software that utilizes police-generated data to predict police-dependent outcomes is ultimately predicting police activity, not crime. I proffer that this problem is related to the outcome variable at hand and cannot be sufficiently minimized by data manipulation. Therefore, the police and judicial biases that have always plagued America's criminal justice system will be paralleled in crime prediction software.

Introduction

Reforming America's criminal justice system has been at the forefront of the national agenda for many years. With prisons overflowing with disproportionately black and brown bodies, jails doubling as makeshift mental health hospitals, and courtrooms inadequately substituting social services, there is an obvious need for systemic reform. At the same time, technological advancements coupled with the increased availability of big data have ignited a shift towards data-driven decision-making. These simultaneous movements spawned a new branch of criminology and penology called actuarial justice.

Actuarial justice focuses on the use of predictive analytics to assess situational or individual probabilities of criminal behavior.[1] Advocates of actuarial methods proffer them as the antidote to the police and judicial biases inciting mass incarceration. Thus, they claim that institutionalizing prediction software in the criminal justice system will increase public safety, reduce incarceration rates, and neutralize criminal justice decisions.[2] The desperation for reform has resulted in the imprudent legitimization and incorporation of actuarial software into the criminal justice system without proper vetting. As a result, crime prediction software now wields significant power, informing consequential liberty and surveillance decisions across the country.

At the same time, as the use of the software has increased, so has speculation about it. In the past few years, there has been major backlash from academic, legal, and human rights

---

[1]Jonathan Simon and Malcolm Feely, "The New Penology: Notes on the Emerging Strategy of Corrections and Its Implications," *Criminology* 30 (March 7, 2006): 449, doi:10.1111/j.1745-9125.1992.tb01112.x.

[2]Jon Kleinberg et al., "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics* 133, no. 1 (February 1, 2018): 237, https://doi.org/10.1093/qje/qjx032.

scholars questioning the fairness of such practices.[3] From data privacy standards to technological due process concerns, there is a broad spectrum of interdisciplinary implications surfacing with the use of crime prediction software. Whereas legal scholars have focused on equal protection and due process ramifications, technologists have worked extensively on issues of explainability and transparency while statisticians have concentrated their efforts on defining how algorithmic fairness and accuracy can be measured and assured. Still, public discourse has mainly focused on issues of discrimination and fairness. Since crime prediction software's main selling point is the increased neutrality in the system, its potential for discrimination must be explored. Without disregarding many on the other noteworthy critiques of the software, this paper will narrow its scope to investigate solely on how crime prediction software can discriminate.

The Current State of Affairs

Public debate has been unsuccessful in unifying a consensus on crime prediction software's potential for discrimination for many reasons. First, the technology is often

[3]Kelly Hannah-Moffat, "The Uncertainties of Risk Assessment: Partiality, Transparency, and Just Decisions," *Federal Sentencing Reporter* 27, no. 4 (April 2015): 244–47, https://doi.org/10.1525/fsr.2015.27.4.244; Bernard E. Harcourt, *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age* (Chicago: University of Chicago Press, 2007); Bernard E. Harcourt, "Risk as a Proxy for Race: The Dangers of Risk Assessment," *Federal Sentencing Reporter* 27, no. 4 (April 2015): 237–43, https://doi.org/10.1525/fsr.2015.27.4.237; Brian Netter, "Using Group Statistics to Sentence Individual Criminals: An Ethical and Statistical Critique of the Virginia Risk Assessment Program," *Journal of Criminal Law & Criminology; Chicago* 97, no. 3 (Spring 2007): 699–729; Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, First edition (New York: Crown, 2016); James C. Oleson, "Risk in Sentencing: Constitutionally Suspect Variables and Evidence-Based Sentencing," *SMUL Rev.* 64 (2011): 1329; Sonja B. Starr, "Evidence-Based Sentencing and the Scientific Rationalization of Discrimination," *Stanford Law Review* 66, no. 4 (April 2014): 803–72; Sonja B. Starr, "The New Profiling: Why Punishing Based on Poverty and Identity Is Unconstitutional and Wrong," *Federal Sentencing Reporter* 27, no. 4 (April 1, 2015): 229–36, https://doi.org/10.1525/fsr.2015.27.4.229.

proprietary and protected by trade secrets, rendering the source code and methodology used to it inscrutable.[4] This permits an asymmetry of information between the developers and consumers of the software, restricting the public to only offer vague criticism. Communities are forced to blindly trust that it functions as professed without being able to verify any claims or inspect it for issues. This opaqueness permits peremptorily claims of fairness and resolutions of critiques. [5] More progressive versions of crime prediction software have attempted to be more transparent, but that still does not guarantee a level playing field for debate.[6]

The conflicting expertise of scholars in the legal, technical, and statistical disciplines also obstructs the ability for open and coherent discourse. Legal scholars may not be able to articulate how a certain issue is operationalized in actuarial software due to a lack of understanding of predictive modeling. On the contrary, judges, politicians, and police departments may be quick to naively trust software as scientific evidence or, conversely, may refuse to trust it at all.

Principally though, the biggest hindrance to honest evaluation of the costs and benefits of crime prediction software is the multi-disciplinary definitions of fairness and discrimination. Terms like bias and discrimination are operationalized differently between each discipline and

---

[4]Eric L. Loomis v. State of Wisconsin, No. 16–6387 (United States Supreme Court June 26, 2017); Rebecca Wexler, "When a Computer Program Keeps You in Jail," *The New York Times*, June 13, 2017, sec. Opinion, https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html; Rebecca Wexler, "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System," Stanford Law Review 70 (February 21, 2017), https://papers.ssrn.com/abstract=2920883.

[5]Danielle Keats Citron, "Technological Due Process," *Wash. UL Rev.* 85 (2007): 1265, http://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/walq85&section=38.

[6]Dave Gershgorn and Dave Gershgorn, "Software Used to Predict Crime Can Now Be Scoured for Bias," *Quartz*, March 22, 2017, https://qz.com/938635/a-predictive-policing-startup-released-all-its-code-so-it-can-be-scoured-for-bias/.

when decontextualized, are often in conflict with one another. Valid but mischaracterized or misclassified critiques of the software are then easily discounted. Either way, apprehensions about the use of crime prediction software have not succeeded in curtailing their production and expansion into the criminal justice system.

Objective

It is vital now more than ever to properly scrutinize crime prediction software before it becomes fully institutionalized, engrained, and embraced into American society. A demonstration of different standards of fairness across disciplines will hopefully bridge the gap in communication allowing for more informed and legitimate debate. This thesis seeks to demonstrate how crime prediction software functions, how discrimination can become embedded, and if discrimination can be overcome. Discrimination will then be contextually examined and exemplified through an evaluation of a risk assessment instrument currently used in New York City. Specifically, this study aims to answer the question: can crime prediction software be used in the criminal justice system to accurately predict criminal activity without discriminating against protected classes of individuals?

Brief History of Crime Prediction Software

While the configurations of crime prediction software range in sophistication, functionality, and purpose, the concept behind it is rather consistent and simple: try to understand patterns of when, where, and by whom crime has occurred in the past to predict when, where, and by whom it will occur next. This paper will focus on two types of crime prediction software: predictive policing software and risk assessment instruments (RAIs).

Risk Assessment Instruments

First, risk assessment instruments are predictive tools implemented in courtrooms to assist with judicial decision-making. RAIs can be designed for use in all stages of court proceedings to predict a defendant's risk for any outcome of interest.[7] For example, an instrument could be used pre-trial to predict the likelihood that a defendant will be rearrested while out on bail, in sentencing to determine a convicted defendant's security classification, or at a parole hearing to predict the likelihood that a defendant will recidivate.

Different instruments vary in how they designate riskiness, but often a percentile estimate of risk is converted into a score on a Likert-type scale (low-risk, medium-risk, high-risk) assigned to the defendant. While a certain score may knock a defendant out of the running for diversion programs, shorter sentence lengths, or low-security prisons, the score is generally intended to guide the authorities in their decision-making process.

Despite their recent notoriety, risk assessment instruments actually have an extensive history in criminal justice platforms. Scholars often circumscribe the evolution of risk assessment instruments into four generations.[8] The first generation of assessments, which date back to the 1920s, were nothing more than semi-structured professional or clinical judgments. Criticism of the overly subjective nature of these judgments motivated the second generation of instruments in the 1970s, which were empirically tested and consisted of mainly static predictor

[7]For a discussion on the history of risk assessment instruments, see: Harcourt, "Risk as a Proxy for Race"; Paula Maurutto and Kelly Hannah-Moffat, "Assembling Risk and The Restructuring of Penal Control," *The British Journal of Criminology* 46, no. 3 (2006): 440–446.

[8]D. A. Andrews, James Bonta, and J. Stephen Wormith, "The Recent Past and Near Future of Risk and/or Need Assessment," *Crime & Delinquency* 52, no. 1 (January 1, 2006): 8, https://doi.org/10.1177/0011128705281756.

variables.[9] Two decades later, the Risk-Needs-Responsivity framework was born which inspired the third generation of instruments. These instruments focused on identifying dynamic factors (needs) that could be treated with an intervention (response).[10] This brief movement away from straight prediction, while rightfully implying that criminals were capable and worthy of changing their behavior, focused mainly on individualistic and psychological needs factors, such as anti-social personalities, rather than sociological factors.

Over the past decade, the actuarial justice movement has pushed assessment instruments into almost every step of the criminal justice process. These tools have now been tailored to operate in contexts like probation that are less worried about intervention and more about accurately predicting recidivism. Therefore, in the fourth generation, we see a split in the tools aimed at achieving the highest predicting accuracy versus tools aimed at assigning interventions or treatment. Dynamic factors are generally not the strongest predictors of recidivism and often get cut from prediction-oriented tools.

Despite advancements in predictive modeling techniques, most risk assessment instruments are still very simple models produced through basic regression techniques. Since the instruments are usually manually administered in a court setting, they need to be able to be hand scored quickly and consist of easily accessible and obtainable data points. These type of requirements don't lend themselves to the advanced artificial intelligence and black-box

---

[9]Chelsea Barabas et al., "Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment," *ArXiv:1712.08238 [Cs, Stat]*, December 21, 2017, http://arxiv.org/abs/1712.08238.

[10]D.A. Andrews, James Bonta, and R.D. Hoge, "Classification for Effective Rehabilitation: Rediscovering Psychology," *Criminal Justice and Behavior* 17, no. 1 (March 1, 1990): 20, https://doi.org/10.1177/0093854890017001004.

modeling typically discussed with predictive modeling. Furthermore, risk assessments are developed in both the private and public sector and are often less profit-motivated than other types of crime prediction software.

Predictive Policing

Unlike risk assessment instruments, predictive policing software is a relatively new innovation spawning out of advancements in machine learning and big data analysis. Predictive policing software refers to any software used by police departments that attempts to predict where crime is likely to occur, who is most likely to commit it, and/or who is most likely to be victimized by it. Police departments use predictive policing software to help determine resource allocations and identify potential suspects, perpetrators and/or victims of crime. [11]

Private sector companies typically develop predictive policing software with the goal of turning a profit. Many big data companies, including IBM, Microsoft, Hitachi, and Palantir, are expanding their markets into predictive policing.[12] Due to the commercialization of predictive policing, the software is often described with unwarranted and superfluous rhetoric. For example, the predictive policing company Azavea claims that their software is a "crystal

---

[11]For an overview of predictive policing, see: Azavea, "HunchLab: Under the Hood," 2015, https://cdn.azavea.com/pdfs/hunchlab/HunchLab-Under-the-Hood.Pdf; Danah Boyd, Sarah Brayne, and Alex Rosenblat, "Predictive Policing" (A New Era of Policing and Justice, Washington, D.C., 2015), http://www.datacivilrights.org/pubs/2015-1027/Predictive_Policing.pdf; "How PredPol Works | Predictive Policing," *PredPol* (blog), accessed December 8, 2017, http://www.predpol.com/how-predictive-policing-works/.

[12]"Predictive Policing Software Is More Accurate at Predicting Policing Than Predicting Crime," American Civil Liberties Union, accessed April 17, 2018, https://www.aclu.org/blog/criminal-law-reform/reforming-police-practices/predictive-policing-software-more-accurate.

ball" that "anticipates where crime is likely to emerge." However, the actual analyses are far from magic. [13] In reality, the geographic-based software analyzes previous police department data, such as 911 calls and arrests, in order to calculate patterns of occurrence based on time, date, location, and type of offense. Some companies have tried to differentiate their software from typical heat maps by incorporating more sophisticated or trendy data techniques into their models, such as "risk-terrain modeling" or "resource allocation."[14] Nevertheless, the basic concept behind the software remains consistent throughout brands.

Other forms of predictive policing software focus less on the location of future crime and more on identifying the future perpetrators. Gang databases, strategic subject lists, and social network analyses are individual-level forms of predictive policing aimed at identifying potentially dangerous individuals.[15] While these types of databases have surfaced in California, Chicago, and New York City, less is known about their development. Departments have reported that individuals are analyzed based on factors like known gang affiliations, criminal history, social media posts, and demographic factors.[16] However, due to the immense opacity surrounding this type of technology, not much else is publicly known about it.

A Primer on Predictive Modeling

[13]Azavea, "HunchLab: Under the Hood."

[14]Risk terrain modeling analyzes risk based on the presence and density of certain geographic features, like subway stops and bars, linked with specific crimes. (Azavea, 2015). Police resources are allocated based on "the predicted societal impact of crime at different locations" while "weighting the importance of preventing different crimes appropriately (Azavea, 2015)."

[15]Boyd, Brayne, and Rosenblat, "Predictive Policing."

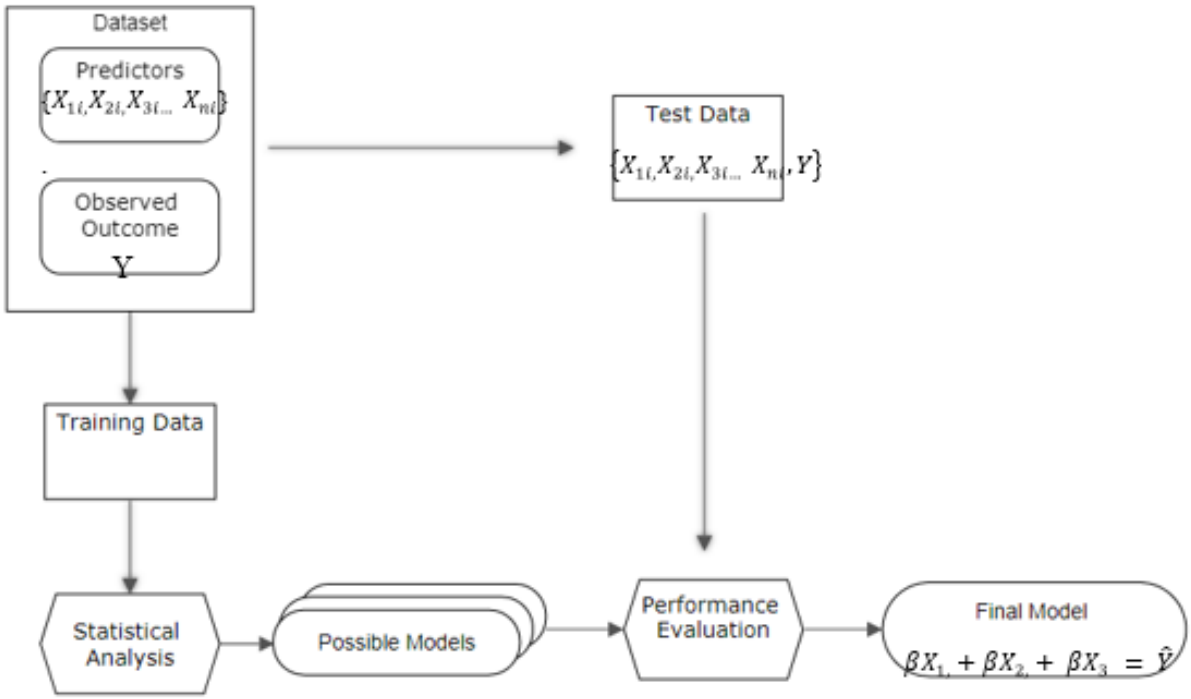
[16]Jack Smith, "Chicago's New Policing Strategy Is Hurting the People It's Supposed to Be Helping," *MIC*, August 17, 2016, https://mic.com/articles/151782/chicago-s-experimental-policing-tool-is-hurting-the-people-it-s-supposed-to-be-helping.

In order to understand how discrimination can manifest in crime prediction software, a basic understanding of predictive modeling is essential. Figure 1 provides an intentionally generalized and simplified example of how predictive models are built. First, a dataset consisting of the observed outcomes for the dependent variable alongside potential independent variables is partitioned into "training data" and "test data". Next, the modeler performs some kind of statistical analysis, which can range from basic linear regression analysis to unsupervised machine learning, on the training data. Regardless of the technique, the analysis seeks to identify the correlations between the independent variables and the outcome of interest in the training data.

Once the strength and direction of the associations between the dependent and independent variables are established, various models can be constructed to summarize the relationships. Models are then tested using the other portion of the dataset, the "test data", to ensure that they perform well. Miscellaneous performance metrics and requirements are often compared when identifying the best possible model for the situation at hand. For example, the need for simplicity and interpretability may be more heavily prioritized when constructing a model for use in a courtroom versus a police department. At the discretion of the modeler, a final version of the model is selected.[17]

[17]Obviously, this process is often not as simplistic and sequential as this example would suggest, but this is nonetheless analogous to the general procedure. The process can be much more sophisticated and the output models can be extremely complicated. As is the case with newer machine learning models, sometimes the model cannot even be explained through an equation but rather a computer identifies the patterns in the data.

Figure 1. Predictive Modeling Process



A couple points should be made here about the nature of predictive modeling and its inherent limitations:

- Predictive models will always be constrained to making associations based on the data fed into them. In the age of big data, these datasets may be far vaster and amorphous than previously thought possible, but they still must be tracked, stored, and input in the model.
- Predictive models are based on correlations, not causation. Predictive modeling is a fundamentally shallow analysis of the relationships between variables. While models seek to understand the weight of the predictor variable's effects on the

dependent variable, they do not seek, nor do they care, to understand why this relationship exists or how it occurs.

- ☐ Predictive models function through a form of quantitative taxonomy, reducing individuals to a product of a few identities and characteristics . Predictions for one individual are made by analyzing the outcomes of all the individuals in the training data with similar identities.
- ☐ Predictive models are fixated on optimization. They require human intervention to constrain them for the contextualized goal of the software, if different from solely optimization.
- ☐ Predictive models are de-contextualized. They cannot detect systematically misrepresentative data if those biases exist in both the training and test data.

Manifestations of Discrimination in Crime Prediction Software

The entire purpose of predictive analytics is to optimize discrimination. Therefore, questioning its potential for discrimination would appear rather oxymoronic. However, human rights law prohibits discrimination based on certain protected-characteristics, like age, race, gender, and socio-economic status. The idea that, "all are equal before the law and are entitled without any discrimination to equal protection of the law, " is defined in Article 7 of The Universal Declaration of Human Rights and mirrored in almost all other human rights doctrines.[18] The prohibition of discrimination has been solidified repeatedly in other

[18]UN General Assembly, *Universal Declaration of Human Rights*, 10 December 1948, 217 A (III), available at: http://www.un.org/en/documents/udhr/ [accessed 11 November 2016]; UN General Assembly, *International Covenant on Civil and Political Rights*, 16 December 1966, United Nations, Treaty Series, vol. 999, p. 171, available at:

conventions as well with respect to specific characteristics, like race, gender, age, and disability.[19] Thus, the question really of interest is – how can crime prediction software discriminate against prohibited characteristics?

International courts have determined that discrimination can occur either directly - when an individual is treated differently based on a protected class characteristic - or indirectly - through policies or laws that result in a disparate impact.[20] Since the human rights regime has arguably the most universal application, this is the definition of discrimination that will be used throughout the paper. It should be noted that just because crime prediction software may discriminate according to this definition, this does not mean that it necessarily violates human rights law. International courts have determined that discrimination by classification is permissible if the distinctions are justified, reasonable, and imposed for an objective and legitimate purpose.[21] Specifically, the Inter-American Court found that equal protection is not violated when "classifications are based on substantial factual differences and there exists a

---

http://www2.ohchr.org/english/law/ccpr.html [accessed 11 November 2016]; UN General Assembly, *International Convention on the Elimination of All Forms of Racial Discrimination*, 21 December 1965, United Nations, Treaty Series, vol. 660, p. 195, available at: http://www2.ohchr.org/english/law/cerd.htm [accessed 11 November 2016]; U.S. Const. amend. XIX. References to Equal Protection can be found in Article 26 of the International Covenant on Civil and Political Rights, Article 24 of the American Convention on Human Rights, and the 14th Amendment of the United States Constitution.

[19]International Convention on the Elimination of All Forms of Racial Discrimination; Convention on the Elimination of All Forms of Discrimination Against Women; Convention on the Rights of Persons with Disabilities; and the Convention on the Rights of the Child

[2020]UN Committee on Economic, Social and Cultural Rights (CESCR), General comment No. 20: Non-discrimination in economic, social and cultural rights (art. 2, para. 2, of the International Covenant on Economic, Social and Cultural Rights), 2 July 2009, E/C.12/GC/20, available at: http://www.refworld.org/docid/4a60961f2.html [accessed 13 October 2017]

[21]United Nations Office of the High Commissioner for Human Rights and International Bar Association. Human Rights In The Administration Of Justice: A Manual On Human Rights For Judges, Prosecutors And Lawyers. United Nations Publications, 2003, 652.

reasonable relationship of proportionality between these differences and the aims of the legal rule under review."[22] While determining if different types of discrimination are legally permissible is outside the scope of this paper, it will emphasize how crime prediction software can discriminate against any protected or vulnerable groups. Clarification on the operalization of discrimination in crime prediction software will inform more legitimate debate on its permissibility.

With that in mind, crime prediction software's potential for discrimination can be sectioned into three categories: discrimination by design, discrimination by training, and discrimination by interaction.

Discrimination by Design

Crime prediction software may discriminate by design, intentionally or unintentionally, if it explicitly makes decision based on a protected characteristic or fails to account for the correlation of that characteristic in the model.

Predictor Variables

The most obvious way in which a model can discriminate is through direct consideration of a restricted attribute. A model can be designed to take into account an individual's race, gender, age, socio-economic status, or any other sensitive attribute when making predictions by explicitly including them as predictor variables. The only thing restricting a model from including an attribute as a predictor variable is 1) whether or not there is data available on that attribute and 2) whether or not the designer chooses to include it. If the attribute is a strong

[22]*Advisory Opinion on Proposed Amendments to the Naturalization Provision of the Constitution of Costa Rica*, OC-4/84, Inter-American Court of Human Rights (IACrtHR), 19 January 1984.

predictor, meaning it is highly correlated with the outcome, then the only reason to exclude it from the model would be ethical or moral concerns.

Race, for example, was often included as a predictor variable in early risk assessment instruments until around the 1970s when it was deemed unethical.[23] The fact that a black individual would be automatically ranked as riskier than a similarly situated white individual, solely because data shows that black individuals historically were "riskier", falls within the definition of explicit racial discrimination. [24] Whether or not this discrimination is permissible or reasonable, though, is subjective. Many argue that it is permissible for some characteristics but not others. Gender and age, for example, are still included as predictors in many risk assessment instruments currently being used in American courtrooms.[25]

Fairness through Unawareness

In an effort to appease critics in regards to controversial predictor variables, many developers and consumers of crime prediction software have fallen victim to the "fairness through unawareness" fallacy. By simply excluding the contentious factor from the training data and subsequent model, developers erroneously conclude that it alleviates the potential for discrimination. On the contrary, exclusion can actually be more problematic than inclusion.

There is still potential for redundant encodings of the attribute through mediating or

---

[23]Harcourt, "Risk as a Proxy for Race," 239.

[24]Harcourt, "Risk as a Proxy for Race,"23; Oleson, "Risk in Sentencing," 1329; Starr, "The Scientific Rationalization of Discrimination,"803; Starr, "The New Profiling,"229;

[25]Oleson, "Risk in Sentencing," 1329.

proxy variables.[26] Consider two individuals asked to predict the likelihood of weight gain based on a food's nutrition facts label; One individual sees the full label and the other sees everything but calorie count. Their predictions are not likely to differ much since other variables on the label – fat, carbohydrate, and sugar content - still account for the relationship between weight gain and calories. So long as other variables associated with calories are included in the prediction, then foods are still going to be discriminated against based on calories.

With crime prediction software, this is not simply a problem of multicollinearity or omitted variable bias. If a sensitive attribute is a good predictor of the outcome variable, then they are somehow correlated. One cannot exclude all variables associated with that attribute from the model without decreasing its accuracy and validity. [27] Accordingly, an accurate prediction of the outcome variable will also be correlated with the attribute. Thus, enforcing decisions based on the prediction will have an effect on anything correlated with

[26] Terms like proxy and mediator have often been conflated in risk assessment literature. To be clear, a proxy variable is a variable that functions as a placeholder for the relationship between a predictor and an outcome, but would not otherwise have a direct relationship to the outcome itself (receiving social security may have a negative correlation with crime because it functions as a proxy for age). The predictor variable has a stronger relationship with the outcome and the proxy functions as a stand in. On the other hand, a mediator variable explains the relationship between the predictor and the outcome variable (police discretion may explain the relationship between race and arrest) whereas a moderator variable changes the direction or strength of the relationship between the predictor and the outcome variable (police density moderates relationship between crime and arrest). There is potential for a sensitive attribute to be influential in a model in all of these ways.

[27]Sorelle A. Friedler et al., "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning,"ArXiv:1802.04422 [Cs, Stat], February 12, 2018, 16, http://arxiv.org/abs/1802.04422.

the outcome, regardless if it is included in the model or not. [28]

Fairness through Awareness

Since many critics of crime prediction software have started to vocalize the trade-off between accuracy and non-discrimination, efforts to achieve "fairness through awareness" have become more common. Despite the knee-jerk reaction to object to the inclusion of controversial predictor variables, it is actually the best way to control for the discrimination. [29] Various methodologies to control for discrimination of a sensitive attribute have been proposed.[30] Each have their own merits and effectiveness but they are also subject to different conceptions of fairness.

For example, statistical parity is one conception of fairness that looks at the classification of individuals based on a sensitive attribute, irrespective of the outcomes. Sometimes referred to as demographic parity, this concept requires that the proportion of individuals classified as high risk is the same for each subgroup, or that the average of the scores over all the strata of each group be the same. [31]

---

[28] Moritz Hardt, Eric Price, and Nathan Srebro, "Equality of Opportunity in Supervised Learning," *ArXiv:1610.02413 [Cs]*, October 7, 2016, 1, http://arxiv.org/abs/1610.02413; Dino Pedreshi, Salvatore Ruggieri, and Franco Turini, "Discrimination-Aware Data Mining," in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM, 2008), 561.

[29] Cynthia Dwork et al., "Fairness Through Awareness," *ArXiv:1104.3913 [Cs]*, April 19, 2011, 11, http://arxiv.org/abs/1104.3913.

[30] Dwork et al., "Fairness Through Awareness"; Friedler et al., "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning"; Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning."

[31] Dwork et al., "Fairness Through Awareness."

While this may equate to a rough idea of group fairness, it fails on an individual level and can have significant cost to accuracy. [32] If certain divisions of a sensitive attribute have different average levels of risk, then equalizing the odds of a prediction for each division will synthetically bolster the risk score for some while lessening the risk score for others.[33] Since this fabricates artificial changes in the score but not the outcome, there will be high error rates in the predictions.

For example, controlling for an attribute like gender in pretrial crime prediction will undeservedly increase female risk scores while decreasing male scores. If no changes in recidivism occur, then more females will be unnecessarily detained while more males will be incorrectly released. This type of predictive affirmative action is better suited for a field where equalized opportunity ensures equal access to benefits rather than punishment. Algorithms that revoke constitutional protections are not the place for a reparations framework since they do not reverse historical errors but rather create new ones.

Statistical parity is not the only regulation that has been proposed to counter discrimination. However, the number of predictor variables that need to be protected can exceed the abilities of the regulation. Very few fairness-aware algorithms can formally handle multiple sensitive attributes at the same time and there is a limit to the number that a

---

[32] Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Calif. L. Rev. 671 (2016): 721; Toon Calders and Sicco Verwer, "Three Naive Bayes Approaches for Discrimination-Free Classification," *Data Min. Knowl. Discov.* 21, no. 2 (September 2010): 290, https://doi.org/10.1007/s10618-010-0190-x.; Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning."

[33]Corbett-Davies

model can control for before losing its functionality and purpose.[34] This is problematic since, even though most critiques have focused on constitutionally protected-characteristics, ethical arguments could be made against the inclusion of almost any predictor variable. For instance, some take issue with risk assessment instruments that consider the delinquency of an individual's friends or the adequacy of their parental supervision in their models, faulting the individual for the confounding sociological stressors in their life.[35] Furthermore, group taxonomies can arise through a combination of predictor variables, making them less apparent but not any less important. For example, including factors on mental health characteristics and peer relations can isolate certain personality types and subject individuals who stray from the norm to higher scrutiny. Point being, unintended group classifications can arise through predictive modeling that, if more visible, would be controversial. Rather than debate the morality of these codifications, these examples are intended to showcase the extent to which typecasts can become amorphous and abstractly discriminated against by the model.

Interpretability & Transparency

Complicating matters further, in cases of black-box machine learning, decisions about which variables to include in a model may be completely relegated to automation. Advancements in machine learning and artificial intelligence have augmented the computational capacity to

[34] Friedler et al., "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning."; Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," ArXiv:1609.05807 [Cs, Stat], September 19, 2016, http://arxiv.org/abs/1609.05807.

[35] R.D. Hoge and D.A. Andrews, "YLS/CMI™ - Youth Level of Service/Case Management Inventory," accessed April 17, 2018, https://www.mhs.com/MHS-Publicsafety?prodname=yls-cmi.

analyze big data and identify more complex relationships between variables. Even though these methods create better models, the gains in accuracy are diminished by the loss in explainability and comprehensibility. Invaluable research is occurring simultaneously to create mechanisms to audit black-box models for discrimination, but these methods may be inefficient to identify all variations of algorithmic discrimination within the criminal justice context.[36]

Detecting discrimination in human-derived models can still be undermined if the developers choose not to be transparent. Transparency decisions about whether or not to report on the data used to construct the software, the model's weights and factors, and the statistical analysis all play a role in masking discrimination. Not surprisingly, most crime prediction software is only evaluated by the same people who developed it, subject to their same oversights, biases, and profit-motive.[37] For the developers of proprietary software, transparency threatens the capitalist bottom-line and exposes them to critiques that can otherwise be avoided. While not implying that developers are deliberately trying to mask discrimination, opaque software allows them to hide behind an impenetrable veil of neutrality that can conceal both calculated and subconscious discrimination. Surely techniques exist to test models for discrimination that do not require knowledge of the software's inner-workings. However, creative manipulation, like instituting decision thresholds, can still camouflage discrimination.

Decision Thresholds

---

[36]Niels Bantilan, "Themis-Ml: A Fairness-Aware Machine Learning Interface for End-To-End Discrimination Discovery and Mitigation," Journal of Technology in Human Services 36, no. 1 (January 2, 2018): 15, https://doi.org/10.1080/15228835.2017.1416512.

[37] Sarah Desmarais and Jay Singh, "Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States," 2013, csgjusticecenter.org/wp-content/uploads/2014/07/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf.

Another way discrimination can manifest in crime predictions software is in the conversion from risk estimates to category labels. Crime prediction software is often designed to transform the model's raw output- probability estimates of risk- into categorical thresholds. There are many rationales for this type of transformation: the results are easily digestible, the risk categories are pre-contextualized, and the treatments options are pre-defined.

Figure 2. Decision Thresholds

However, these categorical labels are not only stigmatizing but can also be easily manipulated. For example, a developer can decide to convert the model's predictions into three risk categories - low risk, medium risk, and high risk – which contextually may designate specific treatment options – release, set bail, detain. A model could be tactfully designed to appear well-calibrated while never putting a certain strata of individual above the medium risk threshold.[38] In the case of opaque software, this algorithmic redlining may be undetectable since

[38]Sam Corbett-Davies et al., "Algorithmic Decision Making and the Cost of Fairness"(ACM Press, 2017), 804, https://doi.org/10.1145/3097983.3098095.; Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning," 3315.

the model would still appear well calibrated if all the strata have an equal likelihood of the outcome.[39]

In summation, significant decisions are made when designing crime prediction software that can make it prone to explicitly discriminate against protected classes according to human rights standards. While the arithmetic involved in predictive modeling can deceive us into believing sensitive attributes are not being considered, exclusion from a model does not equate to protection. On the contrary, the omission of protected characteristics coupled with fancy mathematical techniques and trade secret citations can adroitly conceal direct discrimination. Assuming a model is made that does not explicitly discriminate against protected groups, it can still be trained to indirectly discriminate.

Discrimination by Training

Crime prediction software can have a disparate impact if it systematically makes poorer or riskier predictions for certain groups. This can happen as a result of the training data used to build the model, which, if biased in some way, can result in invalidity. Since predictive models have no contextual understanding of the data they are fed, they are just as susceptible to the historical prejudices that influenced the dataset. This is especially consequential in the criminal justice system where the decisions at every juncture have been subjected to human biases.

Subgroup Validity

The extent to which a model accurately predicts what it is supposed to predict is called predictive validity. Predictive validity is checked based on two performance metrics: calibration and discrimination. Calibration refers to how well a tool's predictions of risk agree with actual

[39]Corbett-Davies et al., "Algorithmic Decision Making and the Cost of Fairness," 804.

observed risk. In the case of group fairness, this means that a score reflects the same likelihood of failure irrespective of the individual's group membership. Discrimination, on the other hand, describes how well an instrument is able to separate those who went on to engage in the outcome from those who did not. [40] Tests of discrimination and calibration are equally important when assessing predictive validity.[41]

While overall predictive validity ensures that a model is not making completely wild and random predictions, it does not ensure that a model makes equally valid predictions across all strata of a sensitive attribute. Certifying that a model has equal predictive validity for all subgroups (subgroup validity) is essential to preventing a model from having a discriminatory impact.

Problematically, crime prediction software is usually advertised based on its overall predictive validity but it is uncommon for software to check or report on subgroup validity. According to a meta-analysis conducted in 2013, most risk assessment instruments currently used in courtrooms have not actually tested their predictive validity among different subgroups.[42] Likewise, predictive policing software is typically marketed directly to police departments who, judging by their history, are likely not concerned with subgroup validity as much as about overall

---

[40]Jay P. Singh, "Predictive Validity Performance Indicators in Violence Risk Assessment: A Methodological Primer: Performance Indicator Primer," Behavioral Sciences & the Law 31, no. 1 (January 2013): 8, https://doi.org/10.1002/bsl.2052.

[41]Corbett-Davies et al., "Algorithmic Decision Making and the Cost of Fairness," 8; To understand the importance, imagine a model is designed to predict a crime that occurs in 1% of communities. An instrument that predicts that crime will occur in 0% of communities will still be 99% accurate and thus, well-calibrated. On the other hand, the instrument demonstrated no discriminatory ability to separate the communities where the crime occurred from those where it did not.

[42] Desmarais and Singh, "Risk Assessment Instruments Validated in the U.S.," 19.

accuracy. If non-discrimination is not a conscious concern for the software, then subgroup validity is likely to go unchecked; This is because extremely poor predictions for members of a small minority group would not cause a substantial decline in the overall predictive accuracy of the model, despite having serious consequences for those group members. Not surprisingly, a substantial amount of empirical research has now demonstrated that RAIs currently in use are less accurate for females, people of color, youth, and the economically disadvantaged.[43]

Instruments may lack validity for a myriad of reasons. For one, a model may fail to understand the complexities of relationships between variables and simply conform to the majority group's predictors. This can happen through the aforementioned omitted variable bias, having insufficient or unequal training data, or favoring simplicity over validity.

If different strata of a group have significantly different predictors of an outcome, then not accounting for that in a model will greatly reduce the accuracy for the minority group(s).

---

[43]Ibid., 51; Kelly Hannah-Moffat, "Actuarial Sentencing: An 'Unsettled' Proposition," *Justice Quarterly* 30, no. 2 (April 1, 2013): 14.; Tracy L. Fass et al., "The LSI-R and the COMPAS: Validation Data on Two Risk-Needs Tools," *Criminal Justice and Behavior* 35, no. 9 (September 2008): 1095, doi:10.1177/; John Monahan, Jennifer Skeem, and Christopher Lowenkamp, "Age, Risk Assessment, and Sanctioning: Overestimating the Old, Underestimating the Young.," Law and Human Behavior 41, no. 2 (2017): 191; Michael D. Reisig, Kristy Holtfreter, and Merry Morash, "Assessing Recidivism Risk across Female Pathways to Crime," *Justice Quarterly* 23, no. 3 (September 1, 2006): 384.; Craig S. Schwalbe et al., "Classifying Juvenile Offenders According to Risk of Recidivism: Predictive Validity, Race/Ethnicity, and Gender," *Criminal Justice and Behavior* 33, no. 3 (June 2006): 321, doi:10.1177/0093854806286451; Patricia Van Voorhis, "Classification of women offenders: Gender-Responsive Approaches to Risk/Needs Assessment," *Community Corrections Report on Law and Corrections Practice* 12, (2005): 19-20; Cheryl Marie Webster and Anthony N. Doob, "Classification without Validity or Equity: An Empirical Examination of the Custody Rating Scale for Federally Sentenced Women Offenders in Canada," *Canadian Journal of Criminology and Criminal Justice* 46 (2004): 395; Kevin W. Whiteacre, "Testing the Level of Service Inventory–Revised (LSI-R) for Racial/Ethnic Bias," *Criminal Justice Policy Review* 17, no. 3 (September 2006): 330, doi:10.1177/0887403405284766; Ivan Zinger, "Actuarial Risk Assessment and Human Rights: A Commentary," *Canadian Journal of Criminology and Criminal Justice* 46 (2004): 607;

This idea was corroborated by a study of the popular LSI-R risk assessment instrument, which found that it only correctly classified female offenders when they offended similarly to men or not in a gendered way.[44] Many parallel studies indicate that instruments generally produce more accurate predictions for male offenders than female offenders.[45] If predictors of crime vary by gender, then subgroup validity would necessitate the construction of a model sophisticated enough to account for different gendered criterion. In courtroom settings, where risk assessment instruments often take the form of hand-scored surveys, this need is likely to go unrealized.

Models may also suffer from subgroup invalidity if there is insufficient training data on minority groups in the population.[46] For example, two studies found that a risk assessment tool used to determine security classifications was not culturally competent and had no predictive validity for the female aboriginal population it was being used on. As a result, it unjustly overclassified minority women into higher levels of scrutiny.[47]

Using an instrument on a population or situation for which it was not designed may also reduce validity. Often referred to as mission creep, recycling crime prediction software is rather

---

[44]Reisig, Holtfreter, and Morash, "Assessing Recidivism Risk across Female Pathways to Crime," 384.

[45] Van Voorhis, " Classification of women offenders," 19; Schwalbe et al., "Predictive Validity, Race/Ethnicity, and Gender," 321; Chi Meng Chu et al., "The Utility of the YLS/CMI-SV for Assessing Youth Offenders in Singapore," *Criminal Justice and Behavior* 41, no. 12 (2014): 1437–57, https://doi.org/10.1177/0093854814537626.

[46]Fass et al., "The LSI-R and the COMPAS," 1095; Fons Van De Vijver and Norbert K Tanzer, "Bias and Equivalence in Cross-Cultural Assessment: An Overview," *European Review of Applied Psychology* 54, no. 2 (June 2004): 121, https://doi.org/10.1016/j.erap.2003.12.004.

[47] Webster and Doob, "Classification without Validity or Equity.".; Ivan Zinger, "Actuarial Risk Assessment and Human Rights: A Commentary," *Canadian Journal of Criminology and Criminal Justice* 46 (2004): 607.

common in the criminal justice system due to restricted budgets and a penchant for quick fixes. The push for actuarial decision-making has caused many counties to prematurely adopt a risk assessment instrument that proved successful in an entirely different jurisdiction without re-validating it for variations in their own county. Similarly, instruments constantly get repurposed into different contexts within the same population, like from predicting dangerousness in the pre-trial context to determining parole decisions. [48]

Certainly some statistical controls can help alleviate the extent to which a model lacks predictive validity – like matched pairs designs and computerized assessments. However, this first requires crime prediction software to identify how criminal predictors vary across all possible attributes and then take steps to account for the variation. Secondly, guaranteeing equal validity for multiple factors would require models likely too intricate to be manually administered quickly. Nevertheless, creating a succinct and generalizeable model, the goal of most crime prediction software, requires averaging the predictions for the entire population; In which case, minority groups will inevitably suffer the brunt of the invalidity.

The Base Rate Problem

Even if a model does have subgroup validity, it can still result in a disparate impact if the groups have different prevalence rates of the outcome. This phenomenon was first pointed out by the now notorious ProPublica and Northpointe debate.

---

[48]Sandy Jung and Edward P. Rawana, "Risk and Need Assessment of Juvenile Offenders," Criminal Justice and Behavior 26, no. 1 (March 1, 1999): 70, https://doi.org/10.1177/0093854899026001004; David Wall, "From Post-Crime to Pre-Crime: Preventing Tomorrow's Crimes Today," Criminal Justice Matters 81, no. 1 (September 1, 2010): 22, https://doi.org/10.1080/09627251.2010.505396.

In 2016, investigative journalists at ProPublica published an article claiming that the popular sentencing tool COMPAS was biased against blacks defendants.[49] In their analysis of the tool, ProPublica claimed that black defendants were misclassified as high risk- meaning they were predicted to recidivate but did not- twice as often as white defendants. Similarly, they found that white individuals were misclassified as low risk - meaning they were not predicted to recidivate but did- 63.2 percent more often than black individuals.[50]

Northpointe, the company that created the COMPAS instrument, rebutted that the difference in misclassification rates is due to the base rates of recidivism.[51] They argue that since a higher proportion of black individuals recidivate, more will be predicted to recidivate, which will lead to a greater, but proportional, number of misclassifications. While Northpointe believes the model is fair since the subgroup validity is equal for both black and white individuals, ProPublica argues that a greater number of false predictions for black individuals will have a disparate impact.

Variations of Predictive Accuracy

Since the COMPAS debate, a large body of work has been devoted to defining and clarifying different theories of fairness and nondiscrimination in the context of predictive

---

[49] Julia Angwin et al., "Machine Bias," *ProPublica*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[50] Jeff Larson et al., "How We Analyzed the COMPAS Recidivism Algorithm," ProPublica, May 23, 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[51]William Dieterich, Christina Mendoza, and Tim Brennan, "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity,"*Northpoint Inc*, 2016, https://assets.documentcloud.org/documents/2998391/ProPublica-Commentary-Final-070616.pdf.

validity.[52] Most focus on various formulations of the confusion matrix- a type of contingency table that displays the predicted outcomes in relation to the observed outcomes. These tables are useful when predicting a binary outcome (like recidivism v. non-recidivism) by juxtaposing the correct predictions to incorrect (see Figure 3).

Figure 3. Confusion Matrix

		OBSERVED	
		Y = 0	Y = 1
PREDICTED	$\hat{Y}$hat= 0	True Negative	False Negative
	$\hat{Y}$hat=1	False Positive	True Positive

Different formulations of the table hold different weight in different contexts to different stakeholders. For example, ProPublica finds false positives particularly egregious in the pretrial context since it dictates being put in jail unnecessarily. On the other hand, judges may be more concerned about true negatives - letting someone free who then commits another crime. These contentions have resulted in multiple proposed controls for group fairness.

[52] Richard Berk et al., "Fairness in Criminal Justice Risk Assessments: The State of the Art," *ArXiv:1703.09207v2 [Stat.ML]*, 2017, https://crim.sas.upenn.edu/sites/crim.sas.upenn.edu/files/2017-1.0-Berk_FairnessCrimJustRisk.pdf; Alexandra Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big Data* 5, no. 2 (June 2017): 153–63, https://doi.org/10.1089/big.2016.0047; Kleinberg, Mullainathan, and Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores,"; Singh, "Predictive Validity Performance Indicators in Violence Risk Assessment."

For example, a white house report called for "equal opportunity by design" which many have interpreted to mean equal opportunity for the advantageous outcome.[53] In the context of the criminal justice system, this would be a prediction of low risk. A stricter version of equal opportunity requires a model to have both equal true positive rates and equal false positive rates across groups.[54] Other conceptions focus on the balance of risk scores, stipulating that the average score assigned to those that engaged in the outcome is the same across groups or the average score assigned to those that did not engage in the outcome is the same for each group. Conversely, balance for the positive class tests that the true positive and true negative rate is the same for each group. Predictive parity, or the positive predictive value, stipulates that among those who were predicted to fail, the proportion of individuals who actually fail is the same for both groups. Alternatively, the negative predictive value tests that the proportion of those that succeeded among those that were predicted to succeed is the same across groups.

All of these formulas have been posited as definitions of predictive fairness. Unfortunately, there is no "one size fits all" approach since many of these conditions of non-discrimination are actually statistically incompatible with one another.

Accurate or Unequal- Impossibility Theorems

Kleinberg et al. (2017) and Chouldechova (2017) found that when an outcome has different prevalence rates across subgroups, it is statistically impossible to satisfy both

---

[53]Big data: A report on algorithmic systems, opportunity, and civil rights. Executive Office of the President, May 2016.; Moritz Hardt, Eric Price, and Nathan Srebro, "Equality of Opportunity in Supervised Learning," *ArXiv:1610.02413 [Cs]*, October 7, 2016, http://arxiv.org/abs/1610.02413.

[54] Moritz Hardt, Eric Price, and Nathan Srebro, "Equality of Opportunity in Supervised Learning," *ArXiv:1610.02413 [Cs]*, October 7, 2016, http://arxiv.org/abs/1610.02413. PAGE 3

calibration and error rate balance (i.e. predict accurately and proportionately). [55] Essentially, this research shows that any situation with unequal base rates across groups will inevitably have a disparate impact. In the COMPAS context, this means that either: a) the model's probability estimates are systematically skewed upward or downward for at least one race b) the model assigns a higher average risk estimate to non-recidivists in one race than the other or c) the models assigns a higher average risk estimate to recidivists in one race than the other.

The takeaway from many academics about the impossibility theorem is that the appropriate stakeholders must decide the appropriate fairness standards in each context. Although this may be a satisfactory solution in situations where base rates truly differ, academics haven fallen short of re-contextualizing the problem within the criminal justice realm, separate from other fields.

Articles like Kleinberg's jump from discussing criminal justice outcomes to medical diagnostics when describing the consequences of the impossibility theorem.[56] However, base rates in the medical field are more reflective of reality due to the extensive data and reporting mechanisms in the industry, the use of controlled experimental studies, and the easier detection of predictive error. Unlike medical diagnostics, crime is only recorded if it is captured and documented by the criminal justice system.

Therefore, if a model is predicting crime accurately, yet disproportionately by subgroup, one of the following realities must be true: 1) there are true base rate differences because of innate differences by subgroup 2) there are true base rate differences by subgroup due to societal

---

[55] Chouldechova, "Fair Prediction with Disparate Impact," 163; Kleinberg et al., "Inherent Trade-Offs in the Fair Determination of Risk Scores," 17.

[56] Ibid.

and sociological factors rather than inherent differences 3) the base rate differences are not true differences but are a product of biased data. To put this back in the COMPAS context, either black individuals innately commit crime more often than white individuals, black individuals commit crime more often than white individuals because of factors/life circumstances that are not controlled for in the model, or crime data is not accurately measuring the true commission of crime.

For the COMPAS developers to argue that their model is not discriminating, they must believe in the first or second theory. Since there is no evidence that criminogenic tendencies vary by race, and few developers would contend that there are innate racial differences in criminal propensity, no further attention will be paid to the first theory. [57] In the case of option two, they believe accurate predictions outweigh controlling for societal issues, meaning black individuals should be faulted for committing more crime regardless of the reasons why they do. On the other hand, the fact that so much attention has been paid to mitigating racial differences insinuates that the majority of the population believes in either theory two or three. Therefore, there is general agreement that the base rate differences are a sociological problem rather than a racial problem. In which case, they believe that people should not be faulted for the societal and sociological reasons that cause them to commit more crime or corrections should be made for biased data. Research on the history of policing in America as well as the accuracy of crime data suggest that the real answer is a combination of options two or three.

[57]Nikolas Rose, "The Biology of Culpability: Pathological Identity and Crime Control in a Biological Culture," *Theoretical Criminology* 4, no.5 (2000): 16, https://doi.org/10.1177/1362480600004001001.

From the beginning of organized crime control in America, the designation of criminals and "high-crime" neighborhoods has never been an impartial process. [58] The early criminal justice system was, in all reality, a cosmetic replacement for slavery that preserved the power imbalance and racism underlying it through the weaponization of criminalization. [59] While chain gangs and convict leasing may seem like a thing of the past, "Stop and Frisk" - a New York City tactic that over-policed communities of color and over-populated their rap sheets - was only ruled unconstitutional in 2013.[60] The oppression of minority communities since America's nascence has also created an entire ecological system that subjugates black individuals disproportionately to consequences like the school-to-prison pipeline. Thus, the difference in "crime rates" observed between black and white individuals is tainted by the history of oppression and differential selection by police.

Even disregarding the sociological reasons for base rate differences, crime data has still been proven time and again to be systematically biased based on the types of quotas, neighborhoods, and tactics used to target it. For example, COMPAS defines crime as "a finger-printable arrest involving a charge and a filing for any uniform crime reporting (UCR) code."[61] Countless studies have shown that uniform crime reports are not representative of all crime but

---

[58] Michelle Alexander and Cornel West, The New Jim Crow: Mass Incarceration in the Age of Colorblindness, Revised edition (New York: New Press, 2012).

[59] Alex Lichtenstein, "Good Roads and Chain Gangs in the Progressive South: 'The Negro Convict Is a Slave,'" The Journal of Southern History 59, no. 1 (1993): 89, https://doi.org/10.2307/2210349.

[60] Floyd et al v. City of New York (S.D.N.Y. 2013)

[61] Tim Brennan, William Dieterich, and Beate Ehret, "Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System," *Criminal Justice and Behavior* 36, no. 1 (January 2009): 26.

have systematically uncounted and overrepresented observations that are not indicative of the real proportions in crime commission.[62] Crime prediction software that relies on police-generated data is incorporating all of these human biases and misrepresentations.[63] Even if one is unconvinced of the reasons for the base rate differences, in order to not discriminate, it is necessary to constrain the model to account for them. However, any proposed controls for the base rate will inevitably cause issues with predictive validity because the outcome variable has not changed.

The Outcome Variable Problem

The fact that developers are trying to control for racial equality while predicting a racially unequal outcome is circular logic. Without changes to the sociological factors causing the differences, the outcomes are still going to be disparate and suffer the same biases - the difference is the model just won't predict them accurately. There is no way to non-discriminatorily predict an outcome that always has been and still is discriminatory.

---

[62]American Civil Liberties Union, "Report: The War on Marijuana in Black and White," accessed April 14, 2018, https://www.aclu.org/report/report-war-marijuana-black-and-white; Steven D. Levitt, "The Relationship Between Crime Reporting and Police: Implications for the Use of Uniform Crime Reports." *Journal of Quantitative Criminology* 14, no. 1 (1998): 61.; Michael D Maltz, "Bridging Gaps in Police Crime Data" (U.S. Department of Justice, n.d.), https://www.bjs.gov/content/pub/pdf/bgpcd.pdf; Ojmarrh Mitchell and Michael S. Caudy, "Examining Racial Disparities in Drug Arrests."*Justice Quarterly* 32, no. 2 (2015): 288; David B. Wilson, Tammy Rinehart Kochel, and Stephen D. Mastrofski, "Race and the Likelihood of Arrest." In Encyclopedia of Criminology and Criminal Justice, edited by Gerben Bruinsma and David Weisburd, 4245. Springer New York, 2014; Heather Zaykowski, "Racial Disparities in Hate Crime Reporting." *Violence and Victims* 25, no. 3 (June 1, 2010): 378.

[63] Developers have tried to account for this by only looking at well-documented crime, such as violent crimes involving a victim, or convictions instead of arrests. Still, research shows that this data is still biased and misrepresentative.

Even if that is an acceptable answer to some, to at least predict equally across races, it is still not possible for all sensitive attributes. Racial bias has dominated the conversation about discrimination since most progressive reform efforts are aware of and trying to compensate for America's dark history of racial oppression. It should be noted again though, that this issue pertains to any sensitive attribute, or any attribute at all, that appears to have base rate differences in crime data. Compensating for this by equalizing predictions across multiple sensitive attributes is, as mentioned before, under-researched and at some point, will likely render prediction useless.[64]

There are certain criminal justice outcomes that do not suffer from this outcome variable problem. Data on defendant's failed appearances, for example, does not suffer from the same observation and selection biases since they are systematically observed for every case and every defendant. However, datasets would need to be severely manipulated to account for the fact that the outcome variable (failure to appear) is only observable for defendants who judges have decided to release. Not accounting for the counterfactual will result in the model regurgitating the same judicial biases that the system has always carried and thus, discriminating against the individuals the system has always discriminated against.

At the end of the day, risk assessment instruments trained on misrepresentative data will inherent the data's flaws. Conscious efforts to account for the biases are only sensible if the outcome being predicted is not similarly biased. If the bias exists at both ends of the spectrum,

---

[64] Friedler et al., "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning."; Kleinberg, Mullainathan, and Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores."

as is the case of police-generated data, then the model can either succumb to the same biases or operate counterintuitively to the goals of the software.

Discrimination by Interaction

As evidenced through the outcome variable problem, crime prediction software does not necessarily need to be discriminatory in and of itself to produce discriminatory results. Since crime prediction software is retrained, it is responsive to society's interactions with it. Retraining models with the data that resulted from their predictions is standard protocol that allows the model to identify which predictions it got right, which it got wrong, and compensate for any changes in the correlations between variables. However, with retraining comes a slew of issues. Many examples have shown that responsive technology is not immune to existing societal biases; Rather, it learns them quite quickly.[65]

Exacerbating Disparate Impact

This is especially problematic with crime predictions software since it does more than just predict; it prescribes behavior. Risk assessment instruments determine release statuses, sentences, and liberty restrictions. Predictive policing software tells police which neighborhoods to surveil and which individuals to target. As a result, crime prediction software can fall victim

[65]Hannah Devlin, "AI Programs Exhibit Racial and Gender Biases, Research Reveals," The Guardian, April 13, 2017, http://www.theguardian.com/technology/2017/apr/13/ai-programs-exhibit-racist-and-sexist-biases-research-reveals; Latanya Sweeney, "Discrimination in Online Ad Delivery," Queue 11, no. 3 (March 2013): 29, https://doi.org/10.1145/2460276.2460278; Daniel Victor, "Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.," The New York Times, March 24, 2016, sec. Technology, https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html.

to confirmation bias.[66] Stuck in a self-fulfilling feedback loop, the instruments transition from prediction to technological determinism.

For instance, if predictive policing software allocates police to one neighborhood at a slightly higher rate than others, and the police are incentivized to make arrests, then they will inevitably arrest more individuals from that neighborhood. Eventually, when the model is retrained, the new data will depict an uptick in crime in the neighborhoods that were more heavily policed. Thus, the model will confirm that it made the right predictions and will tell police to target that neighborhood even more. This phenomenon has been dubbed "the ratchet effect."[67]

The most compelling research yet on the ratchet effect comes from a study of the popular predictive policing software PredPol. The study showed that predictive policing based on reported crime data not only succumbs to historical biases but actually exacerbates them. In their simulation of the popular PredPol algorithm, Lum and Isaac (2016) found that the software increased the historically disproportionate and illogical distribution of officers to poor and minority communities of color. Even though drug crime is almost equivalent across races and neighborhoods, the use of PredPol would increase the targeting of black individuals to twice the rate of white individuals and concentrate police in low-income communities.[68]

Advocates of crime prediction software claim that this effect will be curtailed if the high offending group changes their behaviors due to the increased police targeting. However, groups

---

[66]Kristian Lum and William Isaac, "To Predict and Serve?," *Significance* 13, no. 5 (October 2016): 16, doi:10.1111/j.1740-9713.2016.00960.x.

[67] Harcourt, Against Prediction.

[68] Lum and Isaac, "To Predict and Serve?,"19.

differ in their elasticity to police presence which has not been proven to decrease group offending rates.[69] In practice, studies have shown that police targeting only increases the likelihood of arrests for the targeted group.

For instance, a pilot predictive policing program in Chicago created a Strategic Subjects List of individuals who were predicted to be likely victims of gun violence. The individuals were then targeted with police interventions. However, a follow-up study of the program found that individuals on the list were not at any higher risk for gun violence. Rather, they ended up being more likely to get arrested for a shooting due to the increased police surveillance.[70]

Risk assessment instruments similarly prescribe behavior. If an RAI consistently predicts one group as higher risk, then more individuals from that group will receive harsher detention, punishment, or sentencing decisions. This reduces the groups' chances of altering the model since fewer group members are given the opportunity to demonstrate the counterfactual.

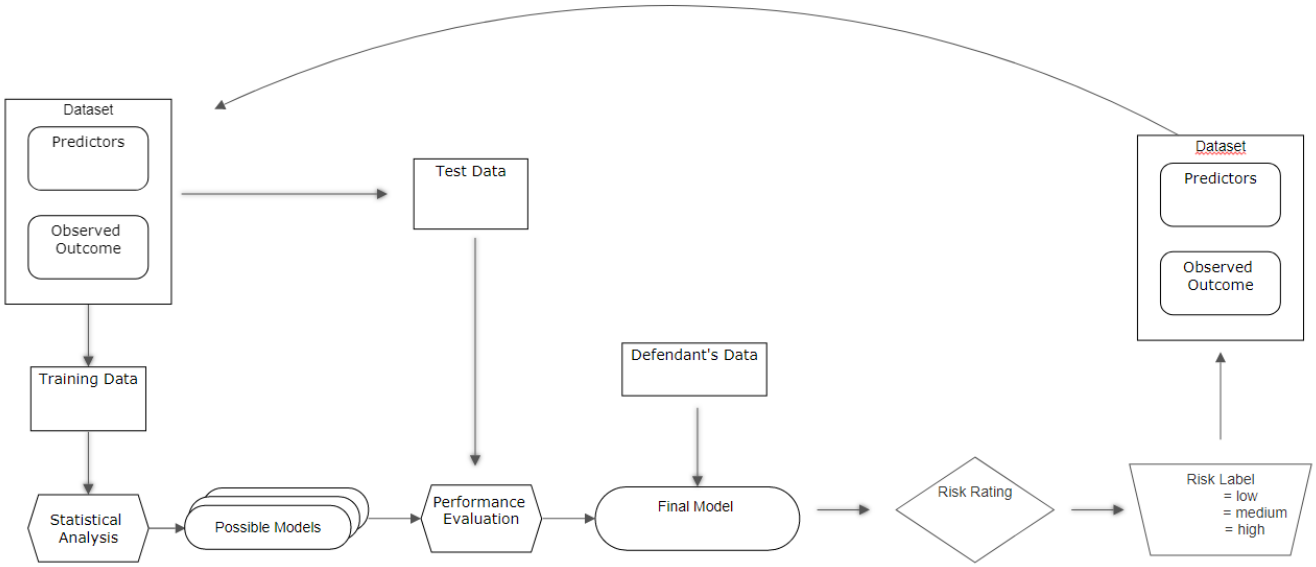
The solution to this problem cannot be to avoid retraining models. Data must be collected on the accuracy of model's predictions to assure its continued validity. Thus, this is a kind of damned if you do, damned if you don't problem. The only true remedy to this issue is starting with a blank slate. Making equal predictions and allocations will better illuminate the true trends in the data instead of the results of the observer effects.

Figure 4. Retraining Model Process

---

[69]Harcourt, Against Prediction, 28; Priscilla Hunt, Jessica M. Saunders, and John S. Hollywood, *Evaluation of the Shreveport Predictive Policing Experiment* (Santa Monica, CA: RAND Corporation, 2014).

[70]Jessica Saunders, Priscillia Hunt, and John S. Hollywood, "Predictions Put into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot,"J*ournal of Experimental Criminology* 12, no. 3 (September 1, 2016): 347.



Algorithmic Authority

Ensuring homogenous decisions in similar situations is the main attraction of crime prediction software. The extent to which this goal can be achieved depends on the extent to which ultimate authority is acquiesced to crime prediction software. In most courtrooms and precincts, crime prediction software is intended to function as a decision aid, not an omnipotent authority. In theory, this accounts for the fact that the model may not have all the information pertinent to the decision. Just the same, any level of human discretion can lead to imbalanced outcomes.

For one, data fundamentalists may give too much credence to the assessments put in front of them and fail to do their due diligence in assessing the individual case. Trusted environments can support undue accreditation and cause an over-reliance on automation. At the same time, judges who have consistently made draconian and discriminatory decisions are unlikely to give more weight to some correlations put in front of them than their practical experiences,

prejudices, or gut instinct.[71] One study found that probation officers would manipulate the variables input into a risk assessment model to receive the outcome they believe the defendant deserved.[72] Likewise, police officers may use predictive policing outputs as parallel construction to conduct the same discriminatory and unjustly targeted searches they did before. Despite algorithmic proselytizers arguing that the software increases transparency and accountability in decisions, splitting authority between the magistrate and the model does not guarantee a departure from the status quo. It can just as easily be used to conceal or justify discriminatory decisions.

At the same time, attempts to equalize the weight of the prediction by limiting human discretion, or removing it altogether, are even more treacherous. While human biases vary based on the individual's experiences and preferences, crime prediction software homogenizes discrimination. Assuming the quantity and quality of injustice is the same, one specific, consistent, central discriminatory practice is more harmful than diverse and localized ones. Indoctrinating crime prediction software that specifically and relentlessly discriminates against one group oppresses them entirely due to the immense individual and societal collateral consequences of a criminal record. In deciding between two evils, diverse human discrimination has less opportunity to condemn an entire group.

---

[71] Angèle Christin, "Algorithms in Practice: Comparing Web Journalism and Criminal Justice," *Big Data & Society* 4, no. 2, (December 2017): 10, https://doi.org/10.1177/20539517; Gina M. Vincent et al., "Impact of Risk/Needs Assessment on Juvenile Probation Officers' Decision Making: Importance of Implementation.," *Psychology, Public Policy, and Law* 18, no. 4 (2012): 554, https://doi.org/10.1037/a0027186; Gina M. Vincent et al., "Risk Assessment Matters, but Only When Implemented Well: A Multisite Study in Juvenile Probation.," *Law and Human Behavior* 40, no. 6 (2016): 684, https://doi.org/10.1037/lhb0000214.

[72] Angèle Christin, "Algorithms in Practice," 10.

It's important to remember that similar promises for increased impartiality were used when imposing mandatory minimums and federal sentencing guidelines within the criminal justice system. The professed intent was to reduce sentencing disparities via the reduction of judicial discretion. Instead of reducing disparities, they contributed to them.[73] Power was not removed but simply shifted to the prosecutors and the legislators of the guidelines. Mandatory minimums for drug crimes, for example, were strategically massaged to enforce much harsher sentences for the crack cocaine used in black communities compared to the powder cocaine used in white communities. [74]

Other rationales used to impose sentencing guidelines have also been mimicked in crime prediction propaganda. Proclamations that even if crime prediction software does not reduce disparities, it will at least shed light on the opaque and intricate rationale behind police and judicial decisions, are common. Past evidence dictates that the guidelines were used as a safeguard for judges and prosecutors to defend the objectivity of their judgments. [75] Advocates of the guidelines - who previously believed their flaws could be improved through incremental reform - have backtracked their statements, arguing against any move to aggregated assessment

---

[73] Fischman Joshua B. and Schanzenbach Max M., "Racial Disparities Under the Federal Sentencing Guidelines: The Role of Judicial Discretion and Mandatory Minimums," *Journal of Empirical Legal Studies* 9, no. 4 (November 6, 2012): 729–64, https://doi.org/10.1111/j.1740-1461.2012.01266.x.

[74] American Civil Liberties Union, "Racial Disparities in Sentencing," Written Submission to the Inter-American Commission on Human Rights, 153rd Session, October 27, 2014, https://www.aclu.org/other/aclu-submission-inter-american-commission-human-rights-racial-disparities-sentencing.

[75] Frank O. Bowman, "The Failure of the Federal Sentencing Guidelines: A Structural Analysis," *Columbia Law Review* 105, no. 4 (2005): 1319, http://www.jstor.org/stable/4099435.

in the criminal justice system. [76] Lessons learned through the failure of the federal sentencing guidelines should not be overlooked when considering the imposition of more regulations thinly veiled as informational aids.

While various other issues come to play with regard to technological due process, they can be summed up with a similar paradox: any interaction of humans with crime prediction software exposes the models to the same human biases as have always existed. However, delegating omnipotent and unwarranted authority to predictive software will almost certainly exacerbate the biases it is trying to alleviate.

Present Study

Transitioning from theoretical to empirical, the present study offers a contextualized demonstration of how crime prediction software can discriminate against a protected group by design, training, and implementation. By examining a never-before-tested risk assessment instrument, this study seeks to ensure that the human rights of New York City residents whose liberty depends on this instrument are protected. Furthermore, this research extends the conversation beyond the COMPAS debate by using an entirely new dataset and checking for heterogeneous definitions of discrimination. Even though only one model is examined, this methodology can be generalized to any model attempting to predict a criminal outcome tied to police activity.

Methodology

The CJA Instrument

[76] Alschuler, Albert. "The Failure of Sentencing Guidelines: A Plea for Less Aggregation." *The University of Chicago Law Review* 58, no. 3 (1991): 911, https://doi.org/10.2307/1599992.

In 2015, the Criminal Justice Agency (CJA) of New York City developed a risk assessment instrument to predict pretrial felony re-arrest (hereafter, CJA tool).[77] The instrument is currently used on defendants in New York City at their first appearance on a misdemeanor or felony case. Defendants are scored based on eight variables that include their age, fulltime activity, and criminal history. Defendants' raw scores ranging from -16 to +18 are then converted into one of five risk categories ranging from low risk to high risk (see Table 1).

The CJA model was designed to determine a defendant's eligibility for "supervised release" – a supervision program offered as an alternative to pretrial incarceration. According to supervised release practitioners, defendants scoring medium-risk and below are considered for a release alternative, defendants in the medium-high range are considered for supervised release if they would otherwise be incarcerated, and defendants scoring in the high range are typically considered too risky. In practice, the risk score can be used by defense attorneys to advocate to the judge for or against certain pre-trial dispositions, like releasing a defendant on their own recognizance.

This tool was selected for the case study because of its transparency, simple design, and lack of independent reviews. First, this tool is one of the rare instruments created by a public city agency so it is not subject to confidentiality or trade secret provisions. The scoring model has been made available to select independent researchers like myself. Secondly, the simplistic design of this RAI, from its development methods to its scoring model, make it an easy example to demonstrate how racial bias can become encoded in an algorithm. Lastly, this model has never been reviewed by anyone other than the agency who created it; and the creators of the

[77]Eion Healy (2015). *Research Report for MOCJ's Pretrial Felony Re-Arrest Risk Assessment Tool*. New York, NY: New York City Criminal Justice Agency. [unpublished]

model never tested it for any types of discrimination or subgroup validity. Nevertheless, versions of the tool are now being considered for use in bail decisions, illustrating a picturesque example of mission creep.[78] It is an opportune time to assess the discriminatory potential of this instrument before it is expanded to other contexts.

Participants

Participants in this study were drawn from a dataset of criminal defendants obtained through The Legal Aid Society. The Legal Aid Society tracks data on every defendant that was arrested and arraigned by their organization.[79] Case data, demographic data, and arraignment paperwork for all Legal Aid clients arraigned for a misdemeanor or felony arrest from December 1, 2015 through December 15, 2015 were pulled from their case management system (N=5823). [80] Since supervised release was not implemented across New York City at this time, multiple data sources were combined to acquire the necessary data for each defendant in order to calculate pseudo risks scores.

[78]New York University School of Law, "Redesigning New York City's Pretrial Risk Assessment and Recommendation System," *The Docket (blog)*, September 18, 2017, http://blogs.law.nyu.edu/docket/events/redesigning-new-york-citys-pretrial-risk-assessment-and-recommendation-system/31861/.

[79]According to citywide arrest data maintained by the Office of the Chief Clerk of New York City Criminal Court, the Legal Aid Society represented the following percentage of defendants in each county of New York City in 2015: Bronx (44.09%), Kings (58.30%), New York (62.37%), Queens (56.76%), Richmond (74.93%).

[80] If participants had two arrests for different incidents within the 15-day period, then only the first arrest was retained in the sample and any subsequent arrests were counted as re-arrests. If a participant had any additional cases open within the 15-day for incidents occurring prior to the time-frame, then only the incident occurring within the time frame was retained in the sample. This may happen if an individual is arrested on new charges or returned on a warrant and a previous case is docketed again.

Participants' demographic data and pretrial release status was coded as it was identified in the Legal Aid Society database. Participants release status was defined into one of three categories: never released, partially released, or always released.[81] Optical Character Recognition technology and data mining scripts were used to scrape participants' criminal history data from their rap sheets and CJA interview forms.[82] Due to the multiple sources of information, all the data points were only available for 17% of the defendants originally identified, reducing the dataset to a final sample of 1012 participants. A detailed description of how data was obtained for each variable is available in Table 2 and the code is available in Appendix A.

Analysis

A three-pronged assessment was used to examine the CJA tool for discrimination based on race. This study first 1) assessed the various standards of predictive accuracy of the tool across racial subgroups; 2) tested the tool for mean score differences across racial subgroups that could result in a disparate impact; and 3) retrained the model based on the new data set.

Information on defendant's outcomes was also needed to check for validity. Participants' re-arrests up until January 15, 2017 were pulled from the Office of Court Administration statewide database. Even though the CJA tool is intended to predict felony re-arrest only during the pre-trial period every defendant in my dataset was given a two-year follow-up period for re-

[81] There was not sufficient data to determine exact time at risk in the partially released group. This variation is hopefully minimized by the longer two-year follow-up period.

[82] All defendants in New York City are interviewed about their full-time activity by the Criminal Justice Agency. These CJA interview forms are included in defendant's arraignment paperwork. The forms include information on defendant's full-time activity.

arrest, regardless of when their case closed.[83] The purpose of this was two-fold: it increased the sample size retention and revoked the need to account for various lengths of time at risk. While this increases the percentage of re-arrests, there is no reason to believe that it alters the subgroup breakdowns of re-arrest.

Results

Predictive Validity

The first aim of this study was to authenticate the reported predictive validity of the CJA tool since no sources have independently verified this claim to date. Following the "fairness through unawareness" technique, the CJA tool was created "race-blind" meaning it did not include race data in the training nor test datasets. Therefore, the predictive validity of the instrument across racial subgroups was never tested. I hypothesized that the instrument will demonstrate acceptable overall predictive validity according to the basic standards set in risk assessment literature. Similarly, I hypothesized that the predictive validity will not significantly vary across these racial subgroups. These hypotheses are based on a few factors. First, CJA reports that they selected the model with the best optimization without constraining for any standards of fairness so there is no reason to believe that the model will not predict sufficiently well for the majority. Additionally, my dataset is sampled from the same New York City population of defendants as the training data, just 6 years later. Lastly, I only had sufficient data to test for racial differences across black and white subgroups. Since black and white defendants

[83] Two-years was selected since CJA gave each case a max two year follow-up period to reach a plea or disposition.

made up a large majority of the training data, invalidity for one group would be highly consequential for the overall validity.[84]

Area Under the Curve

The most commonly reported measure of predictive validity, and the only measure of validity reported by CJA, is the Area Under the Curve (AUC) value. The AUC ranges in value from .5 to 1 and reports the probability that a randomly selected individual who was re-arrested received a higher risk classification than a randomly selected individual who was not re-arrested across cut-off thresholds.[85] The AUC value at which a model is deemed acceptable is subjective; in general, an AUC of .5 to .6 is considered a failure, .6 to .7 is considered poor but fair, 0.7 to 0.8 is acceptable, 0.8 to 0.9 is good, and anything above .9 is excellent.[86]

CJA reports an AUC of .680 for predicting any re-arrest (felony or misdemeanor) and .671 for felony re-arrest. Using my dataset, I calculated an AUC of .760 for any re-arrest and .713 for felony re-arrest when using the raw risk-score as the predictor. Since the risk scores are operationally converted into five categories, I also checked the model using the risk category as an ordinal numeric variable (1 to 5) to ensure that the category conversion does not greatly alter the model's performance. Using this method, I calculated an AUC of .760 for any re-arrest and .701 for felony re-arrest. In both cases, I find that the model has an acceptable discrimination

---

[84] If I had sufficient data to test for validity across the American Indian, Alaskan Native, or Asian subgroups, which did not make up a large majority of the training data, my hypothesis would be different.

[85] Although the AUC value is often incorrectly interpreted as a measure of predictive accuracy, it is actually a global measure of discrimination.

[86] https://ac.els-cdn.com/S1556086415306043/1-s2.0-S1556086415306043-main.pdf?_tid=34b4d0e6-6506-4d99-9028-5af8be6f4979&acdnat=1523215087_c27be5d5fae80e8ec3eba371dc0a47f9

rate on my dataset. In fact, the AUC values are better than those initially reported by CJA, although this is likely due to the fact that all the defendants in my sample had a homogenous two-year period to be re-arrested which is longer on average than the pre-trial period defined by CJA.

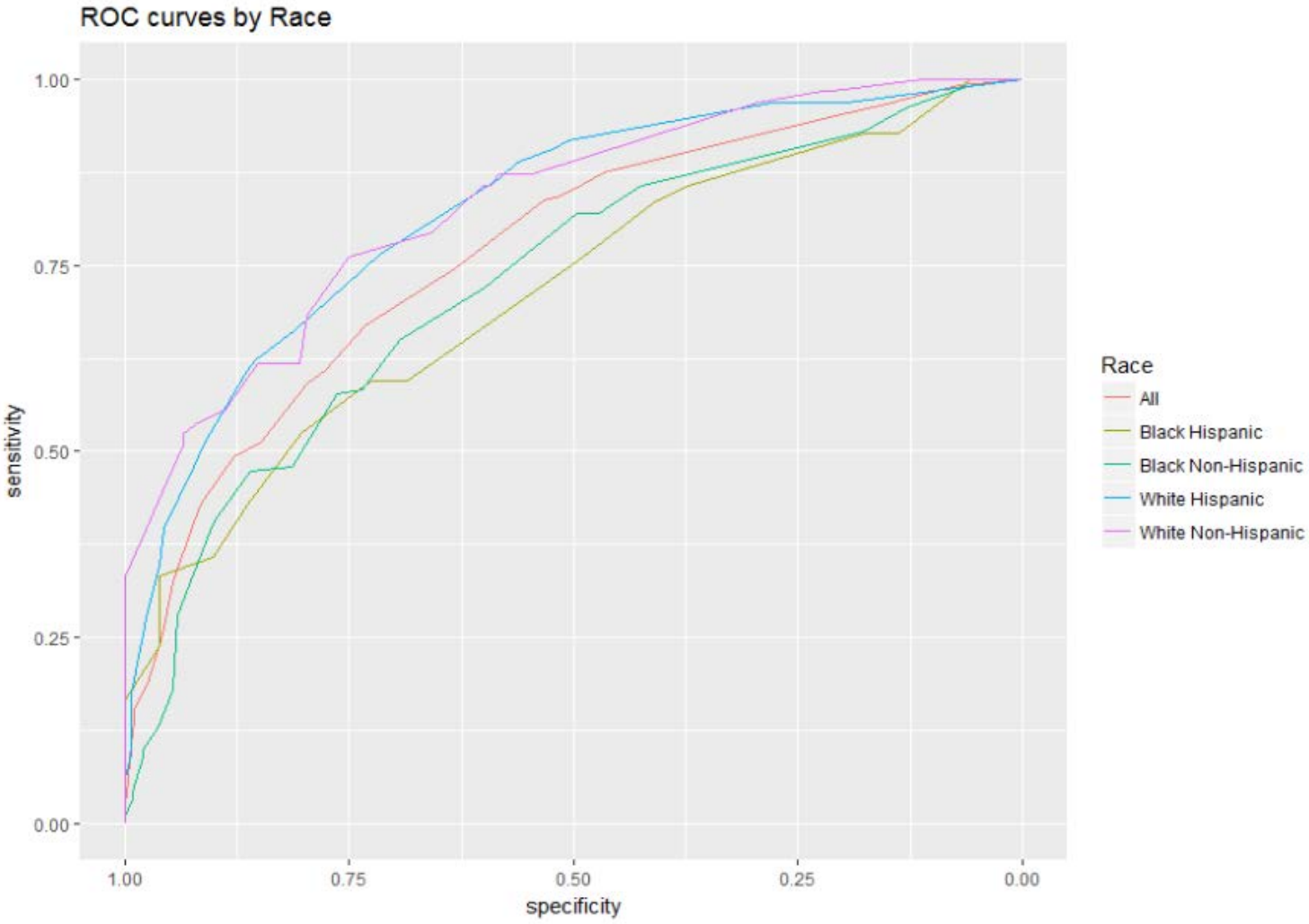
In order to test the overall discrimination by racial subgroup, I calculated the ROC curves for each of the racial strata for both felony re-arrest and any re-arrest using the raw risk score as well as the converted numeric risk category (1 to 5). The AUCs for both felony re-arrest and any re-arrest are listed in Table 3.

Table 3. AUC by Race and Re-arrest Type

Race	**Raw Risk Score**		**Ordinal Risk Category**	
	Any Re-arrest	**Felony Re-arrest**	**Any Re-arrest**	**Felony Re-arrest**
Black	**.723**	**.673**	**.715**	**.659**
Black Hispanic	.713	.634	.705	.641
Black Non-Hispanic	.727	.681	.718	.664
White	**.825**	**.765**	**.815**	**.758**
White Hispanic	.825	.712	.812	.712
White Non-Hispanic	.832	.815	.824	.801
Overall	**.769**	**.713**	**.760**	**.701**

The AUC values were compared across races using the Delong method with permutation tests. While there are significant differences across races when predicting re-arrest (p=.024), I cannot reject the null hypothesis that the AUCs across races are equal when predicting felony re-arrest (p = .422). When differences were tested using binary racial categories without regard for ethnicity, the AUC was significantly greater for white individuals versus black individuals for any re-arrest (p = .0008) but not for felony re-arrest (p = .07). Since I am assessing the tool's predictions for felony re-arrest, this result is inconsequential to our analysis.

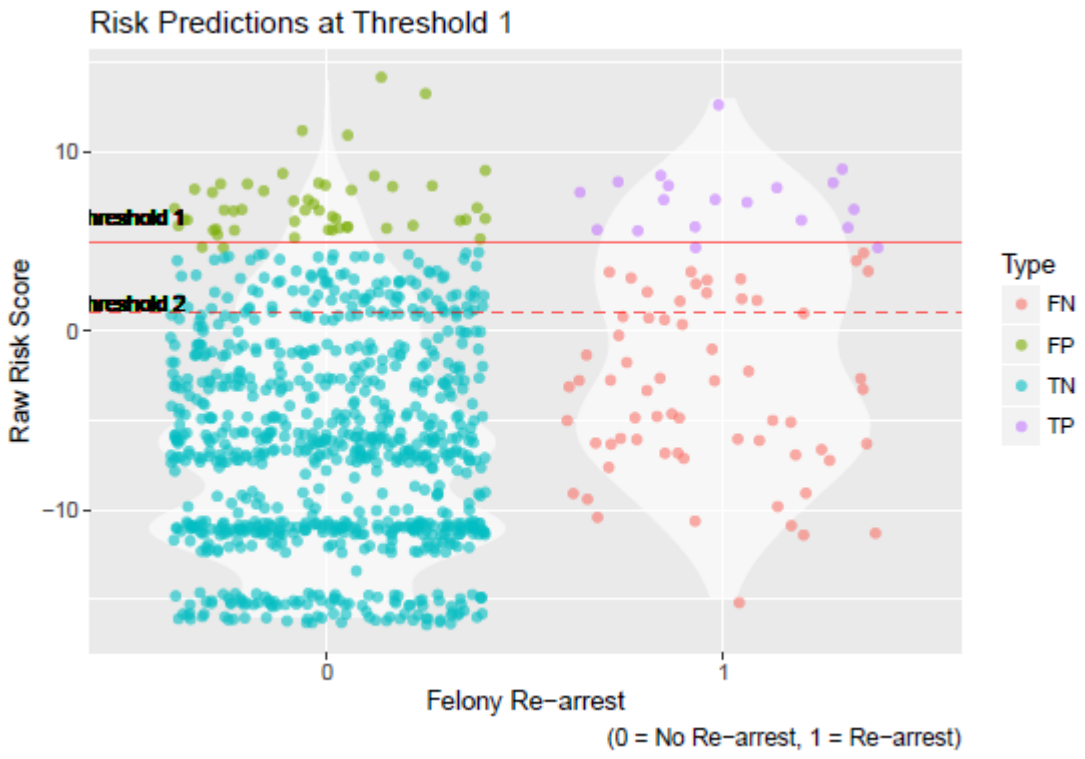
Figure 5. ROC Curves by Race



Performance Indicators Based on Confusion Matrix

Other calculations of discrimination and calibration require a dichotomous prediction. Thus, risk categories were artificially changed to a binary prediction of either felony re-arrest or no re-arrest. Since CJA did not report a single-cut off threshold, different thresholds were tested and findings did not differ significantly based on each threshold (see Figure 5).[87] The threshold reported here divided the cut off at or above the raw score of 5. This was the closest numerical break to picking the top 6.8% of individuals. The RAI reports a 6.8% mean felony re-arrest rate so this is in line with betting the base rate on the highest group.[88]

Figure 5. Risk Predictions at Tested Thresholds



[87] Cut-off points were also tested at the medium-high category and the raw score of 2.

[88] Cja reports about 6.25% of defendants in their training data fell into this category. In my dataset, 6.9% did.

Various discrimination and calibration measures used to check the predictive validity of the instrument were calculated overall and across racial subgroups (see Table 4).

Table 4. Model Performance by Race

Race	Sensitivity (TP/TP+FN)	Specificity (TN/TN+FP)	Positive Predictive Value (TP/TP+FP)	Negative Predictive Value (TN/TN+FN)	Number Needed to Detain (1/PPV)	Number Safely Discharged ((1/1-NPV) – 1)	Diagnostic Odds Ratio (TP x TN/FP x FN)	Accuracy (TP + TN/ TP+FP+TN +FN)
Black (N=546)	0.16	0.95	0.24	0.92	4.12	11.51	3.68	0.88
Hispanic (N=93)	0.20	0.93	0.14	0.95	7.00	20.50	3.42	0.89
Non-Hispanic (N=453)	0.16	0.95	0.27	0.91	3.71	10.54	3.88	0.88
White (N=427)	0.33	0.93	0.30	0.94	3.36	16.73	7.08	0.89
Hispanic (N=256)	0.28	0.95	0.31	0.95	3.20	17.46	7.94	0.91
Non-Hispanic (N=171)	0.40	0.90	0.29	0.94	3.50	15.67	6.27	0.86
Overall (N=973)	**0.23**	**0.94**	**0.27**	**0.93**	**3.68**	**13.33**	**4.97**	**0.88**

When testing generic accuracy, we see that defendants have an 88% chance of being correctly classified by the model. However, as mentioned before, neither the AUC value nor the overall accuracy differentiates between a model's ability to identify high-risk individuals versus low-risk individuals. Since there is only a 6.8% chance of recidivism, this model could have an accuracy of 93.2% simply by not predicting re-arrest for anyone. Contextually, risk assessment in the criminal justice context is more focused on the ability to identify high-risk individuals.

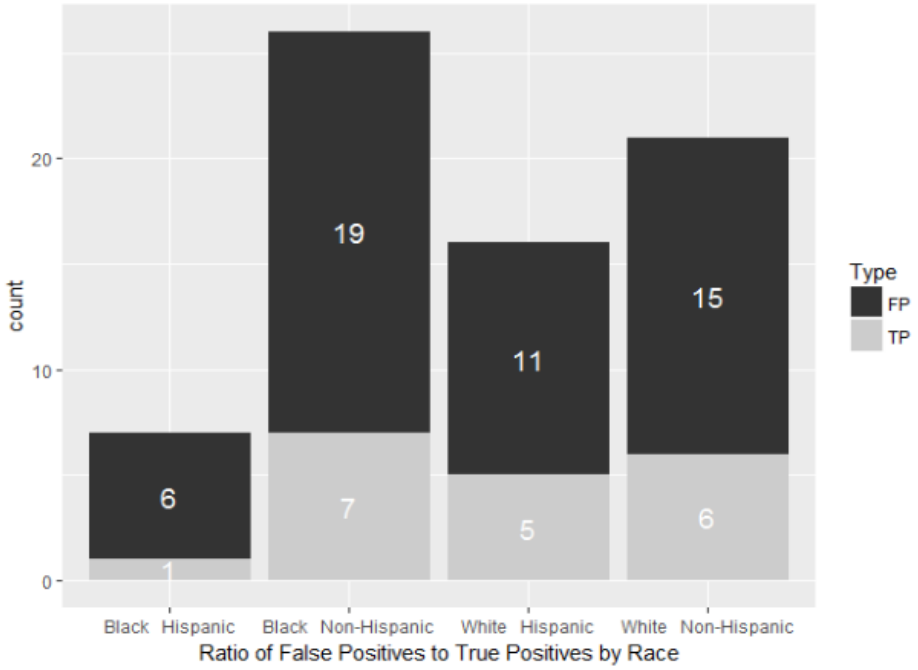
A model's ability to correctly identify high-risk individuals can be measured by its positive predictive value. The positive predictive value is the proportion of those actually re-arrested out of those predicted to be re-arrested. Conversely, the models ability to correctly identify low-risk individuals is the negative predictive value. I find that the CJA tool has an overall positive predictive value of .27 and negative predictive value of .93.

There is no de facto standard for these values since their acceptability must be analyzed in the context for which it is used. By calculating the inverse of the positive predictive value, the number needed to detain (NND), we can come up with a contextualized understanding of the consequences of inaccuracy.[89] The NND represents the number of individuals judged to be at high-risk who must be detained in order to prevent one re-arrest. Assuming all those judged to be at high-risk by the CJA tool are detained, then approximately three to four individuals will need to be detained to prevent one felony re-arrest. In this case, that calculates to saving 19 felonies at the cost of 70 unnecessary detentions.

It's important to ask though, who bears the burden of inaccuracy? We see a large divergence here by race. Seventy-six percent (25/33) of Black individuals were falsely detained compared to only 70% (26/37) of White individuals (see Figure 6). Although these differences do not reach statistical significance in our dataset (df = 67, p-value = 0.31), it is certainly something that should be paid attention to.

Figure 6. True Positive to False Positive Ratio by Race

[89] S. Fleminger, "Number Needed to Detain," *British Journal of Psychiatry* 171, no. 03 (September 1997): 287, https://doi.org/10.1192/bjp.171.3.287a; Singh, "Predictive Validity Performance Indicators," 12.



From a more positive standpoint, the number safely discharged shows how many low-risk individuals can be released before one re-arrest occurs.[90] We see that on average, 13 defendants can be discharged before one is re-arrested. Although contextualized, these values are still a matter of moral and subjective consideration regarding how many people one is willing to detain to prevent one felony re-arrest. Similarly, without regard for the subjectivity of the outcome, they can be used to justify more detention for one race than another under a guise of safety.

Odds Ratios

[90] S. Fazel et al., "Use of Risk Assessment Instruments to Predict Violence and Antisocial Behaviour in 73 Samples Involving 24 827 People: Systematic Review and Meta-Analysis," *BMJ* 345, no. 2 (July 24, 2012): 3, https://doi.org/10.1136/bmj.e4692; Singh, "Predictive Validity Performance Indicators," 12.

The diagnostic odds ratio is also a measure of high-risk discrimination that reports the odds of a true positive relative to the odds of a false positive. The CJA tool has an overall DOR of 3.39, meaning that for every one person correctly predicted to recidivate, at least three people will be predicted incorrectly. The Breslow-Day test was used to test the null hypothesis that the stratum-specific odds ratios are homogenous. Results indicate that the odds ratios are homogenous (X-squared = 1.0919, df = 3, p-value = 0.779).

In order to test the homogeneity of the odds ratios without a single cut-off point, logistic regression models were created for each of the racial stratum using the raw risk score. Consistent with the DOR results, the logistic odds ratios are within the confidence intervals of one another, indicating that there are no significant differences. Both the diagnostic and logistic odds ratio tests confirm that that the odds of a true positive relative to a false positive do not differ significantly by race.[91]

Table 5. Logistic Odds Ratio

Race	Logistic Odds Ratio	95% CI
Black	**1.11**	**1.05-1.16**
Black Hispanic	1.09	0.95- 1.28
Black Non-Hispanic	1.11	1.06- 1.17
White	**1.15**	**1.09-1.21**
White Hispanic	1.13	1.05- 1.22
White Non-Hispanic	1.17	1.08- 1.27
Overall	**1.12**	**1.09-1.17**

Risk Distribution

The risk thresholds of the CJA tool are also an indicator of model performance. Ideally, the risk classifications would result in re-arrest rates that differ substantially for every subsequent

[91] Remember, this is base-rate independent, so it does not mean that there will not be more false positives for one race. It simply means that the odds within one race do not differ.

category. In order to verify that the model optimizes risk distribution between the five risk categories, re-arrests rates were calculated for each risk category (see Table 6). When looking at predictions for any re-arrest, the re-arrest rates appear to be dispersed well across risk categories. The re-arrest rates increase monotonically with each risk classification in the total sample and within each racial subgroup.

On the other hand, predictions for felony re-arrest do not appear to be dispersed well between categories. In the total sample, felony re-arrest rates do increase from 2% in the low risk group to 27% in the high-risk group. However, re-arrest rates were higher in the medium-low risk group than they were in the medium risk group.

Table 6. Re-arrest Rates By Race and Risk Category

Re-arrest Type	Risk Category	Black Hispanic (N=93)	Black Non-Hispanic (N=453)	White Hispanic (N=256)	White Non-Hispanic (N=171)	Total (N=973)	CJA Reported Totals (N=47,370)
Felony	**Total**	**.05**	**.10**	**.07**	**.09**	**.08**	**.07**
	Low	.04	.03	.01	0	**.02**	.03
	Med-Low	.04	.10	.09	.11	**.09**	.06
	Medium	0	.11	.09	.04	**.08**	.09
	Med-High	.11	.14	.04	.21	**.13**	.13
	High	.14	.27	.31	.29	**.27**	.17
Any (Felony or Misdemeanor)	**Total**	**.45**	**.47**	**.63**	**.37**	**.43**	**.18**
	Low	.24	.24	.1	.11	**.18**	.08
	Med-Low	.41	.44	.35	.34	**.39**	.16
	Medium	.44	.49	.61	.40	**.51**	.23
	Med-High	.61	.78	.79	.63	**.74**	.31
	High	1.0	.81	.94	1.0	**.91**	.40

This problem is mimicked in almost all of the racial subgroups as well. While every racial subgroup sees a substantial increase in re-arrest rates from the low risk group to the high risk group, only Black Non-Hispanic defendants show a consistent increase in felony re-arrest

rates with every increase in risk classification. All other racial subgroups have inconsistencies within the medium-low to medium or medium to medium-high risk jumps. While the subgroup inconsistencies may be due to predicting a rare outcome (6.8%) in such a small sample size, the inconsistencies are apparent in the group overall. Since dispersion is similarly poor across subgroups, no tests for significant differences across racial subgroups were calculated.

Even when considering CJA's original findings of well-proportioned re-arrest rates by risk category (reported in Table 6), there is something to be said about the categorical label conversion. "High-risk" has a dangerous connotation that is unlikely to be interpreted as a 17% chance of felony re-arrest. Even defendants labeled "medium-risk" have less than 10% likelihood of felony reoffending. These types of design decisions aimed at translating proportions into action are likely to cause inappropriate discrimination by risk category. A judge may be reluctant to release any defendant labeled high-risk despite the fact that there is a much higher chance of their case being dismissed (41.5%) than them being re-arrested for a felony (8.4%).

Point-Biserial Correlations

To investigate if the strength of the correlation between risk score and felony re-arrest differed significantly by race, Point-Biserial Correlations were calculated (see Table 7). The overall correlation between Risk Score and Re-arrest was rpb = .21. Since the base rate of felony re-arrest is .084 in my sample, the maximum rpb value is .55 . [92] Although the correlation between risk score and felony re-arrest appears stronger for white defendants, the differences were not statistically significant (z = 1.35, p = 0.18).

[92] Mark Gradstein, "Maximal Correlation between Normal and Dichotomous Variables," *Journal of Educational Statistics* 11, no. 4 (1986): 260, https://doi.org/10.2307/1164698.

Table 7. Point Biserial Correlations by Race

Race Correlation Estimate	**Raw Risk Score**	**Numeric Risk Category**
Black	**.18 {.09-.26}**	**.17 {.09-.25}**
Non-Hispanic	*.19 {.1-.28}*	*.18 {.09-.27}*
Hispanic	*.12 {-.08-.32}*	*.13 {-.07-.32}*
White	**.26 {.17-.35}**	**.26 {.17-.35}**
Non-Hispanic	*.32 {.18-.45}*	*.32 {.18-.45}*
Hispanic	*.21 {.09-.32}*	*.21 {.09-.33}*
Total	**.21 {.15-.27}**	**.21 {.15-.27}**

Race as a Moderator

Lastly, in order to test whether race moderates the utility of risk score in predicting felony re-arrest, four logistic regression models were constructed. The first model used race as the sole predictor, the second model used risk score as the sole predictor, the third model used race and risk score, and the fourth model included the interaction between race and risk score.

Table 8. Logistic Regression Models

Model	**BIC**	**AIC**	**Psuedo R-Squared**
Model 1 – Just Race	587.18	567.66	0.005
Model 2 – Just Score	532.73	522.97	0.078
Model 3 – Race & Score	550.37	525.97	0.083
Model 4 – Race * Risk Score	569.86	530.81	0.085

Comparing the BIC and AIC of the models, the model with only risk score as a predictor has the lowest values and is therefore, the preferred model (see Table 8). The coefficient of determination, R-squared, summarizes the proportion of variance in re-arrest that is explained by the predictors. R-squared ranges from 0 to 1 with higher values indicating better models. Here, the McFadden pseudo R-squareds indicate a small increase in the predictive power of each successive model, with the interaction model explaining the largest proportion of variance in re-arrest. However, all our pseudo R-squared values are very low and indicate that there is almost

no relationship, at least not linearly, between the variables in any of the models. Additionally, predictive power does not necessarily indicate that the model is a good fit. Likelihood ratio tests between models two and three (df=3, p = .39), and separately, between models two and four (df=6, p = .66), were also not statistically significant, indicating that race does not improve the model's fit.

Conclusion on Predictive Validity

Taken together, results are consistent with my hypothesis that the CJA tool demonstrates sufficient predictive validity as a whole and across racial strata. Although adequate standards of performance are subjective and should be considered contextually, the model surpasses general statistical standards. Even though the model performed slightly better for white individuals than black individuals across the board, these differences did not reach statistical significance.

Mean Score Differences

A model can demonstrate subgroup predictive validity and still produce disparate outcomes for a subgroup if it consistently scores them as riskier than others. The CJA model was trained using data from 2009 during the height of New York City's stop and frisk era, which disproportionately targeted people of color. Therefore, since the tool predicts re-arrest outcomes based on prior criminal history, I hypothesize that black defendants will receive significantly higher scores on average than white defendants.

ANOVA

First, average risk scores were calculated by race and tested for significant differences using ANOVA and post-hoc pairwise t-tests (see Table 9). Black defendants on average received risk scores of -4.94, which is significantly higher than white defendants who received average scores of -5.78 (p = .022). Although the effect size of this relationship was relatively

small (*d*=.13), it cannot be disregarded in such a context where it demands the removal of an individual from society. [93] When paired with ethnicity, significant differences were found between White Hispanic defendants and both Black Hispanic defendants and Black Non-Hispanic defendants (p =.018, *d*= .26 & p = .012, *d*= .18).[94] There were not statistically significant differences between White Non-Hispanics and Black Hispanic or Black Non-Hispanic individuals.

Table 9. Average Risk Score by Race

	Raw Risk Score		**Risk Category (1 to 5)**	
Race	**Mean Risk Score**	**Standard Deviation**	**Mean Risk Category**	**Standard Deviation**
Black	-4.94	6.29	2.46	1.28
Black Hispanic	*-4.53*	*6.47*	2.54	*1.27*
Black Non-Hispanic	*-5.03*	*6.25*	*2.44*	*1.28*
White	-5.78	6.86	2.30	1.32
White Hispanic	*-6.16*	*6.29*	*2.23*	1.22
White Non-Hispanic	*-5.21*	*7.61*	*2.39*	*1.46*
Total	-5.31	6.55	2.39	1.30

Pearson's Chi-Squared

Next, I checked to see if racial differences still existed when the raw risk scores were combined into risk categories. Pearson's Chi-squared test was used to test the null hypothesis that risk category and race are independent. I find that risk category and race are not independent, meaning the probability distribution of one variable is affected by the other (Chi-Squared = 30.94, p=.002, df=12).

[93] Jacob Cohen, "Statistical Power Analysis," *Current Directions in Psychological Science* 1, no. 3 (June 1, 1992): 98, https://doi.org/10.1111/1467-8721.ep10768783.

[94]These difference between White Hispanic and Black Hispanic individuals was no longer statistically significant when using a Bonferonni adjustment (p=.11)

CJA Mean Differences

Mean differences by race were also tested using the summary statistics reported by the CJA developers. CJA reported the number of defendants per risk category by race, allowing me to create an ordinal risk category variable. Raw scores were not reported. Since their dataset was much larger (N=81734), we are able to see larger differences than can be tested with my sample. It also offers insight into the differences initially known and built into the model.

ANOVA results indicate that race significantly differs by risk category (p = <2e-16, df=3). Post-hoc pairwise t-tests showed that these differences are statistically significant between every racial pairing at the .001 level.[95] Pearson's Chi-Squared test also solidified that risk category and race are not independent (Chi-Squared= 2240.3, df = 12, p-value < 2.2e-16).

Table 10. CJA Average Risk Category by Race

Race	**Mean Risk Score**	**Standard Deviation**
Black	**2.67**	**1.28**
Black Hispanic	*2.59*	*6.47*
Black Non-Hispanic	*2.69*	*6.25*
White	**2.30**	**1.24**
White Hispanic	*2.39*	*6.29*
White Non-Hispanic	*2.13*	*7.61*
Total	**2.53**	**1.28**

Collectively, results indicate that black defendants receive higher risk scores on average compared to white defendants. Although the effect sizes were quite small in my dataset, the CJA dataset confirms the differences with a much larger sample and shows small-medium effect sizes. While overall differences were consistent between the two datasets, they varied when race

[95]Results were all still significant at the .001 level when using a Bonferonni adjustment, Holm adjustment, and Tukey adjustment. Cohen's d shows effect size of .3 for Black v. White, .44 for Black Non-Hispanic v White Non-Hispanic, .16 for Black Hispanic v.White Hispanic.

and ethnicity were combined. In my dataset, Black Hispanics received the highest scores and White Hispanics received the lowest, whereas Black Non-Hispanics received the highest scores in the CJA dataset and White Non-Hispanics received the lowest.

Taken together, tests for predictive validity and mean score differences by subgroup indicate that the CJA tool is similar to the COMPAS instrument in that it predicts accurately yet disproportionately across races.[96] This alludes to the possibility that the model may have a disparate impact by systematically rating black defendants as riskier than white defendants. In order to investigate this further, I retrained the model using my dataset to determine if it would exacerbate the racial bias.

Model Retraining

The last set of tests focused on determining what would happen if the model were retrained using my dataset. Since the original model was trained using a dataset of criminal defendants from 2009, and my dataset was sampled from the same population in 2015, this is emblematic of typical retraining protocol. However, since I did not have access to the data originally used to train the model, my approach here was limited. I focused on determining if there were any changes in the weights of the predictor variables by reconstructing the logistic model on my dataset using the same deviation contrasts method as CJA.

Markedly, I found that all but one of the eight predictor variables no longer had significant predictive utility (□□□□□ = .1). Since previous tests of predictive validity had shown the model functioning relatively well, this is surprising. Interestingly, the only predictor

[96] Rachael T. Perrault, Gina M. Vincent, and Laura S. Guy, "Are Risk Assessments Racially Biased?: Field Study of the SAVRY and YLS/CMI in Probation," Psychological Assessment 29, no. 6 (June 2017): 664, https://doi.org/10.1037/pas0000445; Lowenkamp and Skeem, "Risk, Race, & Recidivism," 705.

that was still significantly predictive of felony re-arrest was whether or not the defendant had any prior arrests (= .002). This finding corroborates our assumption that RAIs are simply predicting police-activity, not crime. The fact that none of the prior conviction data was significant also strongly discredits the counterargument that this is because police are identifying the guilty.

To investigate this further, training data was used to construct a logistic model with number of prior arrests as the sole predictor variable. Different tests were used to find the optimal category thresholds and cut-off point so that the model was constrained to the same type of simple calculation as the CJA model. Remarkably, I found that this model achieved similar validity as the CJA model (See Table 11). There were negligible differences when compared across performance metrics. In other words, all of the predictor variables in the CJA tool function more as placeholders while a defendant's arrest history truly drives the prediction.

Table 11. Re-trained Model Stats by Race

Race	Sensitivity (TP/TP+FN)	Specificity (TN/TN+FP)	Positive Predictive Value (TP/TP+FP)	Negative Predictive Value (TN/TN+FN)	Number Needed to Detain (1/PPV)	Number Safely Discharged ((1/1-NPV) – 1)	Diagnostic Odds Ratio (TP x TN/FP x FN)	Accuracy (TP + TN/ TP+FP+TN+FN)	AUC
Total Arrests Model	0.20	0.95	0.24	0.93	4.2	14.15	4.40	0.89	.70
CJA Model	0.23	0.94	0.27	0.93	3.68	13.33	4.97	0.88	.71

To ensure that this was not a fluke, I also constructed models based on the other predictor variables: total prior convictions, total felony convictions, and total misdemeanor convictions. In all cases, scores were assigned by ranking the defendants high to low based on their number of prior arrests/convictions and then predicting the top 6.8% of the sample to be re-arrested. This

equates to betting the base rate on the most frequently arrested/convicted. Results indicated that the model based only on total number of arrests performs significantly better than those based on convictions.

Combined with other studies of models that predict accurately but disproportionately for subgroups, it appears that the cause it that they are predicting a biased outcome. Researchers found that COMPAS algorithm's 137 features were altogether unnecessary, as just as accurate predictions could be achieved with only two predictors- age and prior convictions.[97] It seems like complicated machine learning and additional predictors function to mask where the algorithms true power comes from, which is learning past police activity.

Limitations

Although these findings are significant, this study is not without limitation. First, the final sample (= 1012) was quite small which did not allow for tests across all racial subgroups. This likely would have produced different results since the CJA model did not take into account any subgroup differences in re-arrest rates and minority groups are typically the ones the model understands the least.

The sample size also limits confidence in the model created based on prior arrests. Since the prevalence rate of felony re-arrest is meager, and there was limited data available to split for test and training, there is no assurance that the predictive validity of the model would hold within the general population. A similar assessment of the instrument should be done using a more substantial dataset to ensure that these findings still stand.

---

[97] Elaine Angelino et al., "Learning Certifiably Optimal Rule Lists for Categorical Data," *ArXiv:1704.01701 [Cs, Stat]*, April 6, 2017, http://arxiv.org/abs/1704.01701; Julia Dressel and Hany Farid, "The Accuracy, Fairness, and Limits of Predicting Recidivism," *Science Advances* 4, no. 1 (January 1, 2018): 4, https://doi.org/10.1126/sciadv.aao5580.

Lastly, every defendant was given a two-year period after their case-open date to be re-arrested for a felony. This is longer on average than the length of a case from arraignment to disposition, which is the time-frame in which the instrument is intended to predict. However, since the risk model does not take potential time at risk into account in its decision, this was deemed a satisfactory approach.

Discussion of Results

Collectively, these findings substantiate what has long been suspected about crime prediction software that utilizes police-generated data - it exploits police activity as an inadequate proxy for criminality. The CJA tool is a quintessential example of the outcome variable problem in which models use systematically misrepresentative data to predict outcomes that are similarly biased. While they can bury the discrimination within their models and promote their AUC values to consumers, they do not, by any means, change the status quo.

Many of the proposed controls for discrimination in predictive modeling focus on data manipulation and reformulations of definitions of fairness. However, no amount of data pre-processing or post-processing can overcome a sociological outcome that is inherently unfair. Incremental reductions in discriminatory predictions, without reductions in discriminatory policing, are useless.

Even in cases where the outcome variable is fairly assessed, crime prediction software is still not likely to live up to the propaganda. The idealized version of risk assessment instruments marketed in academic papers and theorems is often not accustomed to the reality in which they are built. The CJA instrument, for example, was arbitrarily ordered to have five risk classification categories because policymakers believed that would be easily interpretable.

Despite the developers vocalizing their reservations about issues with the training data and model design, the powers lie with the policymakers.

The mere fact that the CJA model is predicting re-arrest typifies how crime prediction software is not intended to be revolutionary. Since the CJA model informs decisions about a defendant's eligibility for a diversion program, it would seem that "risk" should be ancillary to identifying defendants most likely to benefit from the program (those unnecessarily incarcerated) and succeed in it (with intervenable risk factors). Yet, the prediction-oriented tool focuses solely on the risk of re-arrest without any indication that this program will improve the outcomes for the defendant. This is not sophistication or science, its manipulation of the system.

After a comprehensive review of the CJA tool, we can determine that any assessment instrument that uses police-generated data, or any other systematically skewed data, will not be able to avoid discrimination.

Conclusion

The motivation for crime prediction software and the critiques of it admit to the same thing- the current criminal justice system is broken. Nevertheless, predictive modeling uses its same outcomes and condemnations to optimize it, rather than dismantle it.

Both critical and supportive articles on crime prediction software include some version of the adage that even imperfect reform can have meaningful improvements. Even while acknowledging all of its shortcomings, instead of scrapping the idea of actuarial justice, there is a relentless optimism that crime can be prevented if only we were better at predicting it. This is likely because the academic research on it is almost entirely produced by the same individuals invested in its creation. However, the political and economic elite in the field of crime prediction software are often far removed from the individuals whose lives it affects. While they argue in

their echo chambers about the merits of various discrimination reduction formulas, the “false positives” are unnecessarily sitting behind bars because their agency was stolen by a system that deemed them to be nothing more than a mathematical equation.

History has demonstrated that we should be weary of reforms that, in their nascence, claim they can only be improved incrementally. Once institutionalized, reforms quickly become policy and lobbying ploys, puppets for manipulation rather than forces of reformation. With bureaucratic blockages that hinder rapid innovation and quick resolution, there is no assurance that discrimination in crime prediction software will be quickly identified and revised. The most serious threat to the democratic judiciary is the indoctrination of an instrument that legitimizes years of oppression and subjugation under the veil of progressiveness.

Actuarial evangelism has only functioned to delay the urgency for substantive reform. Instead of focusing attention on predicting outcomes, resources must shift to preventing them in a way that does not disburden society’s collective responsibility on the individual. An arbitrary mathematical calculation should not be the justification needed to stop over-policing and over-incarcerating minority communities. Less myopic reform efforts that provide alternatives to social control through surveillance and incapacitation must be considered. Once we can stop debating the different forms of actuarial injustice, revolutionary reform can begin.

Bibliography

Advisory Opinion on Proposed Amendments to the Naturalization Provision of the Constitution of Costa Rica, No. OC-4/84 (Inter-American Court of Human Rights January 19, 1984).

Alschuler, Albert. "The Failure of Sentencing Guidelines: A Plea for Less Aggregation." The University of Chicago Law Review 58, no. 3 (1991): 901–51. https://doi.org/10.2307/1599992.

Angelino, Elaine, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. "Learning Certifiably Optimal Rule Lists for Categorical Data." ArXiv:1704.01701 [Cs, Stat], April 6, 2017. http://arxiv.org/abs/1704.01701;

American Civil Liberties Union. "Racial Disparities in Sentencing." Written Submission to the Inter-American Commission on Human Rights. 153rd Session, October 27, 2014. https://www.aclu.org/other/aclu-submission-inter-american-commission-human-rights-racial-disparities-sentencing.

———. "Report: The War on Marijuana in Black and White." Accessed April 14, 2018. https://www.aclu.org/report/report-war-marijuana-black-and-white.

Andrews, D. A., James Bonta, and J. Stephen Wormith. "The Recent Past and Near Future of Risk and/or Need Assessment." *Crime & Delinquency* 52, no. 1 (January 1, 2006): 7–27. https://doi.org/10.1177/0011128705281756.

Andrews, D.A., James Bonta, and R.D. Hoge. "Classification for Effective Rehabilitation: Rediscovering Psychology." *Criminal Justice and Behavior* 17, no. 1 (March 1, 1990): 19–52. https://doi.org/10.1177/0093854890017001004.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Azavea. "HunchLab: Under the Hood," 2015. https://cdn.azavea.com/pdfs/hunchlab/HunchLab-Under-the-Hood.Pdf.

Bantilan, Niels. "Themis-Ml: A Fairness-Aware Machine Learning Interface for End-To-End Discrimination Discovery and Mitigation." *Journal of Technology in Human Services* 36, no. 1 (January 2, 2018): 15–30. https://doi.org/10.1080/15228835.2017.1416512.

Barabas, Chelsea, Karthik Dinakar, Joichi Ito Madars Virza, and Jonathan Zittrain. "Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment." *ArXiv:1712.08238 [Cs, Stat]*, December 21, 2017. http://arxiv.org/abs/1712.08238.

Barocas, Solon; Selbst, Andrew D. "Big Data's Disparate Impact." *California Law Review* 104, no. 671 (2016): 721. https://doi.org/10.15779/Z38BG31.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. “Fairness in Criminal Justice Risk Assessments: The State of the Art.” *ArXiv:1703.09207v2 [Stat.ML]*, 2017. https://crim.sas.upenn.edu/sites/crim.sas.upenn.edu/files/2017-1.0-Berk_FairnessCrimJustRisk.pdf.

Bowman, Frank O. “The Failure of the Federal Sentencing Guidelines: A Structural Analysis.” *Columbia Law Review* 105, no. 4 (2005): 1315–50.

Boyd, Danah, Sarah Brayne, and Alex Rosenblat. “Predictive Policing.” Washington, D.C., 2015. http://www.datacivilrights.org/pubs/2015-1027/Predictive_Policing.pdf.

Brennan, Tim, William Dieterich, and Beate Ehret. “Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System.” *Criminal Justice and Behavior* 36, no. 1 (January 2009): 21–40. https://doi.org/10.1177/0093854808326545.

Calders, Toon, and Sicco Verwer. “Three Naive Bayes Approaches for Discrimination-Free Classification.” *Data Min. Knowl. Discov.* 21, no. 2 (September 2010): 277–292. https://doi.org/10.1007/s10618-010-0190-x.

Chouldechova, Alexandra. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” *Big Data* 5, no. 2 (June 2017): 153–63. https://doi.org/10.1089/big.2016.0047.

Christin, Angèle. “Algorithms in Practice: Comparing Web Journalism and Criminal Justice.” *Big Data & Society* 4, no. 2 (December 2017). https://doi.org/10.1177/20539517.

Chu, Chi Meng, Hui Yu, Yirong Lee, and Gerald Zeng. “The Utility of the YLS/CMI-SV for Assessing Youth Offenders in Singapore.” *Criminal Justice and Behavior* 41, no. 12 (December 2014): 1437–57. https://doi.org/10.1177/0093854814537626.

Citron, Danielle Keats. “Technological Due Process.” *Wash. UL Rev.* 85 (2007): 1249.

Cohen, Jacob. “Statistical Power Analysis.” *Current Directions in Psychological Science* 1, no. 3 (June 1, 1992): 98–101. https://doi.org/10.1111/1467-8721.ep10768783.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. “Algorithmic Decision Making and the Cost of Fairness,” 797–806. ACM Press, 2017. https://doi.org/10.1145/3097983.3098095.

Desmarais, Sarah, and Jay Singh. “Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States,” 2013. csgjusticecenter.org/wp-content/uploads/2014/07/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf.

Devlin, Hannah. “AI Programs Exhibit Racial and Gender Biases, Research Reveals.” *The Guardian*, April 13, 2017. http://www.theguardian.com/technology/2017/apr/13/ai-programs-exhibit-racist-and-sexist-biases-research-reveals.

Dressel, Julia, and Hany Farid. “The Accuracy, Fairness, and Limits of Predicting Recidivism.” *Science Advances* 4, no. 1 (January 1, 2018). https://doi.org/10.1126/sciadv.aao5580.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. “Fairness Through Awareness.” *ArXiv:1104.3913 [Cs]*, April 19, 2011. http://arxiv.org/abs/1104.3913.

Eric L. Loomis v. State of Wisconsin, No. 16–6387 (United States Supreme Court June 26, 2017).

Fass, Tracy L., Kirk Heilbrun, David DeMatteo, and Ralph Fretz. “The LSI-R and the Compas: Validation Data on Two Risk-Needs Tools.” *Criminal Justice and Behavior* 35, no. 9 (September 2008): 1095–1108. https://doi.org/10.1177/0093854808320497.

Fazel, S., J. P. Singh, H. Doll, and M. Grann. “Use of Risk Assessment Instruments to Predict Violence and Antisocial Behaviour in 73 Samples Involving 24 827 People: Systematic Review and Meta-Analysis.” *BMJ* 345, no. 2 (July 24, 2012). https://doi.org/10.1136/bmj.e4692.

Fischman Joshua B., and Schanzenbach Max M. “Racial Disparities Under the Federal Sentencing Guidelines: The Role of Judicial Discretion and Mandatory Minimums.” *Journal of Empirical Legal Studies* 9, no. 4 (November 6, 2012): 729–64. https://doi.org/10.1111/j.1740-1461.2012.01266.x.

Fleminger, S. “Number Needed to Detain.” *British Journal of Psychiatry* 171, no. 03 (September 1997): 287. https://doi.org/10.1192/bjp.171.3.287a.

Friedler, Sorelle A., Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. “A Comparative Study of Fairness-Enhancing Interventions in Machine Learning.” *ArXiv:1802.04422 [Cs, Stat]*, February 12, 2018. http://arxiv.org/abs/1802.04422.

Gershgorn, Dave, and Dave Gershgorn. “Software Used to Predict Crime Can Now Be Scoured for Bias.” *Quartz* (blog), March 22, 2017. https://qz.com/938635/a-predictive-policing-startup-released-all-its-code-so-it-can-be-scoured-for-bias/.

Gradstein, Mark. “Maximal Correlation between Normal and Dichotomous Variables.” *Journal of Educational Statistics* 11, no. 4 (1986): 259–61. https://doi.org/10.2307/1164698.

Hannah-Moffat, Kelly. “Actuarial Sentencing: An ‘Unsettled’ Proposition.” *Justice Quarterly* 30, no. 2 (April 1, 2013): 270–96. https://doi.org/10.1080/07418825.2012.682603.

———. “The Uncertainties of Risk Assessment: Partiality, Transparency, and Just Decisions.” *Federal Sentencing Reporter* 27, no. 4 (April 2015): 244–47. https://doi.org/10.1525/fsr.2015.27.4.244.

Harcourt, Bernard E. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Chicago: University of Chicago Press, 2007.

———. “Risk as a Proxy for Race: The Dangers of Risk Assessment.” *Federal Sentencing Reporter* 27, no. 4 (April 2015): 237–43. https://doi.org/10.1525/fsr.2015.27.4.237.

Hardt, Moritz, Eric Price, and Nathan Srebro. “Equality of Opportunity in Supervised Learning.” *ArXiv:1610.02413 [Cs]*, October 7, 2016. http://arxiv.org/abs/1610.02413.

“How PredPol Works | Predictive Policing.” *PredPol* (blog). Accessed December 8, 2017. http://www.predpol.com/how-predictive-policing-works/.

Jung, Sandy, and Edward P. Rawana. “Risk and Need Assessment of Juvenile Offenders.” *Criminal Justice and Behavior* 26, no. 1 (March 1, 1999): 69–89. https://doi.org/10.1177/0093854899026001004.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. “Human Decisions and Machine Predictions.” *The Quarterly Journal of Economics* 133, no. 1 (February 1, 2018): 237–93. https://doi.org/10.1093/qje/qjx032.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. “Inherent Trade-Offs in the Fair Determination of Risk Scores.” *ArXiv:1609.05807 [Cs, Stat]*, September 19, 2016. http://arxiv.org/abs/1609.05807.

Larson, Jeff, Julie Angwin, Lauren Kirchner, and Surya Mattu. “How We Analyzed the COMPAS Recidivism Algorithm.” *ProPublica*, May 23, 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Levitt, Steven D. “The Relationship Between Crime Reporting and Police: Implications for the Use of Uniform Crime Reports.” *Journal of Quantitative Criminology* 14, no. 1 (1998): 61–81.

Lichtenstein, Alex. “Good Roads and Chain Gangs in the Progressive South: ‘The Negro Convict Is a Slave.’” *The Journal of Southern History* 59, no. 1 (1993): 85–110. https://doi.org/10.2307/2210349.

Lum, Kristian, and William Isaac. “To Predict and Serve?” *Significance* 13, no. 5 (October 2016): 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x.

Marvell, Thomas B. “Sentencing Guidelines and Prison Population Growth.” *The Journal of Criminal Law and Criminology* 85, no. 3 (1995): 696–709. https://doi.org/10.2307/1144046.

Maurutto, Paula, and Kelly Hannah-Moffat. “Assembling Risk and The Restructuring of Penal Control.” *The British Journal of Criminology* 46, no. 3 (2006): 438–54.

Michael D Maltz. “Bridging Gaps in Police Crime Data.” U.S. Department of Justice, n.d. https://www.bjs.gov/content/pub/pdf/bgpcd.pdf.

Mitchell, Ojmarrh, and Michael S. Caudy. “Examining Racial Disparities in Drug Arrests.” *Justice Quarterly* 32, no. 2 (2015): 288.

Monahan, John, Jennifer Skeem, and Christopher Lowenkamp. “Age, Risk Assessment, and Sanctioning: Overestimating the Old, Underestimating the Young.” *Law and Human Behavior* 41, no. 2 (2017): 191.

Netter, Brian. “Using Group Statistics to Sentence Individual Criminals: An Ethical and Statistical Critique of the Virginia Risk Assessment Program.” *Journal of Criminal Law & Criminology; Chicago* 97, no. 3 (Spring 2007): 699–729.

Oleson, James C. “Risk in Sentencing: Constitutionally Suspect Variables and Evidence-Based Sentencing.” *SMUL Rev.* 64 (2011): 1329.

O’Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. First edition. New York: Crown, 2016.

“Patricia Van Voorhis, ‘Classification of Women Offenders: Gender-Responsive Approaches to Risk/Needs Assessment,’ Commun,” n.d.

Perrault, Rachael T., Gina M. Vincent, and Laura S. Guy. “Are Risk Assessments Racially Biased?: Field Study of the SAVRY and YLS/CMI in Probation.” *Psychological Assessment* 29, no. 6 (June 2017): 664–78. https://doi.org/10.1037/pas0000445.

“Predictive Policing Software Is More Accurate at Predicting Policing Than Predicting Crime.” American Civil Liberties Union. Accessed April 17, 2018. https://www.aclu.org/blog/criminal-law-reform/reforming-police-practices/predictive-policing-software-more-accurate.

Reisig, Michael D., Kristy Holtfreter, and Merry Morash. “Assessing Recidivism Risk across Female Pathways to Crime.” *Justice Quarterly* 23, no. 3 (September 1, 2006): 384–405. https://doi.org/10.1080/07418820600869152.

Rights, United Nations Office of the High Commissioner for Human, and International Bar Association. *Human Rights In The Administration Of Justice: A Manual On Human Rights For Judges, Presecutors And Lawyers*. United Nations Publications, 2003.

Rose, Nikolas. “The Biology of Culpability:: Pathological Identity and Crime Control in a Biological Culture.” *Theoretical Criminology* 4, no. 5 (February 1, 2000). https://doi.org/https://doi.org/10.1177/1362480600004001001.

Saunders, Jessica, Priscillia Hunt, and John S. Hollywood. “Predictions Put into Practice: A Quasi-Experimental Evaluation of Chicago’s Predictive Policing Pilot.” *Journal of*

Experimental Criminology 12, no. 3 (September 1, 2016): 347–71. https://doi.org/10.1007/s11292-016-9272-0.

Schwalbe, Craig S., Mark W. Fraser, Steven H. Day, and Valerie Cooley. "Classifying Juvenile Offenders According to Risk of Recidivism: Predictive Validity, Race/Ethnicity, and Gender." *Criminal Justice and Behavior* 33, no. 3 (June 2006): 305–24. https://doi.org/10.1177/0093854806286451.

Simon, Jonathan, and Malcolm Feely. "The New Penology: Notes on the Emerging Strategy of Corrections and Its Implications." *Criminology* 30 (March 7, 2006): 449–74. https://doi.org/10.1111/j.1745-9125.1992.tb01112.x.

Singh, Jay P. "Predictive Validity Performance Indicators in Violence Risk Assessment: A Methodological Primer: Performance Indicator Primer." *Behavioral Sciences & the Law* 31, no. 1 (January 2013): 8–22. https://doi.org/10.1002/bsl.2052.

Skeem Jennifer L., and Lowenkamp Christopher T. "Risk, Race, and Recidivism: Predictive Bias and Disparate Impact*." *Criminology* 54, no. 4 (November 23, 2016): 680–712. https://doi.org/10.1111/1745-9125.12123.

Smith, Jack. "Chicago's New Policing Strategy Is Hurting the People It's Supposed to Be Helping." *MIC*, August 17, 2016. https://mic.com/articles/151782/chicago-s-experimental-policing-tool-is-hurting-the-people-it-s-supposed-to-be-helping.

Starr, Sonja B. "Evidence-Based Sentencing and the Scientific Rationalization of Discrimination." *Stanford Law Review* 66, no. 4 (April 2014): 803–72.

———. "The New Profiling: Why Punishing Based on Poverty and Identity Is Unconstitutional and Wrong." *Federal Sentencing Reporter* 27, no. 4 (April 1, 2015): 229–36. https://doi.org/10.1525/fsr.2015.27.4.229.

Sweeney, Latanya. "Discrimination in Online Ad Delivery." *Queue* 11, no. 3 (March 2013): 10:10–10:29. https://doi.org/10.1145/2460276.2460278.

"UN General Assembly, International Convention on the Elimination of All Forms of Racial Discrimination, 21 December 1965," n.d.

"UN General Assembly, International Covenant on Civil and Political Rights, 16 December 1966, United Nations, Treaty Seri," n.d.

"UN General Assembly, Universal Declaration of Human Rights, 10 December 1948, 217 A (III), Available at: Http://Www.Un.O," n.d.

"UN General Assembly, Convention on the Elimination of All Forms of Discrimination Against Women, 18 December 1979, Unite," n.d.

"UN General Assembly, Convention on the Rights of Persons with Disabilities : Resolution / Adopted by the General Assembl," n.d.

"UN General Assembly, Convention on the Rights of the Child, 20 November 1989, United Nations, Treaty Series, Vol. 1577, ," n.d.

"U.S. Const. Amend. XIX.," n.d.

Van De Vijver, Fons, and Norbert K Tanzer. "Bias and Equivalence in Cross-Cultural Assessment: An Overview." *European Review of Applied Psychology* 54, no. 2 (June 2004): 119–35. https://doi.org/10.1016/j.erap.2003.12.004.

Victor, Daniel. "Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk." *The New York Times*, March 24, 2016, sec. Technology. https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html.

Vincent, Gina M., Laura S. Guy, Rachael T. Perrault, and Bernice Gershenson. "Risk Assessment Matters, but Only When Implemented Well: A Multisite Study in Juvenile Probation." *Law and Human Behavior* 40, no. 6 (2016): 683–96. https://doi.org/10.1037/lhb0000214.

Vincent, Gina M., Melissa L. Paiva-Salisbury, Nathan E. Cook, Laura S. Guy, and Rachael T. Perrault. "Impact of Risk/Needs Assessment on Juvenile Probation Officers' Decision Making: Importance of Implementation." *Psychology, Public Policy, and Law* 18, no. 4 (2012): 549–76. https://doi.org/10.1037/a0027186.

Wall, David. "From Post-Crime to Pre-Crime: Preventing Tomorrow's Crimes Today." *Criminal Justice Matters* 81, no. 1 (September 1, 2010): 22–23. https://doi.org/10.1080/09627251.2010.505396.

Webster, Cheryl Marie, and Anthony N. Doob. "Classification without Validity or Equity: An Empirical Examination of the Custody Rating Scale for Federally Sentenced Women Offenders in Canada." *Canadian Journal of Criminology and Criminal Justice* 46 (2004): 395. https://doi.org/https://doi.org/10.3138/cjccj.46.4.395.

Wexler, Rebecca. "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System." Stanford Law Review 70 (February 21, 2017). https://papers.ssrn.com/abstract=2920883.v

Wexler, Rebecca. "When a Computer Program Keeps You in Jail." *The New York Times*, June 13, 2017, sec. Opinion. https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html.

Whiteacre, Kevin W. "Testing the Level of Service Inventory–Revised (LSI-R) for Racial/Ethnic Bias." *Criminal Justice Policy Review* 17, no. 3 (September 2006): 330–42. https://doi.org/10.1177/0887403405284766.

Wilson, David B., Tammy Rinehart Kochel, and Stephen D. Mastrofski. “Race and the Likelihood of Arrest.” In *Encyclopedia of Criminology and Criminal Justice*, edited by Gerben Bruinsma and David Weisburd, 4245–51. Springer New York, 2014. https://doi.org/10.1007/978-1-4614-5690-2_245.

Zaykowski, Heather. “Racial Disparities in Hate Crime Reporting.” *Violence and Victims* 25, no. 3 (June 1, 2010): 378–94. https://doi.org/10.1891/0886-6708.25.3.378.

Zinger, Ivan. “Actuarial Risk Assessment and Human Rights: A Commentary.” *Canadian Journal of Criminology and Criminal Justice* 46 (2004): 607.

Appendix A: Tables and Figures

Table 1: CJA Predictor Variables and Risk Points

Variable	Score	Risk Points
Age	16 to 19	6
	20 to 29	1
	30 to 39	-3
	40+	-4
Open Cases	No	-1
	Yes	1
First Arrest	No	3
	Yes	-3
Fulltime Activity	No	2
	Yes	-2
Warrant in last 4 years	No	-1
	Yes	1
Misdemeanor conviction in last 4 years	No	-2
	Yes	2
Felony conviction in last 9 years	No	-1
	Yes	1
Drug conviction in last 9 years	No	-2
	Yes	2

Risk Category	Total Points
Low	-16 to -10
Medium Low	-9 to -5
Medium	-4 to 0
Medium High	1 to 4
High	5 to 18

Source: Eion Healy (2015). Research Report for MOCJ's Pretrial Felony Re-Arrest Risk Assessment Tool. New York, NY: New York City Criminal Justice Agency. [unpublished]

Table 2: Data Sources

Variables	Source
Age	Legal Aid Society Case Management System
Race	
Gender	
Prior Misdemeanor Convictions	Data was extracted from Client's Rap Sheet. If no rap sheet data was not available, variables were inferred based on defendant's criminal history in the OCA database
Prior Felony Convictions	
Prior Drug Convictions	
Prior Warrants	Rap Sheet
Full-Time Activity	CJA Form. If no CJA form was available, LAS database was checked for employment. If employment data available, activity marked as affirmative. If none available, left Null.
Re-arrest	NYSID was run through the OCA database to check for any re-arrests 2 years following case open date.

Figure 7. Risk Category Distribution by Race