ALGORITHMIC DISCRIMINATION IN THE U.S. JUSTICE SYSTEM: A QUANTITATIVE ASSESSMENT OF RACIAL AND GENDER BIAS ENCODED IN THE DATA ANALYTICS MODEL OF THE CORRECTIONAL OFFENDER MANAGEMENT PROFILING FOR ALTERNATIVE SANCTIONS (COMPAS)

by
Yubin Li

A capstone paper submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science in Government Analytics

Baltimore, Maryland
April, 2017

ABSTRACT

The fourth-generation risk-need assessment instruments such as Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) have opened the opportunities for the use of big data analytics to assist judicial decision-making across the criminal justice system in U.S. While the COMPAS system becomes increasingly popular in supporting correctional professionals' judgement on an offender's risk of committing future crime, little research has been published to investigate the potential systematic bias encoded in the algorithms behind these assessment tools that could possibly work against certain ethnic or gender groups. This paper uses two-sample t-test and ordinary least-square regression model to demonstrate that COMPAS algorithms systemically generates a higher risk score for African-American and male offenders in terms of the risk of failure to appear, risk of recidivism, and risk of violence. Although race was explicitly excluded when the COMPAS algorithms were developed, the results showed that such an analytic model still systematically discriminates against African-American offenders. This paper introduced the importance of examining algorithmic fairness in big data analytic applications and offers the methodology as well as tools to investigate systematic bias encoded in machine leaning algorithms. Additionally, the implications of this paper also suggest that simply removing the protected variable in a big data algorithm could not be sufficient to eliminate the systematic bias that can still affect the protected groups, and that further research is needed for solutions to thoroughly address the algorithmic bias in big data analytics.

Key words: Big Data, Algorithm, Justice, Discrimination, COMPAS

TABLE OF CONTENTS

INTRODUCTION

Big data analytics are becoming increasingly popular as a quantitative tool in both public and private sector. Centers for Medicare and Medicaid Services (CMS) has developed a predictive analytics system for Medicare fraud prevention, Amazon relies on algorithms to prioritize the targeted geographies to expand its Prime services, law enforcement departments are testing predictive policing to fight against crimes, along with many others. While data analytics open new opportunities in a variety of domains, research and government reports have discussed the possibility of systematic discrimination encoded in big data algorithms that could reinforce the inequality in certain social groups, raising a critical concern regarding the reliance of using big data algorithms for our decision making without fully vetting the algorithmic fairness.

In particular, correctional professionals and judiciary officers in U.S. justice system are increasingly utilizing the fourth-generation risk-need assessment instruments that potentially are the best risk assessment instruments currently available for criminal justice since these tools not only incorporate multi-theoretical criminal risk factors, but also are designed to be integrated into the "selection of intervention modes and targets for treatment.[1] Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is one of the most popular assessment tools used in nationwide criminal justice system[2], producing quantitative evaluations towards the offenders based on the

[1] Blomberg, Thomas, William Bales, Karen Mann, Ryan Meldrum, and Joe Nedelec. "Validation of the COMPAS Risk Assessment Classification Instrument." College of Criminology and Criminal Justice, Florida State University, Tallahassee, FL (2010)

[2] Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. "How We Analyzed the COMPAS Recidivism Algorithm." ProPublica (5 2016).

offender’s characteristics as an important component of judicial supervision and intervention. Using empirically based evidence and big data analytics algorithms, these risk assessment tools predict each offender’s likelihood of committing wrong-doings in the future, aiming to provide an “objective” evaluation on which the judiciary officers can rely for sentencing, deciding probation, and other decision-making. There is no doubt that judicial fairness is one of the core principals in the U.S. justice system. With the new strategy of using algorithm-based assessment tools, there is a need to investigate the fairness at these predictive risk scores among different social groups in order to ensure the assessments that are produced by algorithms are based on the offender’s criminal factors instead of his/her race or gender.

This paper examined the risk scores for 13,186 offenders in Broward County Florida, who were assessed by the COMPAS system between 2013 and 2014. The results of the statistical analyses demonstrate that the COMPAS algorithm systematically predicts a higher risk score towards African-American offenders and male offenders. On average, the system is expected to produce at least 0.20-point higher COMPAS score for African-American offenders as well as for male offenders in terms of risk of failure to appear. In addition, an African-American is expected to receive an approximately 1.0-point higher score from the COMPAS assessment in terms of risk of recidivism and risk of violence; while the COMPAS scores of male offenders are statistically significantly higher than the risk scores of female offenders by 0.2-point higher in terms of risk of recidivism and by 0.7-point in terms of violence.

Although the differences in risk scores among the two ethnic and gender groups are not substantially significant but rather statistically significant, the implication of this paper raise an important concern of the potential algorithmic discrimination encoded in the COMPAS data analytics model.

LITERATURE REVIEW

Existing academic studies have revealed the possibility that systematic discrimination against certain protected groups could be encoded in data analytics algorithms. While acknowledging that machine learning, a common use of big data analytics, could improve predictions in domains such as employment, education, and even criminal justice, Hardt et al. pointed out that "its effect on existing biases is not well understood."[3] In addition, the Big Data Working Group in Obama's Administration in May 2014 released a report noting that algorithmic bias can sometimes even "be the inadvertent outcome of the way big data technologies are structured and used" due to the "encoding discrimination in automated decisions.[4]" Furthermore, using case studies in a variety of practices of big data analytics, Obama's Administration published a report in May 2016 that has explicitly acknowledged the challenges raised by algorithmic systems that can "perpetuate, exacerbate, or mask harmful discrimination."[5]

DISCRIMINATION ENCODED IN BIG DATA ALGORITHMS

In a recent statement, the Association for Computing Machinery argued that automated decision–making, powered by big data algorithms, can cause "harmful

[3] Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of Opportunity in Supervised Learning." In Advances in Neural Information Processing Systems, pp. 3315-3323. 2016.

[4] United States. Executive Office of the President, and John Podesta. Big Data: Seizing Opportunities, Preserving Values. 2014.

[5] Munoz, Cecilia, Megan Smith, and D. J. Patil. "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights." Executive Office of the President (2016).

discrimination" on disadvantaged individuals[6]. For example, researchers found that target ads that are produced by big data algorithms would discriminate against low-income consumers[7]. These individuals that could otherwise eligible for better offers (such as lower interest for personal loans) may never receive the information as the result of the algorithmic selection for the ad targeting.

Existing literature has also discussed reasons that could potentially cause algorithmic bias in the practice of big data analytics, even sometimes without any human errors. One argument maintains that the learning algorithms can capture the stereotypes and biases from input data and generate algorithmically-biased outcomes[8]. Barocas and Selbst also noted that big data analytics "can reproduce existing patterns of discrimination" and reinforce the "existing inequalities by suggesting that historically disadvantaged groups actually deserve less favorable treatment.[9]" In addition, since algorithms are defined by humans, they could inadvertently inhere the human biases that are incorporated at the programming of the algorithms[10]. For instance, a study on statistical discrimination in labor economics found that employers would have incentives to "easily use observable characteristic such as sex and race" as the proxy to predict the productivity of the workers based on the gender and racial group that they belong to, and therefore workers of the discriminatory group who are as equally productive as workers

[6] ACM U.S. Public Policy Council. "Statement on Algorithmic Transparency and Accountability." January 2017. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

[7] Federal Trade Commission. "Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues." January 2016

[8] Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. "Quantifying and Reducing Stereotypes in Word Embedding." arXiv Preprint arXiv:1606.06121 (2016)

[9] Barocas, Solon, and Selbst, Andrew D.. "Big Data's Disparate Impact." (2016).

[10] Kirkpatrick, Keith. "Battling Algorithmic Bias: How Do We Ensure Algorithms Treat Us Fairly?." Communications of the ACM 59, no. 10 (2016): 16-17.

of the non-discriminatory group can experience differences in wage.[11] Moreover, even without absorbing the biases from the input data or the human developers, due to the biases of omission, big data analytics could still produce biased classifications and decisions "because the data is implicitly biased by virtue of who is represented and who is omitted."[12]

Empirical studies have discovered a variety of algorithmic biases in the real-world applications of big data analytics in both public and private sector. In a study from a Harvard researcher, she found that Google's online advertising algorithm would be 25% more likely to deliver an ad suggestive of an arrest record when the input queries are identified as the names associated with African-Americans[13]. In addition, a study in 2015 also discovered that the online advertising platform in Google systematically displayed fewer target ads for high paying jobs to female users than it did to male users[14]. In 2016, Bolukasi et al conducted a research of an analogy puzzle[15] that analyzed a dataset of 3 million words trained on a corpus of text from Google News, and concluded that the algorithm returns "*ASSAULTED*" as the closest word to the query "*BLACK MALE*" while the responses to "*WHITE MALE*" is "*ENTITLED TO*".[16] These studies produced evidence

[11] Romei, Andrea, and Salvatore Ruggieri. "A Multidisciplinary Survey on Discrimination Analysis." The Knowledge Engineering Review 29, no. 05 (2014): 582-638.

[12] Lipton, Zachary. "The Foundations of Algorithmic Bias." (2016)

[13] Sweeney, Latanya. "Discrimination in Online Ad Delivery." Queue 11, no. 3 (2013): 10.

[14] Datta, Amit, Michael Carl Tschantz, and Anupam Datta. "Automated Experiments on Ad Privacy Settings." Proceedings on Privacy Enhancing Technologies 2015, no. 1 (2015): 92-112.

[15] An analogy puzzle (in the format of a:b :: c:d) is a data analytic model that selects the most appropriate d (which is the dependent variable) given the a, b, and c (the independent variables).

[16] Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. "Quantifying and Reducing Stereotypes in Word Embeddings." arXiv Preprint arXiv:1606.06121 (2016).

of the notion that data mining can reinforce the biases that inherits the prejudice of the existing data[17].

In recent years, scholars refined the definition of non-discrimination (or fairness) in the context of big data analytics. One aspect of the algorithmic fairness is introduced by Dwork et al as the notion of "individual-based fairness," arguing that similar individuals in terms of the non-protected characteristics should be treated similarly by the big data analytics models[18]. In addition, Zliobaite points out the other aspect of the fairness for machine-learning is to avoid "redlining", which refers the different predictions among different groups of individuals "can only be as large as justified by the non-protected characteristics." [19] These two aspects cover the conditions of algorithmic fairness from individual level to the group level.

ALGORITHMIC FAIRNESS IN U.S. JUSTICE SYSTEM

Data analytics have also been widely used in support of the decision making in U.S. judiciary system in the past two decades. Scholars have acknowledged the increasing importance as well as popularity of using actuarial, objective, risk-need assessments in the field of criminal justice.[20] Using data modeling and empirical risk factors, these tools are developed and designed to predict the offender's likelihood of recidivism, thereby assisting judicial agencies to decide the level of sentencing

[17] Barocas, Solon, and Selbst, Andrew D.. "Big Data's Disparate Impact." (2016).

[18] Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through Awareness." In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214-226. ACM, 2012.

[19] Zliobaite, Indre. "A Survey on Measuring Indirect Discrimination in Machine Learning." arXiv Preprint arXiv:1511.00148 (2015).

[20] Holsinger, Alexander M., Christopher T. Lowenkamp, and Edward J. Latessa. "Ethnicity, Gender, and the Level of Service Inventory-Revised." Journal of Criminal Justice 31, no. 4 (2003): 309-320.

accordingly[21]. Over the past few years, the empirically based risk assessment instruments such as Level of Service Inventory-Revised (LSI-R) model and the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) model are becoming “an integral part” of the judicial system in U.S.[22]

LSI-R is a third-generation assessment model that evaluates the characteristics of offenders including: criminal history, education/employment, finances, marital status, accommodations, recreation preferences, social companions, alcohol/drug, emotional/personal status, and attitude orientation.[23] With these static and dynamic factors, LSI-R provides an assessment regarding the offender’s likelihood of re-offending a crime in the future in order to assist the correctional officers in making decision such as sentencing, levels of supervision, and release from institutional custody.[24]

Although race is explicitly excluded as an input in the LSI-R model, a variety of scholars have conducted studies to investigate the ethnic bias encoded in the LSI-R predictive classification on the offenders’ “risk scores” of recidivism. Whiteacre’s study in 2006 found that the LSI-R classification system has a tendency towards more errors of over-classification (false positives) for African American offenders than either Caucasians or Hispanics offenders.[25] Holsinger et al also discovered that the LSI-R is

[21] Flores, Anthony W., Christopher T. Lowenkamp, Alexander M. Holsinger, and Edward J. Latessa. "Predicting Outcome with the Level of Service Inventory-Revised: The Importance of Implementation Integrity." Journal of Criminal Justice 34, no. 5 (2006): 523-529.

[22] 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 (2008).

[23] Andrews, D. A., & Bonta, J. “Level of Service Inventory–Revised (LSI-R): User’s Manual.” North Tonawanda, NY: Multi-Health Systems (2001).

[24] Kroner, Daryl G., and Jeremy F. Mills. "The Accuracy of Five Risk Appraisal Instruments in Predicting Institutional Misconduct and New Convictions." Criminal Justice and Behavior 28, no. 4 (2001): 471-489.

[25] Whiteacre, Kevin W. "Testing the Level of Service Inventory–Revised (LSI-R) for Racial/Ethnic Bias." Criminal Justice Policy Review 17, no. 3 (2006): 330-342.

likely to predict "significantly higher scores" to Native Americans (non-White individuals) than to Non-Native Americans on the likelihood of recidivism.[26]

COMPAS, developed by Northpointe Inc. (a private company now called Equivant), is one of the "best known" fourth generation systems[27] that predict offenders' likelihood of re-offending to assist judicial decision making in U.S. Instead of simply providing one risk score toward each offender (like the conventional risk assessment systems), COMPAS generates separate risk predictions, ranging from 1 (very low risk) to 10 (very high risk), in terms of violence, recidivism, failure to appear, and community failure.[28] Several published data and literature have discussed the predictive validity of the COMPAS assessment tool. Northpointe Inc. has conducted a few internal studies to illustrate the validity of the COMPAS, such as the report by Brennan et al in 2008 arguing the predictions produced by COMPAS recidivism risk model are "equal [to] or exceed similar" fourth generation judicial assessment instruments.[29] However, a study in University of California, in contrast, found little evidence on "interrater reliability, predictive utility, and construct validity" in the COMPAS's prediction on recidivism.[30]

[26] Holsinger, Alexander M., Christopher T. Lowenkamp, and Edward J. Latessa. "Ethnicity, Gender, and the Level of Service Inventory-Revised." Journal of Criminal Justice 31, no. 4 (2003): 309-320.

[27] Andrews, Don A., James Bonta, and J. Stephen Wormith. "The Recent Past and Near Future of Risk and/or Need Assessment." Crime & Delinquency 52, no. 1 (2006): 7-27.

[28] 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 (2008).

[29] 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 (2009): 21-40.

[30] Skeem, J., and J. Eno Louden. "Assessment of Evidence on the Quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)." Unpublished Report Prepared for the California Department of Corrections and Rehabilitation. Available at: https://webfiles. uci. edu/skeem/Downloads. html (2007).

Furthermore, less research has been conducted to investigate the ethnic and/or gender bias encoded in the COMPAS analytics model. In 2016, a study by ProPublica analyzed the predictive risk scores generated by the COMPAS model for 7,000 offenders in Florida, and found that African-Americans who were "labeled as higher risk" (risk score of 8-10) but did not actually reoffend are twice as likely as White individuals[31]. Northpointe Inc., subsequently published their own study[32] to counter ProPublica's conclusions on the racial bias encoded in COMPAS and criticize the statistical methodology that Angwin et al utilized at ProPublica's report.

PURPOSE OF THIS STUDY

In the context of the judicial system, algorithm bias from data analytics would present a statistical inference that is misleading and can eventually hinder the fair judgement in criminal sentencing[33] against the disadvantaged groups. Moreover, although the fourth-generation risk assessment instruments (like COMPAS) are increasingly common in courtrooms across the nation[34], a limited amount of research has been published that examines any systematic bias against minority groups in the predictions generated by these big data analytic algorithms. The main goal of this study is to investigate whether there is any systematic discrimination in the risk scores produced by the COMPAS model towards African Americans and/or female offenders.

[31] Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." Pro Publica (2016).

[32] Dieterich, William, Christina Mendoza, and Tim Brennan. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Technical report, Northpointe, July 2016. http://www. northpointeinc. com/northpointe-analysis, 2016.

[33] Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine bias." Pro Publica (2016).

[34] Dieterich, William, Christina Mendoza, and Tim Brennan. COMPAS risk scales: Demonstrating Accuracy Equity and Predictive Parity. Technical report, Northpointe, July 2016. http://www. northpointeinc. com/northpointe-analysis, 2016.

DATA AND METHODS

This paper analyzes the original dataset that Larson et al received via the Freedom of Information Act (FOIA) from the Broward County Sheriff's Office in Florida and used at their study[35] at ProPublica on the algorithmic bias in the COMPAS system. The dataset consists of 18,610 individual offenders in Broward County, Florida, who were assessed by the COMPAS system in 2013 and 2014. According to Larson et al, Broward County is "a large jurisdiction using the COMPAS tool" in its criminal justice system and has a strong open-records law, it is reasonable to serve as the data sample geography for the research topic at this paper.

Based on the dataset, the COMPAS algorithm produced three risk scores for each defendant in terms of "Risk of Recidivism", "Risk of Violence" and "Risk of Failure to Appear." The COMPAS scores for each individual, which is the dependent variable of this study, are from 1 (the lowest risk) to 10 (the highest risk). In addition, the dataset also includes the independent variables of interest such as the defendant's ethnic and gender as well as the potential controlling variables such as the offender's age when the screening was performed, marital status, and legal status.

Among the 18,610 offenders at the raw dataset, 2,373 (or 12.8%) of them are classified as the races (Hispanic, Asian, etc.) other than African-Americans or Caucasians. In addition, after removing the individuals who either have an invalid COMPAS score or have a marital status classified as "other" in the dataset, a total of 13,186 offenders remain as a valid data sample for the statistical analyses.

[35] Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. "How We Analyzed the COMPAS Recidivism Algorithm." ProPublica (5 2016) (2016).

Zliobaite[36] provided a theoretical framework with regard to testing discrimination in big data predictions using statistical tests, which refers to the hypotheses testing models such as the OLS regression and t-tests to measure whether there is a statistically significant difference of the means in different groups. Thus, in order to test and measure the ethnic and gender bias encoded in the COMPAS algorithms, a variety of statistical tests have been conducted at this paper. First, t-tests will be performed to investigate whether there is any statistically significant relationship between the means of the COMPAS scores by risk category (risk of recidivism, risk of violence, risk of failure to appear) and race/gender. Specifically, the following hypotheses will be tested at the t-tests:

H_{01} = There is no difference between the means of COMPAS scores in African American defendants and Caucasian defendants
H_{A1} = There is difference between the means of COMPAS scores in African American defendants and Caucasian defendants

H_{02} = There is no difference between the means of COMPAS scores in Female defendants and Male defendants
H_{A2} = There is difference between the means of COMPAS scores in Female defendants and Male defendants

In addition, several ordinary least squares (OLS) regression models with various control variables will be used in the second phase to further test the relationship and measure the magnitude of the ethnic/gender bias (if any) in COMPAS scores by risk category. Specifically, the OLS regression tests will investigate the following hypotheses:

[36] Zliobaite, Indre. "A Survey on Measuring Indirect Discrimination in Machine Learning." *arXiv preprint arXiv:1511.00148* (2015).

H_{03} = There is no relationship between the COMPAS scores and the defendant's race (African American versus Caucasian)

H_{A3} = There is a relationship between the COMPAS scores and the defendant's race (African American versus Caucasian)

H_{04} = There is no relationship between the COMPAS scores and the defendant's gender (male versus female)

H_{A5} = There is a relationship between the COMPAS scores and the defendant's gender (male versus female)

Furthermore, this paper will use the raw decile COMPAS scores (ranged from 1 to 10) as the dependent variable instead of the risk classification level (low, medium, high) for all the statistical analyses since the categorical risk classification cannot accurately reflect the scale of differences in COMPAS scores. For example, offenders with a score of 1 and offenders with a score of 4 would both be classified as "low risk" while the difference of their COMPAS scores were 3 points.

Additionally, the predictive validity of COMPAS is not the primary concentration of this study. Hence, the statistical analyses at this paper will not investigate the accuracy of the COMPAS scores in comparison to the offender's actual record of re-arrest by race or gender. Instead, the focus of this paper is examining the fairness of the risk scores directly generated by the COMPAS algorithms between different ethnic groups and gender groups. Based on Zliobaite's definition of discrimination in machine learning[37], if COMPAS produces fair predictions, offenders that are similar in terms of the "non-protected characteristics" (all the risk factors expect race or gender) should receive similar predictive risk scores.

[37] Zliobaite, Indre. "A Survey on Measuring Indirect Discrimination in Machine Learning." *arXiv preprint arXiv:1511.00148* (2015).

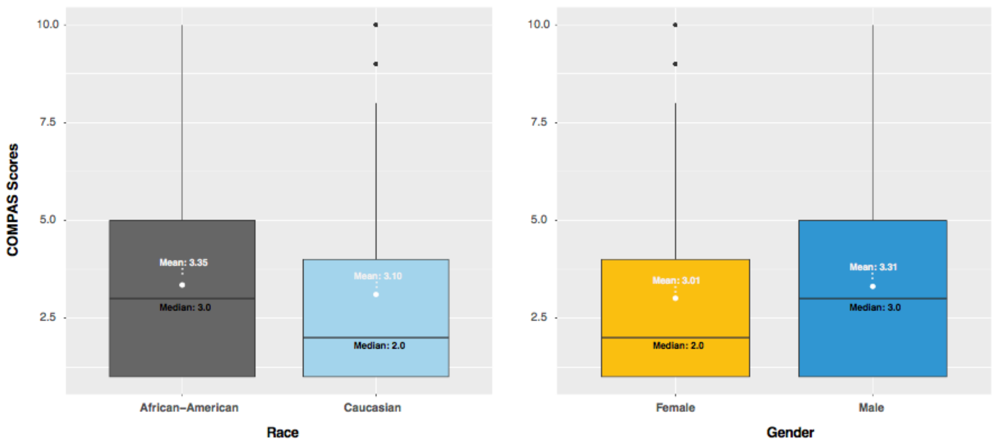
RESULTS

In general, this paper found that, although COMPAS algorithms have explicitly excluded "race" from the questionnaire that collects the offender's data as the input of its modeling, this assessment tool still generates discriminatively higher risk scores for African-American defendants in terms of risk of failure to appear, risk of recidivism, and risk of violence. Additionally, statistical analyses also indicated that difference of the COMPAS scores between male and female offenders cross those three types of risk category is also statistically significant.

RISK OF FAILURE TO APPEAR

Figure 1 presents the six-number summary (minimum, first quartile, median, mean, third quartile, maximum) of COMPAS scores in terms of the offender's risk for failure to appear by race (African-American versus Caucasian) and by gender (female versus male). The mean and median of COMPAS scores of African-American offenders (respectively 3.35 and 3.0) are higher than Caucasian offenders (mean: 3.10; median 2.0). In addition, the mean and median of the male offender's COMPAS scores (respectively 3.31 and 3.0) are also both higher than that of female offender's scores.

Figure 1: COMPAS Scores (Risk of Failure to Appear) Boxplots by Race & Gender



The results in t-tests (Table 1) show that the average COMPAS score of African-American individuals is statistically significantly higher than the score of Caucasian individuals in terms of the offender's risk of failure to appear. Specifically, it's at 95% confidence level that, on average, African-Americans would have a higher COMPAS score than Caucasians by 0.16 to 0.32 points. In addition, the difference between male offender's risk of failure to appear score and the female offender's is also statistically significant at the 99% confidence level.

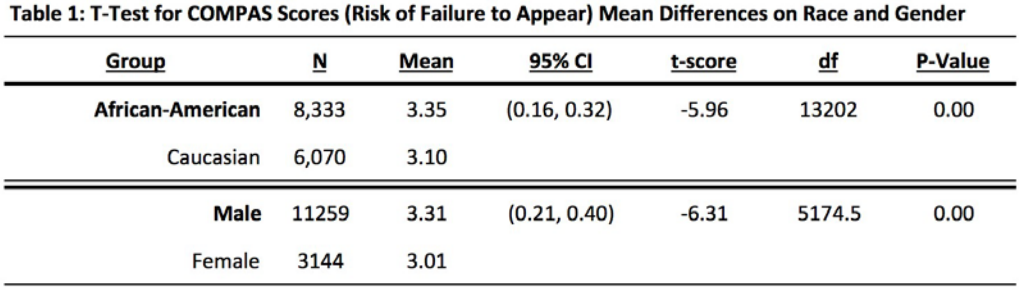


Table 1: T-Test for COMPAS Scores (Risk of Failure to Appear) Mean Differences on Race and Gender

Group	N	Mean	95% CI	t-score	df	P-Value
African-American	8,333	3.35	(0.16, 0.32)	-5.96	13202	0.00
Caucasian	6,070	3.10				
Male	11259	3.31	(0.21, 0.40)	-6.31	5174.5	0.00
Female	3144	3.01				

The results of the OLS regression models, presented in Table 2, indicate that being an African-American would generally results in a 0.24-point higher risk of failure to appear score than being a Caucasian. Moreover, the COMPAS algorithm systematically produces a higher risk score to male offenders than it does for female offenders by 0.3-point in terms of risk of failure to appear. Furthermore, holding the individual's marital status and age constant, the results from the regression model show that the COMPAS tool, on average, generates a 0.4-point higher risk of failure to appear score to African-Americans and a 0.26-point higher score to male offenders in comparison to the risk scores towards Caucasians and female offenders, respectively, at greater than 99% confidence level. However, the r-squared value in the multivariate regression model is only 0.07, indicating the model is not fully specified.

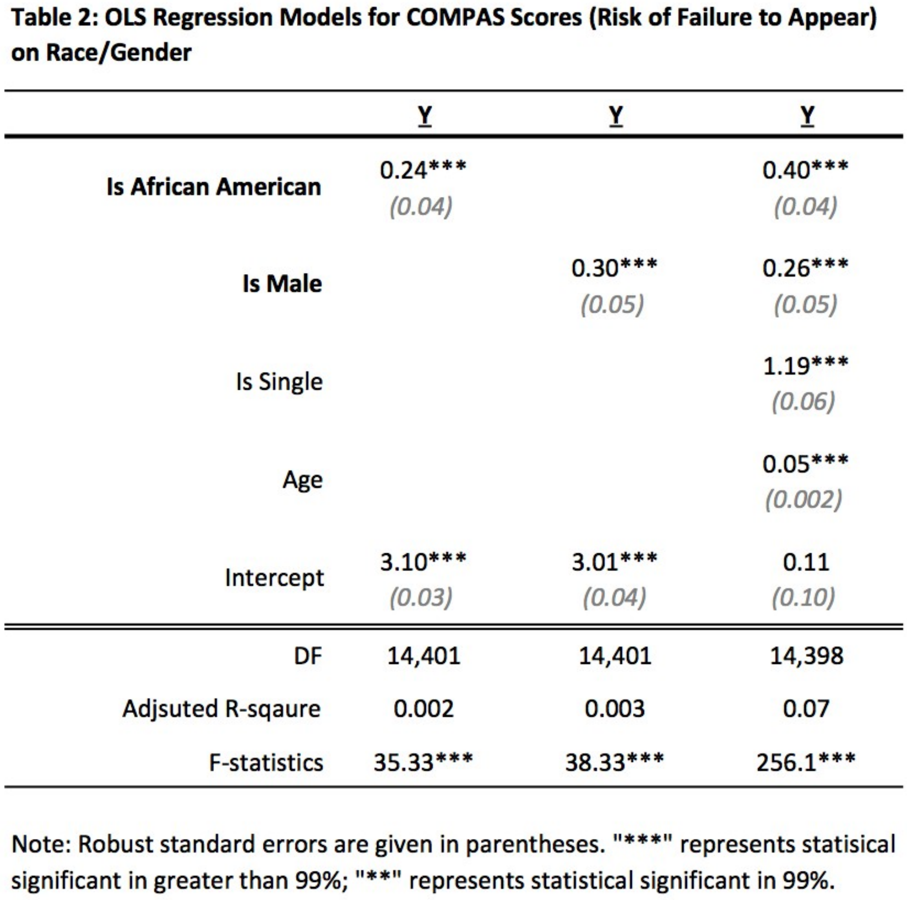


Table 2: OLS Regression Models for COMPAS Scores (Risk of Failure to Appear) on Race/Gender

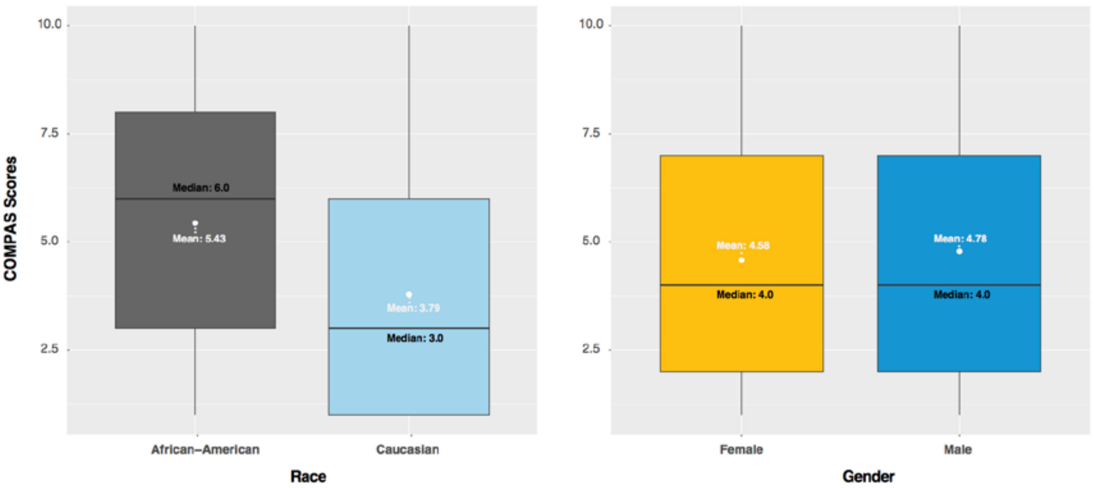
	Y	Y	Y
Is African American	0.24*** *(0.04)*		0.40*** *(0.04)*
Is Male		0.30*** *(0.05)*	0.26*** *(0.05)*
Is Single			1.19*** *(0.06)*
Age			0.05*** *(0.002)*
Intercept	3.10*** *(0.03)*	3.01*** *(0.04)*	0.11 *(0.10)*
DF	14,401	14,401	14,398
Adjsuted R-sqaure	0.002	0.003	0.07
F-statistics	35.33***	38.33***	256.1***

Note: Robust standard errors are given in parentheses. "***" represents statisical significant in greater than 99%; "**" represents statistical significant in 99%.

RISK OF RECIDIVISM

The six-number summary at Figure 2 illustrates the mean and median of the COMPAS score among African-American offenders (respectively 5.43 and 6.0) are both higher than those two statistics of the risk scores among Caucasian offenders (mean: 3.79; median: 3.0) in terms of the individual's risk of recidivism. In addition, the interquartile rage (IQR) of African-American's risk scores is also apparently higher than the IQR of Caucasian's COMPAS scores in terms of the offender's risk of recidivism. In contrast, the boxplots at Figure 2 doesn't show any significant differences of the recidivism risk scores between male and female offenders.

Figure 2: COMPAS Scores (Risk of Recidivism) Boxplots by Race & Gender



Furthermore, Table 3 presents that the average level of African-American offender's COMPAS score is statistically significantly higher than Caucasian offender's by 1.56 to 1.74-point at a 95% confidence level in terms of the risk of recidivism. Although the mean difference of the risk of recidivism scores between male and female offenders is also statistically significant, it's not substantially significant since the t-test shows the difference is between 0.09 to 0.31-point.

Table 3: T-Test for COMPAS Scores (Risk of Recidivism) Mean Differences on Race and Gender

Group	N	Mean	95% CI	t-score	df	P-Value
African-American	8,322	5.43	(1.56, 1.74)	-35.76	13,612	0.00
Caucasian	6,059	3.79				
Male	11,242	4.78	(0.09, 0.31)	-3.65	5,388	0.00
Female	3,139	4.58				

The OLS regression models (results listed at Table 4) show that, if an offender is African-American, he/she is expected to receive a 1.65-point higher risk score in terms of recidivism from the COMPAS algorithm than Caucasian offenders. In addition, on average, COMPAS systematically generates a 0.20-point higher recidivism risk score to male offenders than it does to female offenders. Moreover, controlling the offender’s marital status and age, it’s statistically significant that the COMPAS algorithm still produces a 1.17-point higher risk to African-American individuals and a 0.20-point higher score to male in terms of the offender’s risk of recidivism. However, the r-squared value of the specification #3 at regression model is only 0.23, which is relatively low and the model is not fully specified.

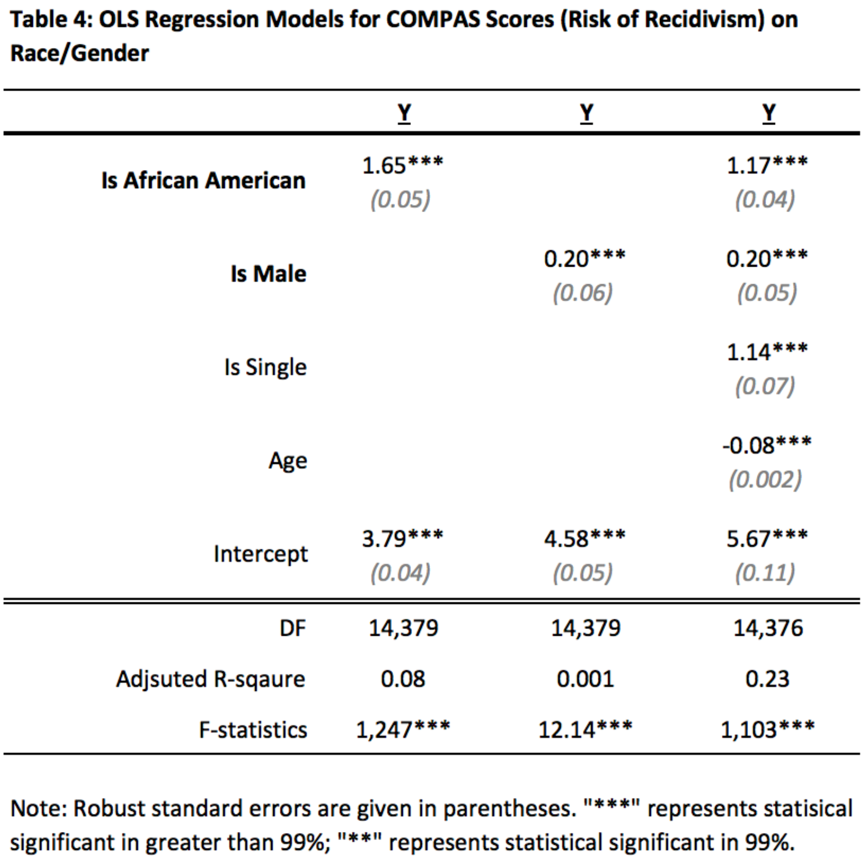


Table 4: OLS Regression Models for COMPAS Scores (Risk of Recidivism) on Race/Gender

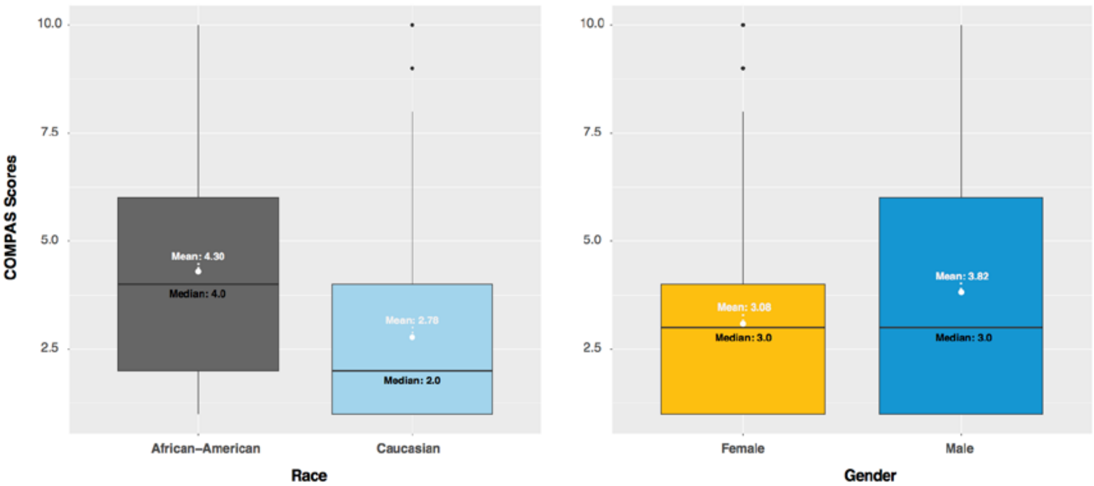
	Y	Y	Y
Is African American	1.65*** *(0.05)*		1.17*** *(0.04)*
Is Male		0.20*** *(0.06)*	0.20*** *(0.05)*
Is Single			1.14*** *(0.07)*
Age			-0.08*** *(0.002)*
Intercept	3.79*** *(0.04)*	4.58*** *(0.05)*	5.67*** *(0.11)*
DF	14,379	14,379	14,376
Adjsuted R-sqaure	0.08	0.001	0.23
F-statistics	1,247***	12.14***	1,103***

Note: Robust standard errors are given in parentheses. "***" represents statisical significant in greater than 99%; "**" represents statistical significant in 99%.

RISK OF VIOLENCE

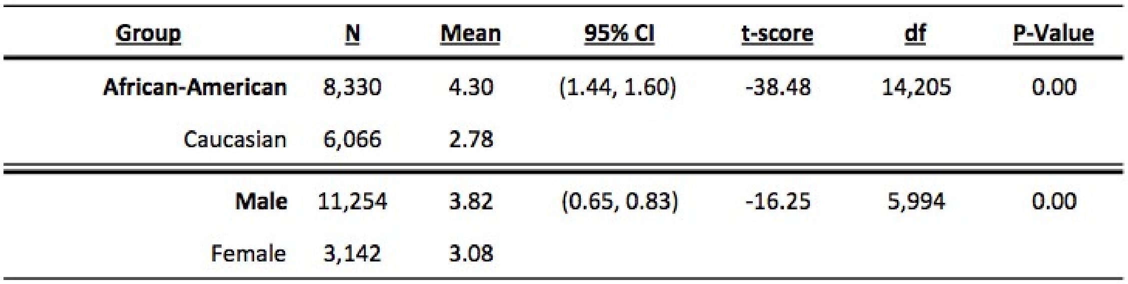
In terms of the offender’s risk of violence, Figure 3 demonstrates that the mean and median of the COMPAS scores in African-American offenders (respectively 4.3 and 4.0) are also higher than that in Caucasian offenders (mean: 2.78; median: 2.0); while it also illustrates that the IQR of the male offender's COMPAS score is apparently wider than that of the female offender's, indicating that the COMPAS scores among male offenders have a bigger spread towards a higher risk score.

Figure 3: COMPAS Scores (Risk of Violence) Boxplots by Race & Gender



The results of the two-sample t-tests (listed in Table 5) not only show the mean difference of the COMPAS scores is statistically significant between the two ethnic groups, but also indicate the average risk score in African-Americans is systemically higher than that in Caucasians by 1.44 to 1.60-point in terms of the offender's risk of violence. Similar to the risk scores in terms of recidivism, although the t-test proves the average level of the COMPAS scores is statistically significantly higher among the male offenders (compared to the female offenders), the difference is in the range of 0.65 to 0.83-point, which is not substantially significant.

Table 5: T-Test for COMPAS Scores (Risk of Violence) Mean Differences on Race and Gender



Group	N	Mean	95% CI	t-score	df	P-Value
African-American	8,330	4.30	(1.44, 1.60)	-38.48	14,205	0.00
Caucasian	6,066	2.78				
Male	11,254	3.82	(0.65, 0.83)	-16.25	5,994	0.00
Female	3,142	3.08				

Based on the regression analyses (results listed at Table 6), on average, if the offender is an African-American, the COMPAS model is expected to produce a 1.52-point higher score in terms of the risk of violence. Moreover, it's statistically significant that male offenders would receive a higher violence risk score by 0.74-point than female offenders would do. Additionally, holding the offender's marital status and age constant, the regression model indicates that being an African-American offender, on average, would receive a 0.93-point higher COPMAS score; while being a male offender is expected to be evaluated as 0.74-point higher risk to conduct violent crime by the COMPAS algorithm. All effects of ethnic and gender on COMPAS scores are statistically significant in greater than 99% confident level. However, the coefficient for being an African-American offender is lower in the multivariate regression model than that in the bivariate regression model, indicating part of the violence risk scores are explained not by the race variable but rather by the other control variables (including marital status and age). Last but not the least, the r-squared value in the multivariate regression model is only 0.38, indicating the model is not fully specified.

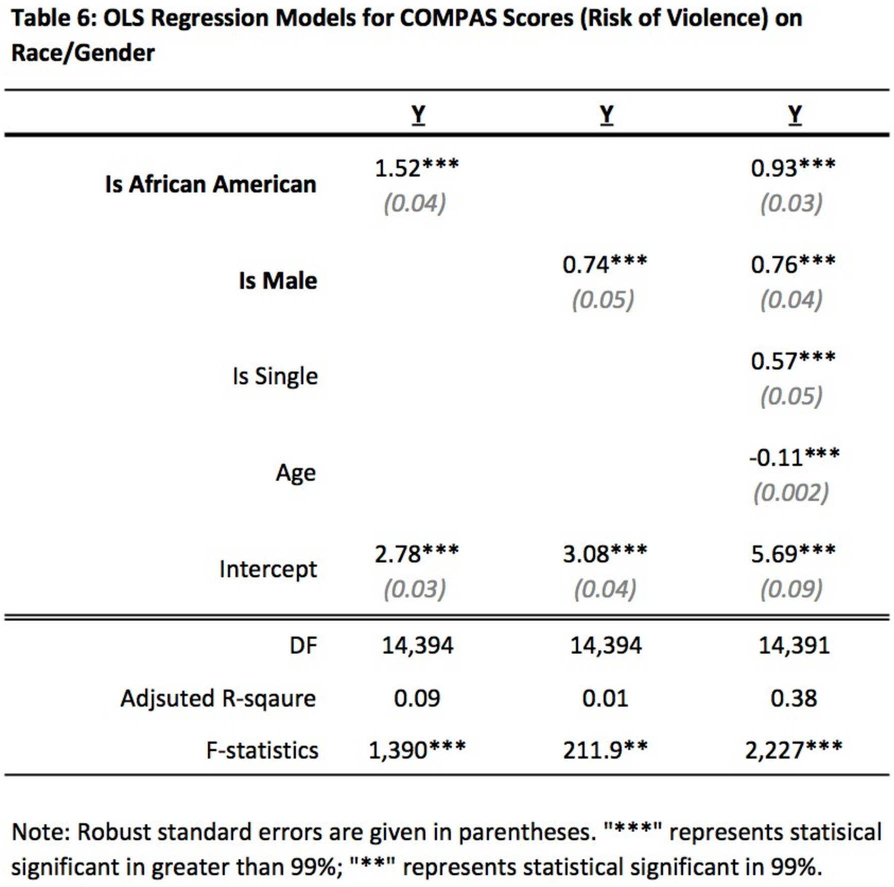


Table 6: OLS Regression Models for COMPAS Scores (Risk of Violence) on Race/Gender

	Y	Y	Y
Is African American	1.52*** *(0.04)*		0.93*** *(0.03)*
Is Male		0.74*** *(0.05)*	0.76*** *(0.04)*
Is Single			0.57*** *(0.05)*
Age			-0.11*** *(0.002)*
Intercept	2.78*** *(0.03)*	3.08*** *(0.04)*	5.69*** *(0.09)*
DF	14,394	14,394	14,391
Adjsuted R-sqaure	0.09	0.01	0.38
F-statistics	1,390***	211.9**	2,227***

Note: Robust standard errors are given in parentheses. "***" represents statisical significant in greater than 99%; "**" represents statistical significant in 99%.

CONCLUSION

Using statistical analyses to examine the fairness of the risk scores generated by the COMPAS system towards 13,186 offenders in Broward County FL, this paper found that the COMPAS algorithm systematically predicts a higher risk score towards African-American and male offenders. As Table 7 illustrates, although the differences in risk score between the two race (African-American versus Caucasian) and gender (male versus female) groups are all statistically significant in terms of all three types of risk category (risk of failure to appear (FTA), risk of recidivism, and risk of violence), the scale of the differences in risk scores varies in different type of risk category.

Overall, the results indicate that the effect of race and gender on the COMPAS scores in terms of risk of FTA is relatively low (less than 0.5-point). However, African-American offenders are expected to receive generally 1.0-point higher risk scores in terms of recidivism and violence; while the risk scores of male offenders are approximately 0.7-point higher in terms of these two risk types, raising an important concern regarding the algorithmic discrimination towards certain ethnic and gender groups encoded in the COMPAS assessment system.

Since it’s increasingly popular in the U.S. justice system to use these assessment instruments that produce a quantitative evaluation based on offender’s empirical characteristics as part of the judiciary decision making process, it’s particularly important to ensure these assessment tool powered by big data analytics algorithms to generate a fair prediction among all social groups that is solely based on the offender’s criminal factors instead of his/her race or gender. Without doubt, all judicial decisions have an important impact on relevant individual’s personal life as well as the fairness in the justice system, any potential bias encoded in these assessment tools that could discriminate against certain ethnic or gender groups must be investigated and addressed before any risk scores can be used in the court system to assist the decision making of the judicial officers and correctional professionals

While the findings at this paper are supported by careful designed methodology as well as statistical analyses, there are still some limitations and hence can be improved at future research. Specifically, the data sample at this paper is just limited to the Broward County in Florida so a bigger and more diverse data sample that cover more geographies could help further examine the fairness of COMPAS scores towards the two ethnic and gender groups. In addition, the r-square values at the regression models are relatively low because the control variables used at this paper are limited due to the data availability. Future research can employ a dataset that includes more control variables (such as education, household income, and more) for the statistical modeling in order to further investigate the effect of race and gender on the COMPAS scores while holding more potentially related characteristics constant.

Finally, the implications of this paper indicate the need for future research further analyzing the algorithmic discrimination in different big data analytics applications as well as the approach to address or even eliminate the bias encoded in the algorithms. This paper introduces the methodology and tools to investigate algorithmic bias and can be applied to examine the fairness of other big data analytics practices such as fraud prevention, recruiting, education admission, predictive policing, and more that utilize algorithms to assist human being's decision-making. Furthermore, although race is explicitly excluded as an input variable when the COMPAS algorithm was developed, the results at this paper suggest the system still produces discriminatory results for African-American offenders. Thus, simply removing the protected variable in building the algorithm is not sufficient to address the potential bias against the protected characteristics. Future research is needed to identify feasible approaches to address the algorithmic discrimination and provide best practices in big data analytics applications to ensure discriminatory bias are either thoroughly disclosed or appropriately justified.

REFERENCE

- ACM U.S. Public Policy Council. “Statement on Algorithmic Transparency and Accountability.” January 2017. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

- Andrews, D. A., & Bonta, J. “Level of Service Inventory–Revised (LSI-R): User’s Manual.” North Tonawanda, NY: Multi-Health Systems (2001).

- Andrews, Don A., James Bonta, and J. Stephen Wormith. "The Recent Past and Near Future of Risk and/or Need Assessment." Crime & Delinquency 52, no. 1 (2006): 7-27.

- Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." Pro Publica (2016).

- Barocas, Solon, and Selbst, Andrew D.. "Big Data's Disparate Impact." (2016).

- Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. "Quantifying and Reducing Stereotypes in Word Embeddings." arXiv Preprint arXiv:1606.06121 (2016).

- 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 (2009): 21-40.

- Datta, Amit, Michael Carl Tschantz, and Anupam Datta. "Automated Experiments on Ad Privacy Settings." Proceedings on Privacy Enhancing Technologies 2015, no. 1 (2015): 92-112.

- Dieterich, William, Christina Mendoza, and Tim Brennan. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Technical report, Northpointe, July 2016. http://www. northpointeinc. com/northpointe-analysis, 2016.

- Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through Awareness." In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214-226. ACM, 2012.
- 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 (2008).
- Flores, Anthony W., Christopher T. Lowenkamp, Alexander M. Holsinger, and Edward J. Latessa. "Predicting Outcome with the Level of Service Inventory-Revised: The Importance of Implementation Integrity." Journal of Criminal Justice 34, no. 5 (2006): 523-529.
- Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of Opportunity in Supervised Learning." In Advances in Neural Information Processing Systems, pp. 3315-3323. 2016.
- Holsinger, Alexander M., Christopher T. Lowenkamp, and Edward J. Latessa. "Ethnicity, Gender, and the Level of Service Inventory-Revised." Journal of Criminal Justice 31, no. 4 (2003): 309-320.
- Kroner, Daryl G., and Jeremy F. Mills. "The Accuracy of Five Risk Appraisal Instruments in Predicting Institutional Misconduct and New Convictions." Criminal Justice and Behavior 28, no. 4 (2001): 471-489.
- Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. "How We Analyzed the COMPAS Recidivism Algorithm." ProPublica (5 2016) (2016).
- Lipton, Zachary. “The Foundations of Algorithmic Bias.” (2016).
- Munoz, Cecilia, Megan Smith, and D. J. Patil. "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights." Executive Office of the President (2016).
- Skeem, J., and J. Eno Louden. "Assessment of Evidence on the Quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)." Unpublished Report Prepared for the California Department of

Corrections and Rehabilitation. Available at: https://webfiles. uci. edu/skeem/Downloads. html (2007).

- Sweeney, Latanya. "Discrimination in Online Ad Delivery." Queue 11, no. 3 (2013): 10.
- United States. Executive Office of the President, and John Podesta. Big Data: Seizing Opportunities, Preserving Values. 2014.
- Whiteacre, Kevin W. "Testing the Level of Service Inventory–Revised (LSI-R) for Racial/Ethnic Bias." Criminal Justice Policy Review 17, no. 3 (2006): 330-342.
- Zliobaite, Indre. "A Survey on Measuring Indirect Discrimination in Machine Learning." arXiv Preprint arXiv:1511.00148 (2015).

Yubin (Lee) Li

EDUCATION

Johns Hopkins University, Washington, DC

- Master of Science in Government Analytics, Expected May 2017
- **GPA: 3.9 / 4.0**

Loyola University Chicago, Chicago, IL

- Master of Public Policy, May 2014
- **GPA:** 3.9/4.0

EXPERIENCE

Western Washington University — Bellingham, WA

Research Associate, March 2017 – Present

- Assist with the data collection, data management, and quantitative analyses regarding a research project for the economic impact of foreign direct investment (FDI) in U.S.
- Research pull and push factors that affect FDI inflows and outflows on a global scale.

Invest In the USA (IIUSA) — Washington, DC

Policy Analyst, June 2014 – March 2017

- Conduct quantitative analysis on policy proposals for EB-5 Regional Center Program; compile policy evaluations and data analyses to inform IIUSA advocacy, education, and industry development efforts.
- Consolidate databases for EB-5 industry intelligence, collect and analyze data from various U.S. governmental agencies, and manage documents that IIUSA obtained via *Freedom of Information Act.*
- Research public policy related to the EB-5 Program including laws, regulations, administrative interpretations, case law, and more on a global scale to assist IIUSA in maintaining and advancing the best practices of the Program in overseas investor markets.
- Utilize data and quantitative analyses to illustrate the latest industry trends that empower IIUSA as the authoritative voice for the EB-5 Regional Center industry.

Invest In the USA (IIUSA) — Chicago, IL

Data Management Intern, November 2013 – June 2014

- Maintained various databases, managed EB-5 industry data, and conducted basic data reports.
- Assisted with various administrative tasks, including composing bilingual membership publicities in English and Chinese, translating EB-5 policy memos, and providing I.T. support.

Loyola University Chicago, Public Policy and Urban Affairs — Chicago, IL

Graduate Research Assistant, September – December 2013

- Assisted with data collection, data management, and statistical analysis for a study of inequality in U.S.
- Compiled and managed a large, national time series dataset, aggregating data from numerous federal and private sources, including the U.S. Census and the Bureau of Labor Statistics.

Taobao Marketplace — Hangzhou, Zhejiang, China

Social Media Marketing Intern, August 2010 – 2011

- Managed and analyzed original marketing data via SPSS; delivered actionable plans to supervisors

VOLUNTEER and SEMINARS

- US-China Education Policy Seminar with Ambassador Gary F. Locke *(2011)*, Guangzhou, China
- Study of the U.S. Institute on International Affairs for Asian Student Leaders *(2011)*, Washington D.C.