*Gregory T Adams*

*Bachelors Thesis, Polished Draft*

*February 8, 2017*

*All code used can be found at https://github.com/gregorytadams/BA*

*Risk Analysis and Criminal Justice*

A Study of Bias in the Implementation of Machine-Learning Recidivism Risk Algorithms at the State and Local Level

**Abstract:**

This study aims to provide a detailed analysis and evaluation of the implementation of recidivism risk algorithms at the state and county level. First, it discusses the high-level issues surrounding the implementation of these algorithms, from media response to constitutionality. It then progresses to evaluate a widely-circulated statistical analysis of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm, endorsing the study’s findings as both accurate and robust, and extending its methods of analysis of bias to gender and other races. It finds that there is evidence of bias not just against black defendants, but against multiple different races and women (using the same definition of bias as Propublica) as a result of base-rate differences. Finally, it suggests a framework for evaluating these algorithms based around rigorous statistical investigation and community feedback that can be used by policy implementers in search of a tailored, engaging approach to their implementation effort.

**Introduction:**

For years, state and local criminal justice systems have been using recidivism risk algorithms to judge the likelihood of repeat criminal offenses by defendants (in criminal trials) and inmates (in prisons). While these algorithms show immense promise—in areas as wide ranging as reducing bias and promoting equity to providing an avenue through which localities can streamline criminal justice procedures—there have been relatively few systematic analyses of the implementation and effectiveness of these methods. This paper aims to begin to rectify that problem.

The primary goal of this paper is to study the implementation of recidivism risk algorithms. These algorithms are used during sentencing to estimate the likelihood that a defendant is going to return to prison upon release. Given in the form of “score”—often between 0 and 10—the judge can use or weigh the estimation however they wish, including as an aggravating factor.

In order to accomplish that goal, this paper aims to find out, first, whether policies that implement these recidivism risk algorithms effectively address the problems they try to fix and, second, what unintended consequences may have resulted. Initially, these algorithms were implemented to address systemic problems in the criminal justice system: from judicial bias to inconsistent sentencing and alternative resource allocation problems (such as who gets drug rehabilitation facility spots). Recently, however, there have been criticisms alleging, among other things, that these scores are biased and ineffective.

Ultimately, this paper will explore the impact of technology within the policy world and expound upon the broader significance of technological advancement in how the government functions. Technology is changing how governmental policy is designed and implemented at every level. Both specific programs, like Chicago’s predictive policing initiatives, and general policy theory, such as evidence-based policy, are more and more drawing on newly accessible analysis methods, like machine learning, to identify and shape the best practices. But evaluation methods have lagged behind the spread of new practices, sometimes leaving public servants without a way to understand the impact of their policies. Accordingly, this paper aims to develop an evaluative framework through which policy designers (and perhaps the general public) can understand the implications of the policies they support.

**Analysis Plan:**

First, I address the pervasive media buzz that has surrounded the implementation of these algorithms. As with any new technology, journalists and the general public initially struggle to understand the impacts of these often opaque and complex instruments. Because public input and reactions are critical to any policy implementation, I begin my analysis with a discussion of how these methods are perceived. Some people highlight the positive: the potential of an egalitarian, efficient, and objective criminal justice system. Others, however, point out more controversial aspects of these algorithms: differential impact and bias (as alleged by ProPublica’s article, *Machine Bias*), the opacity of these algorithms’ evaluations, and allegations of double jeopardy. This give-and-take is critical to understanding how the public at large is reacting to these algorithms’ implementations. Because the public shapes the democratic institutions through which these policies are developed, any discussion of implementation needs to start with public perception.

After addressing public perception, I move on to discuss the potential benefits of these algorithms. These benefits—from increased efficiency to reductions in existing bias--provide valuable information necessary to effectively weigh the relative merits of criticisms levelled against the algorithms. Moreover, they provide benchmarks against which the success of different policies can be measured.

Third, this paper looks at where similar policies have been used or suggested, and to what effect. Police use predictive modelling to make their officers more efficient, credit companies use machine learning to (among other things) minimize lending risk, and some compelling research suggests that schools could use predictive algorithms to identify at-risk students that would benefit most from intervention. Each of these policies rely on very similar—or, for all we know, identical—technologies as recidivism risk algorithms, so the lessons learned through exploring those strategies are likely to be highly transferrable.

From there, this paper addresses concerns about the constitutionality of and legal limits on recidivism risk algorithms. There have been a variety of court cases arguing that these algorithms are unconstitutional as used, and the decisions have imposed important limits on practices. Accordingly, this paper briefly discusses the relevant case law in order to elucidate the boundaries within which policies are viable.

After that background and discussion of recidivism risk algorithms generally, this paper explores the low-level, technical aspects of this predictive technology, focusing in on one specific algorithm that has been particularly controversial: the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm. The topic of a particularly poignant long-form journalism piece from ProPublica (an article informed by much academic research), it has become a focal point for discussion about the implementation of recidivism risk algorithms. Accordingly, I first go over some of the academic literature relating to the COMPAS algorithm. Then, using the data provided by ProPublica, I recreate their analysis, changing their methods to test the robustness of their assertion of racial bias. Finally, I extend their analysis by repeating their methods across gender groups and different races (beyond white/black), testing the possibility of biases missed by the initial analysis.

**Background:**

*How to Define Bias*

The media has been largely ambivalent about these recidivism risk algorithms. On one hand, many people are skeptical of their impact and fear Minority-Report-style justice.[[1]](#footnote-1) On the other hand, some see these algorithms as holding the promise of a more egalitarian criminal justice system.[[2]](#footnote-2) The tension between these ideas is critical to the public perception of these policies.

One prominent critic of these algorithms is Harvard-trained mathematician Cathy O’Neil. A former hedge fund analyst, O’Neill recently released a book called *Weapons of Math Destruction* in which she argues (among other things) against the inequity of using recidivism models based on historical data that, in her view, is biased.[[3]](#footnote-3) In a National Public Radio (NPR) interview promoting her book, she uses the example of an engineering firm to illustrate a possible source of inequity:

“imagine [...] an engineering firm that decided to build a new hiring process for engineers and they say, OK, it's based on historical data that we have on what engineers we've hired in the past and how they've done and whether they've been successful. Then you might imagine that the algorithm would exclude women, for example. And the algorithm might do the right thing by excluding women if it's only told just to do what we have done historically. The problem is that when people trust things blindly and when they just apply them blindly, they don't think about cause and effect.”[[4]](#footnote-4)

By relying on biased data, O’neill argues, there is only one outcome: the perpetuation of bias. CNN, in an article (somewhat sensationally) entitled *Math is Racist: How Data is Driving Inequality*, further explains this argument, explaining that while zip codes, credit scores and grammar are all used to evaluate customers in the private sector, “zip codes are also a stand-in for race, credit scores for wealth, and poor grammar for immigrants.”[[5]](#footnote-5)

As a counterpoint to this argument, National Review columnist David French critiqued the definition of racism and inequity presented by O’Neill and those covering her work. He disputes the idea that a “fair result” definitionally “breaks down precisely along demographic lines.”[[6]](#footnote-6) In this view, the underlying bias of the data is irrelevant. These algorithms, ostensibly meant to accurately predict and reflect the real world, should use whatever data allows them to do that best. Whereas O’Neill believes differential outcomes are necessarily problematic, French argues that they could be fair, so long as the differences are already present in the real world.

Weighing in on this argument, the Ford Foundation published an article, *Can Computers Be Racist? Big Data, Inequality, and Discrimination*, proposing a solution. The article quotes two prominent professors, Latanya Sweeney of Harvard and Alvaro Bedoya of Georgetown, who work on predictive algorithms.[[7]](#footnote-7) Sweeney published a study alleging racial bias in advertisements on Google’s search engine; she found that, when one searches for a name “‘racially associated’ with the black community,” ads implying the subject had previously been incarcerated were much more likely to appear.[[8]](#footnote-8) Professor Bedoya is the executive director of the Center on Privacy and Technology.[[9]](#footnote-9) They both claim that predictive algorithms should account for the potential bias in underlying data and compensate for it. Ultimately, the article suggests three steps organizations can take to address the controversy: getting more people interested, publishing relevant data (“algorithmic transparency,” as the author calls it), and updated regulations on personal data use.[[10]](#footnote-10)

*Potential Benefits*

Beyond the critiques of bias, there have been a number of articles talking about the potential of these algorithms to revolutionize the way police forces and criminal justice systems are able to allocate their resources. From the National Institute of Justice to congressional press releases, multiple sources have praised the possible impact of these methods on inefficient bureaucracies and less-than-objective criminal justice processes.[[11]](#footnote-11)[[12]](#footnote-12)

The National Institute of Justice, the research wing of the United States Department of Justice, published an article detailing the implementation of recidivism risk modelling in Philadelphia. The report, *Predicting Recidivism Risk: New Tool in Philadelphia Shows Great Promise*, talks about the gains in efficiency and effectiveness the Philadelphia Police Department was able to make due to the new methods. [[13]](#footnote-13) The tool allows the police to “base their personnel and policy decisions on a scientifically proven method,” rather than subjective impressions.[[14]](#footnote-14) The Philadelphia Corrections Department was able to operate more efficiently because they could “concentrate resources on a small number of probationers who require more active supervision, rather than on those who are unlikely to reoffend regardless of how they are supervised.”[[15]](#footnote-15) And addressing some of the concerns raised in other articles in the media, the author argues that, in any criminal justice system, predictions already happen. “Everyone involved in the criminal justice system [...] is making judgments, essentially predictions, about the relative risk of an offender,” the author points out, suggesting that a scientific method is better than one based off of personal impressions.[[16]](#footnote-16) This promise is a large part of why predictive algorithms are able to gain traction; they imply the possibility of reduced costs, quicker decision times, and more well-informed decisions.

Congress, riding this potential, recently passed a law encouraging the use of recidivism risk algorithms, called the Sentencing Reform and Corrections Act (SRCA). Partnered with recidivism reduction strategies like employment assistance and transition programs, Senator Sheldon Whitehouse (D-RI) claims that the same recidivism risk algorithms that the SRCA encourage helped reduce recidivism in Rhode Island by 17 percent.[[17]](#footnote-17)[[18]](#footnote-18)[[19]](#footnote-19) Taken as another concrete example of the potential of these algorithms to make the criminal justice system more fair and efficient, this claim provides compelling evidence in the media to push forward with the development of these algorithms.

*Other Uses of Similar Technology*

Recidivism risk algorithms, while relatively new, draw on technology that the private sector and government already use for other purposes. Predictive policing is one area where prediction has had success. Richmond, Nashville and Kansas City all have had success combining simple and complex predictive methods to reduce crime. Another area is in estimating credit-worthiness. Credit card companies have been using automated algorithms to reduce their non-repayment rates for years, and have even faced many of the same accusations that are now being leveled against the recidivism-risk algorithms. A third area is education. At the leading edge of technology, work is being done to predict which students are most in need of help.

In policing, predictive technology is used in a variety of ways, from guiding patrols to identifying suspects. Three big technologies used are spatio-temporal analysis, regression modelling and machine learning.[[20]](#footnote-20) All of these methods are likely used in some way in recidivism risk modelling, though how much each is used is impossible to tell, given the models’ proprietary nature.

Spatio-temporal analysis is any predictive policing method that leverages time and/or spatial data to reveal patterns.[[21]](#footnote-21)  This can include, but is not limited to, hotspot mapping, grid-based patrol patterns, and density-estimation models.[[22]](#footnote-22) Richmond used this type of analysis to increase police seizures of weapons by 246% for one day. Every New Year’s Eve for years before their predictive policy was implemented, there had been widespread reports of random gunfire.  By mapping the city and identifying hotspots, they were able to most effectively distribute manpower, reducing gunfire by almost half and saving $15,000 in the process. [[23]](#footnote-23)

Regression modelling, able to identify a moderate diversity of relationships (e.g. linear, nonlinear, splines), has also been used to predict crime frequency and guide officers in responding to criminal behavior. In Nashville, Tennessee, a statistician named Ronald Wilson was able to cross reference disparate types of crime data, such as traffic stops and assault reports, to predict other types of crimes, such as drunk driving.[[24]](#footnote-24) The police were able to use this analysis to reduce fatal crashes in the city by over 30% and increase drunk driving arrests by 72.3%.  By integrating and building statistical models out of this data, the Nashville police department pioneered what is now called the data-driven approach to crime and traffic safety. [[25]](#footnote-25)

Machine learning modelling is at the leading edge of predictive policing, and is the method most likely used in building recidivism risk algorithms.  Though adoption has been slow--not many police officers are equipped to understand the complex methods--it shows promise, and has been embraced by police departments in cities as large as Chicago and Kansas City.[[26]](#footnote-26)

There are few rigorous academic studies analyzing the effectiveness of these machine learning methods (just as there are few analyses of recidivism risk algorithms), but they have gotten a large amount of press.  Chicago has used network analyses and data mining to compile a list of individuals they believe are likely to be involved in a fatal shooting. [[27]](#footnote-27)  Despite the recent rise in violent crime, CPD deputy police chief Jonathan Lewin has praised the models as a way to “inform” the policing process.[[28]](#footnote-28) Kansas City is using social media data and unsupervised clustering to identify “crews,” warning them that they will target the whole group for investigation if any of them commit a crime.[[29]](#footnote-29)  Controversial in its public image, the police department nonetheless believes it will be instrumental in reducing criminal activity.  “We have a moral reason to do a better job at addressing violence in the community” the local prosecutor in Kansas City said; “we need to try.”[[30]](#footnote-30)

In the private sector, predictive technologies are widely used to estimate credit-worthiness. From a few high-profile startups aiming to use social media and cellphone data to large institutions using advanced analytics, the applications of these algorithms have been growing for years. While it is difficult to identify specific users of a given technology—the data and methods are mostly proprietary—it is very likely widespread. One analysis from a team at the Massachusetts Institute of Technology used data from six major commercial banks in the U.S. to leveraged to build a model that predicted credit delinquency. Though the banks themselves were not named (the federal regulator which provided the MIT team with the data forbade disclosure of the data source), the conclusion of the paper suggests overhauling regulatory methods based on the individualized prediction techniques.[[31]](#footnote-31)

Along with these techniques, suggestions of bias have surfaced. One example was in 2015, when a man named Kevin Johnson had his credit limit reduced by almost 70% because he “had been shopping at stores frequented by people deemed by the credit card company to have a poor repayment history. “[[32]](#footnote-32) A result of these predictive algorithms, the practice drew comparisons with a racist housing practice called “redlining,” where different lending patterns led to effectively segregated cities, especially in the American Southeast.[[33]](#footnote-33)

One new area where this technology has been suggested is schools. It aims to address two common problems teachers and administrators face: first, identifying which students are most in need of help and, second, prioritizing which students are most likely to benefit from different intervention strategies. One piece of research out of the University of Chicago’s Data Science for Social Good tests a variety of machine learning methods and classifiers on a dataset of nearly 200,000 students, building a technical framework for future work.[[34]](#footnote-34) They have somewhat considerable success predicting which students are most likely to drop out of high school before graduating, as they are able to build a classifier that identifies almost 90% of the students that will drop out with only a 20% false-positive rate.[[35]](#footnote-35) The results not only suggests immense potential for these methods, but it signals a willingness by school officials to explore the new technology.

*Legal Limits and Constitutionality*

One big issue with using recidivism risk algorithms for sentencing is the controversy about double jeopardy. Presumably, a recidivist would have already been punished for the crimes they had committed previously. It could be argued that considering those crimes—even as evidence of future risk—indicates that a defendant is being punished twice for the same crime. More than that, using a prediction of future crimes as an aggravating factor in sentencing seems to be punishing a defendant in the present for a crime that may happen in the future. These two issues have been brought before various courts in the United States. Despite some controversy, the supreme court has ruled that both recidivism and underlying conduct of the defendant (i.e. other crimes that the defendant may have committed but was not convicted of) can both be used as aggravating factors in sentencing. And though recidivism risk has not been tested in the U.S.’s highest court, the Wisconsin Supreme Court has held that they too can be used as aggravating factors.

In *Almandarez-Torres v. United States*, a 1998 Supreme Court case addressing the validity of using recidivism as an aggravating factor in sentencing, the court upheld the Fifth Circuit Court of Appeals’ decision that recidivism is a valid consideration for sentencing under a law that prescribes a penalty for repeat offense.[[36]](#footnote-36) In the case, a man was sentenced to 85 months in prison for illegally returning to the United States after having been deported as a result of an “aggravated felony.”[[37]](#footnote-37) He argued that sentencing based on past crime was double jeopardy. The court rejected that argument, concluding that judges may apply additional sentencing penalties under laws which prescribe them.[[38]](#footnote-38)

A related case to Almendarez had come up the previous year, in 1997. Though not dealing specifically with recidivism, *United States v. Watts* held that a defendant’s underlying conduct can be used as an aggravating factor for sentencing under the preponderance of evidence standard (rather than beyond a reasonable doubt), even if they had been acquitted of the charge(s) during trial.[[39]](#footnote-39) Though not directly related to recidivism, the case has been used to argue that, for sentencing, there need not be definitive evidence of a crime for a judge to consider what evidence there is while sentencing the defendant for a different crime.[[40]](#footnote-40) In fact, this case was cited in the arguments against *Almendarez-Torres* to justify using recidivism as a factor.

The above two cases set precedents that were relevant for the Wisconsin Supreme Court’s decision to uphold the use of recidivism risk algorithms in sentencing. In *Wisconsin V. Loomis*, the state charges Eric L. Loomis with five crimes related to a drive-by shooting. Rather than go to trial, Loomis accepted a plea deal whereby he would plead guilty to two of the lesser charges and the evidence for the greater charges (as well as a COMPAS risk assessment) would be presented to the judge at sentencing.

Loomis argued (among other things), that the COMPAS algorithm, as an opaque and unusual piece of evidence, violated his right to due process. The state, on the other hand, countered by asserting that, as it was used in conjunction with other factual pieces of evidence, the COMPAS score was not unduly determinative.[[41]](#footnote-41) The court agreed with the state, saying:

“We determine that because the circuit court explained that its consideration of the COMPAS risk scores was supported by other independent factors, its use was not determinative in deciding whether Loomis could be supervised safely and effectively in the community. Therefore, the circuit court did not erroneously exercise its discretion.”[[42]](#footnote-42)

This was decision was further supported by a disclaimer that accompanied the COMPAS risk assessment score:

“For purposes of Evidence Based Sentencing, actuarial assessment tools [such as the COMPAS score] are especially relevant to: 1. Identify offenders who should be targeted for interventions. 2. Identify dynamic risk factors to target with conditions of supervision. 3. It is very important to remember that risk scores are not intended to determine the severity of the sentence or whether an offender is incarcerated.”[[43]](#footnote-43)

It is important to recognize the limits imposed here by the Court: first, that “other independent factors” need to be considered along with recidivism risk scores; second, that risk scores should not be used to decide whether or not an individual should be incarcerated; and third, that the algorithms can be used to suggest alternative sentencing to prison.

The implications of each of these cases reveal different aspects of the legal use of recidivism risk algorithms. *Almendarez-Torres* shows that recidivism is a valid sentencing consideration. *Watts* shows that underlying conduct of a defendant (as one could consider the COMPAS score to signal) is allowed to be considered. And the *Loomis* case shows many of the limits of these algorithms while upholding their constitutionality.

**Literature Review:**

**Initial article:**

[**https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing**](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing)

**Evaluation study:**

[**http://journals.sagepub.com/doi/pdf/10.1177/0093854808326545**](http://journals.sagepub.com/doi/pdf/10.1177/0093854808326545)

**Northpointe Response:**

[**http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica\_Commentary\_Final\_070616.pdf**](http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf)

**ProPublica Responses to the response:**

**In article form:**

[**https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story**](https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story)

**Technical response:**

[**https://www.propublica.org/article/technical-response-to-northpointe**](https://www.propublica.org/article/technical-response-to-northpointe)

**Rando critiqued Propublica’s technical response, so ProPublica annotated the critique, responding point-by-point:**

[**https://www.documentcloud.org/documents/3248777-Lowenkamp-Fedprobation-sept2016-0.html**](https://www.documentcloud.org/documents/3248777-Lowenkamp-Fedprobation-sept2016-0.html)

**Stanford researchers:**

**“impossible to satisfy both definitions of fairness”**

**article:**

[**https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say**](https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say)

**mathematical proof (lol):**

[**https://arxiv.org/abs/1609.05807**](https://arxiv.org/abs/1609.05807)

**Methods:**

I follow very similar methods to those laid out in the ProPublica analysis of racial bias in the COMPAS algorithm.[[44]](#footnote-44) They use four primary techniques to break down the problem. First, they perform descriptive analysis of the dataset, developing a nuanced picture of everything from demographics to score distribution. They then move on to regression analysis to find the relative risk of each race in terms of getting scored either medium/high or low. After the regressions, they run a Cox Proportional Hazard model to test predictive accuracy (which, they note in both their article and analysis comments, is the same test Northpointe ran in their validations study), and, finally, move on to an analysis of errors, comparing the differences between black and white defendants.

I follow many of the same analysis methods as ProPublica in order to directly compare the results of my analysis to theirs. I go through and change different variables based on the method, showing whether or not the analysis is a reflection of a general trend or the specific parameters the analysts used to measure impact. For example, in the linear regressions, ProPublica used a binomial logistic regression of scores “high” and “medium” against score “low.” While this is undoubtable a valid method, certain other methods such as multinomial logistic regression or binomial regression with medium and low grouped together are also valid. My analysis will test these design decisions, and see how they impact the final results (if at all). Many of the design decisions I test are informed by Northpointe’s critiques of the initial analysis and ProPublica’s response; many are my own.

Once I’ve tested the robustness of ProPublica’s analysis, I begin to extend it. One design decision ProPublica made was to isolate the analysis to only black and white defendants. Though such a decision is not necessarily wrong, it isolates the direct relevance of the study’s conclusions to a single group. I therefore, first, include the other ethnicities present in the dataset and, second bring in gender as a variable.

The final thing I do is combine the previous two sections to attempt to draw more granular conclusions about the bias present in the algorithm. Having broken the data down separately by both race and gender, I test the differences between the cross-tabulated groups (e.g. black men and white women). Often, disaggregated data will have a different relationship than aggregated data because of on lurking variables.[[45]](#footnote-45) By breaking down the analysis in this way, I hope to be able to identify if any lurking variables have had a significant impact on my or ProPublica’s conclusions.

**Analysis:**

**Policy Recommendations:**

There are three primary things that someone implementing recidivism risk algorithms needs to consider before moving forward. While not comprehensive, these recommendations are currently missed or underemphasized in the existing implementation literature and are important for policy professionals to consider in order to both implement effective policy and mitigate whatever unexpected consequences or impacts may arise out of the policy:

1. Clarify and emphasize the instructions given to judges along with the algorithms.
2. Rigorously pilot and evaluate the algorithms before implementation
3. Publicize data regarding the use of these algorithms periodically

*Clarify Instructions*

It is critical that judges know how they should and shouldn’t use these algorithms. In the Wisconsin Supreme Court case, the opinion (affirming the validity of recidivism risk as a valid sentencing tool/guideline) hinged on how the algorithm was used. Nonetheless, judges get very little training or guidance with regard to what exactly the numbers they get mean.

For example, the difference between a 4 and a 6 rating may be much smaller--or much larger--than the difference between a 6 and an 8 rating with regard to the danger of recidivism or violence. But judges aren’t given enough information or training to make an informed decision about that difference. As a result, they’re forced to make their own judgement calls--decisions that legal training in no way, shape or form provides you with tools to effectively make--that have almost arbitrary outcomes and massive impacts on people’s lives. Judges need to be given more information and training to mitigate that process.

*Rigorously Evaluate*

With most government programs, evaluation is important to measure effectiveness. But it has a high cost: data gathering, entry and cleaning is troublesome; analysis techniques need to be developed anew each time, and each time are subject to controversy; the cost of qualified experts only increases as previous evaluations are used and the evaluators are proven to be high quality; and political ramifications of failure can be huge. But with high-technology, the marginal cost of effective evaluation tends to be rather low. Impact analysis can often be mostly automated once the data is gathered. While effectively integrating data is often difficult, once it is done once (as many recidivism-risk policies necessitate), using the data tends to be rather straightforward. And interpretation of the results, while it is complex, requires few people and can be easily standardized. There are few reasons why these algorithms should be so opaque to the very governments that use them.

*Publicize Data*

Because the scores are part of ongoing, public criminal justice trials, the scores reside in the public domain. Nonetheless, companies like Northpointe resist releasing data that can be used by independent experts to evaluate the algorithms. To them, the data is proprietary.

For state and local governments, however, the data is public and can be requested under the Freedom of Information Act.[[46]](#footnote-46) But rather than forcing citizens to hire a lawyer to get to the data, governments should make access as simple as possible. Judges are subject to elections every so often in order to ensure they are responsive to the public they serve (and to allow constituents to remove them if they dislike the results of the judge’s work). There is no similar process for these algorithms, which have a huge effect on sentencing.

It is possible to provide responsive government when implementing technical solutions by making data easily accessible to people who can understand the impact of the new technology. Chicago already makes much government-held data public on their data portal.[[47]](#footnote-47) ProPublica made their dataset available on Github, a cloud-based hosting service commonly used by people with the skillset to effectively use the data.[[48]](#footnote-48)

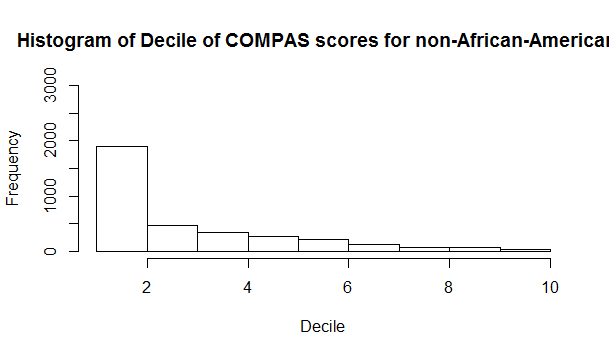
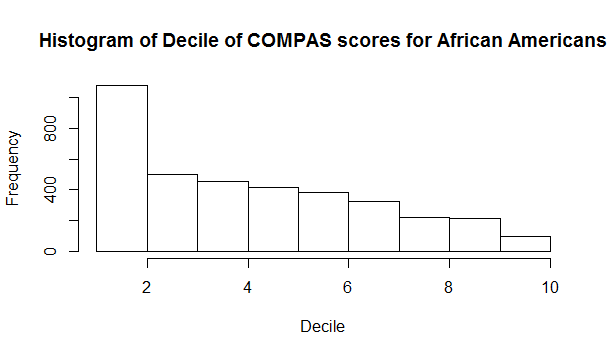
*Implications*

The implications of these recidivism risk algorithms have been substantial. They have helped determine the fate of a huge number of defendants over the past twenty years. They have had significant unintended consequences--years of litigation, allegations of bias, media attention and differential impact--and have nonetheless continued to thrive. Going forward, governments implementing these algorithms need to better clarify parameters, evaluate the algorithms (and publish the results) and publicize the data that they have.

Where will you describe the categories that the algorithms use? Will this go in the lit review? Or your methods section?

**Analysis:**

The initial and most intuitive way to look for bias in the recidivism risk algorithms is to see the difference in scores for each group, African American and non-African American, and compare them to the overall rate and to each other.

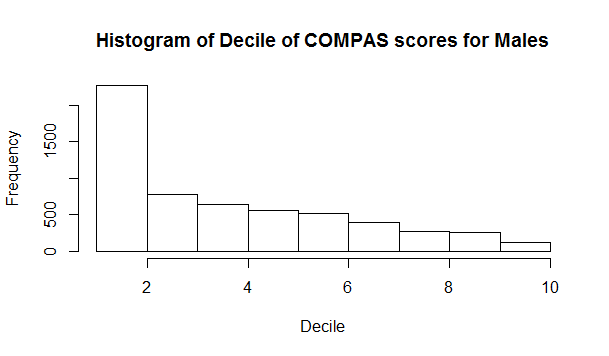
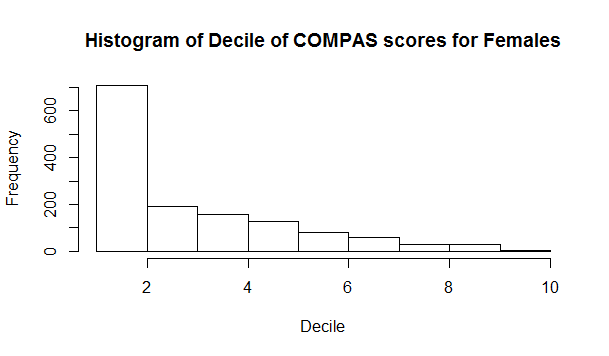


Comparing the relative rates of higher scores to lower ones, it becomes apparent that African Americans are more likely to be given higher scores using this algorithm. Overall, only 15.6% of African Americans scored by the COMPAS algorithm were given scores in the bottom decile, while 37.9% of non-African-American respondents were given bottom-level scores.

In your discussion, you’ll probably want to talk about how your results could tie back to the definitions of bias you cite in your lit review.

How would you go about assessing this?

This cursory finding does not necessarily imply that the algorithm is biased, just that the scores are not equally distributed across racial lines. Moreover, this gives us no evidence that this difference is systematic or non-random. With lurking variables like education, income and geography, it is hard to tell whether the difference in scores is significant, or even particularly relevant.

This finding is also mirrored in scores across gender groups. 

It seems that females are likely to get lower scores than males, in a similar dynamic to the racial breakdown. 24.7% of men are given bottom-level scores, while 34.0% of women are given those scores. Again, this does not necessarily imply bias, but it does indicate this is a possible avenue for future analysis.

What other kinds of analyses will you do with these variables? Are you able to combine them with different variables? What about age? Can you break the analysis down further by racial/ethnic group? You’ll want to describe your approach pretty thoroughly in the Methods section. Right now the Methods/Analysis sections are very light on details.

**Reference section?**

Greg –

This is off to a very good start. I don’t have a lot to add to my comments above. Mainly, just make sure that in the polished draft you walk us through your methodological approach (what are the parameters of your data set, what kinds of statistical analysis you’ll be using, etc.) before you jump into the analysis itself, and be sure to indicate to us how your analysis of these data is different from that of ProPublica. I look forward to seeing what you find.

Grade: 10/10

1. Rawlins, Aimee. “Math Is Racist: How Data Is Driving Inequality.” *CNNMoney*, September 6, 2016. <http://money.cnn.com/2016/09/06/technology/weapons-of-math-destruction/index.html>. [↑](#footnote-ref-1)
2. Nancy Ritter, "Predicting Recidivism Risk: New Tool in Philadelphia Shows Great Promise," NIJ Journal, no. 271 (February 2013): , accessed November 28, 2016, https://www.ncjrs.gov/pdffiles1/nij/240696.pdf. [↑](#footnote-ref-2)
3. McEvers, Kelly. "'Weapons Of Math Destruction' Outlines Dangers Of Relying On Data Analytics." NPR. Accessed November 28, 2016. http://www.npr.org/2016/09/12/493654950/weapons-of-math-destruction-outlines-dangers-of-relying-on-data-analytics. [↑](#footnote-ref-3)
4. ibid. [↑](#footnote-ref-4)
5. Rawlins, *Math is Racist* [↑](#footnote-ref-5)
6. French, David. "No, Math Isn't Racist." National Review. September 08, 2016. Accessed November 28, 2016. <http://www.nationalreview.com/corner/439846/no-math-isnt-racist> [↑](#footnote-ref-6)
7. “Can Computers Be Racist? Big Data, Inequality, and Discrimination.” *Ford Foundation*. Accessed October 18, 2016. <https://www.fordfoundation.org/ideas/equals-change-blog/posts/can-computers-be-racist-big-data-inequality-and-discrimination/>. [↑](#footnote-ref-7)
8. Ibid. [↑](#footnote-ref-8)
9. Ibid. [↑](#footnote-ref-9)
10. Ibid. [↑](#footnote-ref-10)
11. Ritter, Nancy. "Predicting Recidivism Risk: New Tool in Philadelphia Shows Great Promise." National Institute of Justice. Accessed November 28, 2016. http://www.nij.gov/journals/271/pages/predicting-recidivism.aspx. [↑](#footnote-ref-11)
12. Judiciary Committee Clears Sentencing Reform and Corrections Act." U.S. Senator Sheldon Whitehouse of Rhode Island. Accessed February 03, 2017. https://www.whitehouse.senate.gov/news/release/judiciary-committee-clears-sentencing-reform-and-corrections-act. [↑](#footnote-ref-12)
13. Ibid. [↑](#footnote-ref-13)
14. Ibid. [↑](#footnote-ref-14)
15. Ibid. [↑](#footnote-ref-15)
16. Ibid. [↑](#footnote-ref-16)
17. Waugh, Christopher I. "Prison by Algorithm." The Atlantic. Accessed February 03, 2017. http://www.theatlantic.com/politics/archive/2016/06/congress-takes-on-recidivism/488741/. [↑](#footnote-ref-17)
18. "Judiciary Committee Clears Sentencing Reform and Corrections Act." U.S. Senator Sheldon Whitehouse of Rhode Island. Accessed February 03, 2017. https://www.whitehouse.senate.gov/news/release/judiciary-committee-clears-sentencing-reform-and-corrections-act. [↑](#footnote-ref-18)
19. "RI Delegation Announces $500K for Providence-Cranston Partnership to Reduce Recidivism." Whitehouse.senate.gov. July 01, 2016. Accessed February 03, 2017. https://www.whitehouse.senate.gov/news/release/ri-delegation-announces-500k-for-providence-cranston-partnership-to-reduce-recidivism. [↑](#footnote-ref-19)
20. Perry, Walter L., Brian McInnis, Carter C. Price, Susan C. Smith, and John S. Hollywood. Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. Rand Corporation. Ncjrs.gov. 2013. Accessed November 4, 2016. <https://www.ncjrs.gov/pdffiles1/nij/grants/243830.pdf>. [↑](#footnote-ref-20)
21. Perry et al., xxi-xxiii [↑](#footnote-ref-21)
22. Perry et al., 48 [↑](#footnote-ref-22)
23. Perry et al., 46 [↑](#footnote-ref-23)
24. Wilson drew on a criminological concept called “deviant place theory,” which holds that the individual characteristics that lead people to commit crimes are not limited to certain criminal categories. For example, the impulse that leads one to shoplift is the same that leads one to drink and drive, or burgle a home (Perry et al., 69) [↑](#footnote-ref-24)
25. Perry et al., 29-37 [↑](#footnote-ref-25)
26. Hvistendahl, Mara. "Can ‘predictive Policing’ Prevent Crime before It Happens?" Sciencemag.com. Accessed November 4, 2016. <http://www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens>. [↑](#footnote-ref-26)
27. Davey, Monica. "Chicago Police Try to Predict Who May Shoot or Be Shot." Nytimes.com, May 23, 2016. Accessed November 4, 2016. <http://www.nytimes.com/2016/05/24/us/armed-with-data-chicago-police-try-to-predict-who-may-shoot-or-be-shot.html>. [↑](#footnote-ref-27)
28. Ibid. [↑](#footnote-ref-28)
29. Eligon, John, and Timothy Williams. "Police Program Aims to Pinpoint Those Most Likely to Commit Crimes." Nytimes.com, September 24, 2015. Accessed November 4, 2016. <http://www.nytimes.com/2015/09/25/us/police-program-aims-to-pinpoint-those-most-likely-to-commit-crimes.html>. [↑](#footnote-ref-29)
30. Ibid. [↑](#footnote-ref-30)
31. Butaru, Florentin, Qingqing Chen, Brian Clark, Sanmay Das, Andrew Lo, and Akhtar Siddique. "Risk and Risk Management in the Credit Card Industry." 2015. Accessed February 4, 2017. doi:10.3386/w21305. [↑](#footnote-ref-31)
32. Alloway, Tracy. "Big data: Credit where credit’s due." Ft.com. January 4, 2015. Accessed February 04, 2017. https://www.ft.com/content/7933792e-a2e6-11e4-9c06-00144feab7de. [↑](#footnote-ref-32)
33. Ibid. [↑](#footnote-ref-33)
34. Lakkaraju, Himabindu et al.. A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes∗. Data Science for Social Good. ACM Digital Library. Accessed February 3, 2017. <https://dssg.uchicago.edu/wp-content/uploads/2016/04/montogmery-kd2015.pdf> [↑](#footnote-ref-34)
35. Ibid. [↑](#footnote-ref-35)
36. "Almendarez-Torres v. United States 523 U.S. 224 (1998)." Justia Law. Accessed February 04, 2017. https://supreme.justia.com/cases/federal/us/523/224/case.html. [↑](#footnote-ref-36)
37. "Almendarez-Torres v. United States." *Oyez,* https://www.oyez.org/cases/1997/96-6839. Accessed 4 Feb. 2017. [↑](#footnote-ref-37)
38. Ibid. [↑](#footnote-ref-38)
39. Chicago-Kent College of Law at Illinois Tech. "United States v. Watts." Oyez. https://www.oyez.org/cases/1996/95-1906 (accessed February 4, 2017). [↑](#footnote-ref-39)
40. "United States v. Watts 519 U.S. 148 (1997)." Justia Law. Accessed February 04, 2017. https://supreme.justia.com/cases/federal/us/519/148/case.html. [↑](#footnote-ref-40)
41. State of Wisconsin V. Eric L. Loomis (July 13, 2016). <https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690> [↑](#footnote-ref-41)
42. Wisconsin V. Loomis, 6 [↑](#footnote-ref-42)
43. Wisconsin V. Loomis, 7 [↑](#footnote-ref-43)
44. The entire ProPublica technical analysis, which I reference throughout the following sections, can be found here: https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb [↑](#footnote-ref-44)
45. One such example of this is Simpson’s Paradox, where the relationship between aggregated and disaggregated data is opposite. An example of this might look like, in a given organization, women making more money than men on average, but make less than a man in each job. The because they tend to be in higher-paying jobs. Without breaking down the analysis into specific groups (e.g. type of job), it would be very difficult to identify such a scenario. [↑](#footnote-ref-45)
46. *Machine Bias* [↑](#footnote-ref-46)
47. "City of Chicago | Data Portal." Chicago. Accessed February 03, 2017. https://data.cityofchicago.org/. [↑](#footnote-ref-47)
48. Propublica. "Propublica/compas-analysis." GitHub. July 29, 2016. Accessed February 03, 2017. https://github.com/propublica/compas-analysis. [↑](#footnote-ref-48)