The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities

Bo Cowgill*

December 5, 2018

Working paper

Abstract

How do judges use algorithmic suggestions in criminal proceedings? I study bail-setting in criminal cases in Broward County Florida, where judges are provided predictions of defendants' recidivism using an algorithm derived from historical data. The algorithm's output is continuous, but is shared with judges in rounded buckets (low, medium and high). Using the underlying continuous score, I examine judicial decisions close to the thresholds using a regression discontinuity design. Defendants slightly above the thresholds are detained an average extra one to four weeks before trial, depending on the threshold. Black defendants' outcomes are more sensitive to the thresholds' than white defendants. When I link jail decisions to outcomes, I find that the extra jail-time given to defendants above the thresholds corresponds to a small increase in recidivism within two years. These results suggest that algorithmic suggestions have a causal impact on criminal proceedings and recidivism.

*Thanks to Will Dobbie, Raphael Ginsburg, Sharad Goel, Catherine Tucker and participants at the 2018 NBER Economics of AI Conference.

1 Introduction

In July 2016, *ProPublica* published an explosive report about algorithms and criminal justice (Angwin et al., 2016). Using data from Florida, the authors argued that a widely-used algorithmic tool ("COMPAS") for guiding bail decisions was biased against blacks. The primary evidence was that black defendants were more likely to be classified as "high risk" for recidivism, even though black defendants were *not* more likely to recidivate. *ProPublica*'s finding was widely discussed in the media and a key example in the burgeoning algorithmic fairness literature across several academic disciplines (Larson et al., 2016).

The *ProPublica* analysis – and subsequent analysis using the same data – overlooked an important feature of the data. COMPAS scores were not an arms-length prediction post-trial outcomes. The scores were an ingredient into judicial decision-making. The scores may have affected outcomes. The *ProPublica* piece even raises this possibility to motivate their study.

Judges' use of the scores complicates the evaluation of COMPAS. Were non-recidivating, high-risk defendants truly misclassified? Or did judges intervene to affect recidivism – possibly by utilizing COMPAS' recommendation?[1]

At the center of these issues are counterfactual questions: How much are judges influenced by content of the COMPAS scores? How does this affect judicial decisions heterogeneously, and particularly along racial lines? If judges' are exceedingly deferential to COMPAS, how does this impact later life outcomes for defendants (such as recidivism)?

I examine these issues in the context of Broward County Florida's criminal court, where judges are provided guidance about defendants' recidivism risk using a predictive algorithm. The algorithm's output is continuous, but is shared with judges in rounded buckets (low, medium and high). Using the underlying continuous score, I examine judicial decisions close to the thresholds using a regression discontinuity design.

Defendants slightly above the thresholds are detained an average extra one to four weeks before trial, depending on the threshold. Black defendants' outcomes are more sensitive to the thresholds' than white defendants. When I link jail decisions to outcomes, I find that the extra jail-time given to defendants above the thresholds corresponds to a small increase in recidivism within two years of release. These results suggest that algorithmic suggestions have a causal impact on criminal proceedings and recidivism.

"Feedback loops" and "self-fulfilling prophecies" are a frequent concern in the public discourse around algorithmic bias. If an algorithm influences decisions, these decisions may effect outcomes. These outcomes are later codified into training data for future algorithms.[2]

[1] In particular, judges may have intervened by making harsher decisions. These harsher decisions may have deterred future crimes. Simply staying in jail longer would have prevented recidivism, as the defendant would have lacked the liberty necessary for additional crime.

[2] For example: Version 1.0 of an algorithm may predict that men are more likely to become successful attorneys, based on historical patterns, and may begin suggesting a company hire more men. After six months of algorithmic hiring, engineers re-train their models with the recent data. However thanks to the algorithm's deployment, the historical data has now been even further contaminated (favoring men). Version 2.0 of the algorithm will not only reflect historical bias prior to Version 1.0, but also Version 1.0's favoritism towards men. In theory, this feedback loop could iterate indefinitely towards men unless some adjustment was made to account for the algorithm's historical intervention.

Causal inferences about "algorithmic feedback loops" are inherently difficult. In many cases, individuals labeled as "greater risk of recidivism" may actually be truly more likely to be charged with future crimes – even without algorithmic labeling for judges.

Attributing this propensity to re-offend back to algorithmic intervention requires quasi-experimental variation in algorithm's deployment. This paper is, to our knowledge, the only well-identified causal evidence of the "feedback loop" phenomena. The design exploits quasi-experimental variation in the algorithm that arises from arbitrary thresholds in sentencing. In our setup, labeling a defendant "medium risk" – simply because he/she falls slightly above or below an arbitrary threshold – appears to have a causal effect on whether those defendants are charged with a later crime in the next two years.

Showing the labels to judges effects whether the original assessment was "correct" by traditional predictive-accuracy measures. Labeling a defendant "higher risk" today – possibly for arbitrary reasons – may exert a causal influence on future behavior that affects the label's accuracy.

Because of the secrecy of the COMPAS algorithm, we cannot know whether Northpointe takes feedback loop into account in training the next generation of their algorithms. Similarly, we also do not know whether doing so would have a meaningful impact on their algorithm's suggestions. Even without corrections for feedback loops, it is posible that Northpointe's suggestions are more fair than a counterfactual judge. Arnold et al. (2017) uses random assignment to judges to suggest that human bail decisions – the same decision studied by *ProPublica* – are already biased.[3]

However, a 2014 government report (Austin, 2014) to Broward County policymakers recommended that "the COMPAS system could be easily replaced with a customized risk assessment scale [...] tailored to Broward County." If this happened, the recidivism outcomes caused by COMPAS could find its way into training datasets used for future algorithms.

In addition, the feedback loop effects the conclusions drawn in academic research. The emerging computer science literature about fairness in has extensively utilized the *ProPublica* COMPAS data. This literature uses COMPAS-influenced recidivism outcomes as "ground truth" for training and evaluating new methods, rather than as contaminated data (influenced by earlier algorithmic interventions). This literature generally makes no adjustment for judges use of COMPAS in the bail decisions. Multiple earlier papers suggests that longer bailtime exerts a causal influence on defendants' outcomes, including recidivism. These papers fail to incorporate the distinction above, and thus perpetuates the feedback loop into the conclusions of research papers.

The remainder of this paper proceeds as follows. Section 1.1 discusses related literature. Section 2 discusses the empirical setting for this analysis, and Sections 3 and 4 covers the data and estimation strategy. I review the main results in Section 5 and conclude in Section 6.

1.1 Related Literature

A well-known paper by Kleinberg et al. (2017) develops an algoritm for judicial decisions. The authors then exploit the random assignment of judges to simulate a counterfactual using causal

[3]Like *ProPublica*, the Arnold et al. (2017) paper specifically examines bail decisions in county from the Miami metropolitan area.

inference methods (Lakkaraju et al., 2017). They find exactly that their algorithm, trained on historical data, "is a force for racial justice" compared to the human substitute. One policy simulation shows crime reductions "up to 24.7% with no change in jailing rates, or jailing rate reductions up to 41.9% with no increase in crime rates."

As the authors write, "in practice algorithms would be decision aids, not decision makers." By contrast, the empirical results in Kleinberg et al. (2017) simulate full algorithmic replacement of human discretion.

This paper attempts to pick up on this topic in studying judicial compliance with algorithmic risk assessments. Stevenson (2017a) examines compliance with judicial algorithms in Kentucky, where algorithmic guidance is shown to judges but do not limit judicial discretion. Using sharp pre/post variation around 2011, when algorithmic scoring in Kentucky became mandatory, she finds only small changes in average pretrial release outcomes nearly the beginning of adoption. These changes eroded over time as judges returned to their earlier habits. Berk (2017) uses a similar regression discontinuity design in the context of parole settings, and concludes that the use of algorithms "led to reductions in re-arrests for both nonviolent and violent crimes."

In Section 4, I discussion and compare the main estimands in this paper – and their interpretation – to those in related studies. The estimates in this paper are about *marginal defendants* – those on the cusp of a low/medium/high threshold. The empirical strategy identifies defendants about whom judges may be persuadable – those on the margin of a low/medium or medium/high – and isolates the impact of the pretrial algorithms on these decisions.

Other estimates in this literature examine changes in judicial decisions on different margins. In Section 4, I compare the empirical strategy of this paper (RD) to empirical strategies based on other margins. Stevenson (2017a) estimates changes for all defendants – both those around thresholds and those in the middle – around margin in 2011 in which algorithms became mandatory. The intervention in this paper also features low/medium/high bucketing. Defendants in these categories differ by as much as 25 percentage points in the probability of non-financial release before trials. This suggests there are potentially high effects for defendants near the threshold.

This paper is also related to the literature on how pre-trial detention effects defendant outcomes. Although algorithms are increasingly used in court to guide bail and pre-trial detention outcomes, the existing literature examines variation coming from non-algorithmic sources. Several recent papers (Gupta et al., 2016; Stevenson, 2017b; Leslie and Pope, 2017; Dobbie et al., 2018; Didwania, 2018) use judge leniency instruments to examine the causal effect of pretrial detention on trial outcomes. They all find that higher pre-trial assessments increase the likelihood of adverse outcomes for the defendant in the trial.

Gupta et al. (2016); Dobbie et al. (2018) examine effects on later recidivism; Gupta et al. (2016) finds no small effects, and Dobbie et al. (2018) finds some. This paper finds a similar source of exogenous variation in pre-trial sentencing, coming from algorithmic discontinuities. The paper currently has no data about effects on employment or trial outcomes, but it does find small effects on recidivism.

In addition, many researchers have found that marginal jail-time – including before and after trial – actually *causes* greater recidivism (Chen and Shapiro, 2007; Durlauf and Nagin, 2011; Rose and Shem-Tov, n.d.). To these scholars, jail-time stigmatizes the labor prospects of the incarcer-

ated, leaving them with fewer outside options. Jail thus acts as a training and recruiting site for professional criminals. Others find the opposite (Bhuller et al., 2016): That jailtime decreases the propensity for crime. Either way, these studies all suggest that jail-time has a causal impact on later recidivism – a possibility ignored by the *ProPublica* evaluation and subsequent authors.

Much of the this literature measures recidivism as being charged with a later crime. However, charges, arrests and accusations are the product of defendants' *and* law enforcement's decisions. Increases in "recidivism" may thus may *not* necessarily be caused by additional criminal behavior, but greater monitoring and enforcement by police and prosecutors.

In particular: The studies cited above suggest that additional pretrial detention leads to more guilty verdicts. These verdicts would increase a defendant's count of prior convictions. Prior convictions are publicly observable variables to law enforcement officers who are considering arresting a subject or bringing charges. In fact, the COMPAS algorithms itself directly uses prior convictions to suggest pre-trial detainment.

Even if a subject did not increase his or her criminal activity, "higher priors" may lead to greater monitoring or enforcement in way that increases probability of a future charge.

Finally, an emerging computer science literature about fairness in machine learning has studied the *ProPublica* COMPAS database in many papers. Nearly all computer science research about COMPAS examines the recidivism outcomes in the *ProPublica* dataset as ground-truth data to be used – and *not* a downstream outcome of an intervention.

Several recent computer science papers evaluate new proposed algorithms or approaches, both purely algorithmic (Zafar et al., 2017; Corbett-Davies et al., 2017) and incorporating human discretion (Tan et al., 2018; Dressel and Farid, 2018). The effectiveness of these approaches is then measured against the "ground truth" in the *ProPublica* dataset – as if these outcomes were not contaminated by Broward judges' use of COMPAS.

Similar issues arise evaluating prediction technology in other domains. For example, corporate prediction markets (Gillen et al., 2017; Cowgill and Zitzewitz, 2015) attempt to help executives forecast company outcomes. If managers utilize to these forecasts, they interact with the reality the markets are designed to predict. This changes the informational content of the prices (Siemroth, 2015). A prediction market could therefore appear "wrong" to a naive *ex-post* observer, even if it has given managers highly actionable information.[4]

2 Empirical Setting

The setting of this paper is Broward County, Florida, one of three counties in South Florida that make up the Miami metropolitan area. Broward county's population 2017 was estimated as 1.9M. According to the 2015 5-year American Community Survey, the median income for a household in Broward County was $51K, and the median income for a family was $61K. The per capita income for the county was $28K. In the 2010 Census, whites made up 42% of the population, and African-

[4]For example: Suppose a market forecasts disaster with 90% probability. Managers react to this forecast by changing their plans, thus averting the disaster. The ex-post 90% forecast may appear wrong to a naive observer because the disaster was avoided, but it was premised on the state of the world before the intervention.

Americans and Hispanic/Latinos makeup about 26% of the population each. The largest city is Fort Lauderdale. 30.9% of the county's population were foreign born.

For an overview of the pre-trail bail and detention system in the United States, see Dobbie et al. (2018). In Broward County, the use of COMPAS for pre-trial assessment began in May 2008. COMPAS was developed by a private company called Northpointe. Scores in COMPAS are based on answers of a survey containing 130+ questions.[5] The survey is completed by pre-trial services in cooperation with the defendant after his or her arrest. Part of the survey is answered based on administrative data. A sample of the COMPAS survey is available online.[6]

COMPAS does not ask directly about race, but does solicit data that may be correlated with race (educational background, employment status, gang membership, friends' arrest records and residential stability). The questionnaire also asks arrestees to agree or disagree with statements such as "A hungry person has a right to steal."

COMPAS' models were not trained on a Broward -specific subsample. The company does not disclose the details of the algorithm. Even with perfect judicial compliance, the COMPAS algorithm may not yield the results in Kleinberg et al. (2017) because of differences in how these algorithms were trained.[7]

These survey answers are inputs for algorithms that score defendants. Northpointe offered several scores, but two of its most popular are its "General Recidivism Risk" and "Violent Recidivism Risk" measures. These two scores are what I primarily analyze in this paper.[8]

In Broward County, the use of the scores are used for bail and pre-trial detention decisions. The process described in a government report by Austin (2014).

> *All of the First Appearance Court hearings are conducted via a live but limited video feed from the jail facility to the magistrate judge's chambers. [...] In a minute or two, Judge Hurley asks the defendant a set of basic questions regarding aspects of their crime, social and family relationships. He conducts a quick case review to determine release options - usually a few minutes. It is noteworthy that the Judge does not inquire about the COMPAS Risk Assessment though it is part of the packet of materials the he has available to review[.]*

Following this hearing, the bail judge sets a bail amount. The defendant is detained until he or she is able to meet bail, either through his/her own funds or through lending. Judges in most US courts do not directly assign pre-trial detention lengths, but influence it through setting either lenient or strict bail amounts.

There are many reasons to doubt whether algorithms can effectively influence judges' bail decisions. Two important conditions are required.

[5] Northpointe claims that not all of these variables are used in recidivism prediction, and some of their models use only six inputs. http://www.equivant.com/blog/official-response-to-science-advances, accessed September 4, 2018.

[6] hhttps://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html, last accessed on September 4, 2018.

[7] The Kleinberg et al. (2017) paper contains empirical results about the difference between their method and a simpler form of logistic regression.

[8] The *ProPublica* data also includes a "Failure to Appear Risk" risk score.

First, COMPAS must disagree with judges' independent instincts. If COMPAS and judges' agreed, the introduction of COMPAS would produce no change. Judge and COMPAS predictions could be highly correlated – both are influenced by historical records of sentencing and recidivism.

Several reports suggests that COMPAS and judges may independently agree. The Austin (2014) report about COMPAS in Broward notes that "questions asked by the Judge are similar to those covered by COMPAS." As the New York Times Liptak (2017) noted in its coverage of a major case about algorithmic sentencing (*State of Wisconsin v. Loomis*), "Mr. Loomis would have gotten the same sentence based solely on the usual factors." Very few papers about algorithmic-assisted decisionmaking attempt to measure or characterize the region of disagreement, in any way.

A second condition is also required: In cases of disagreement, judges must abandon their own instincts and defer to COMPAS. Human experts are unwilling to defer to algorithms in many settings (Dietvorst et al., 2015b,a; Hoffman et al., 2015; Yeomans et al., 2017; Logg, 2017; Stevenson, 2017a); this fact is often problematic for human-computer interaction designers. One sociologist and courtroom ethnographer (Christin, 2016) wrote about COMPAS-like technology,

> *During my observations, [...] risk scores were often ignored. The scores were printed out and added to the heavy paper files about defendants, but prosecutors, attorneys, and judges never discussed them. The scores were not part of the plea bargaining and negotiation process. In fact, most of judges and prosecutors told me that they did not trust the risk scores at all. Why should they follow the recommendations of a model built by a for-profit company that they knew nothing about, using data they didn't control? They didn't see the point. For better or worse, they trusted their own expertise and experience instead.*

Judges may be particularly unwilling to defer to algorithms, by comparison with other decision-makers in society. In the United States,"judicial activism" is a widely-used pejorative for judges' unwillingness to defer to law in favor of personal beliefs. The primary pre-trial judge in Broward County expressed his own personal opinions about COMPAS in the Austin (2014) report, in which he "expressed his concerns regarding the validity and utility of the COMPAS risk assessment." He also encouraged adjustments to COMPAS "to include his [personal] standards for making inmate release decisions." In Florida, the setting of the *ProPublica* study, judges are elected.[9] In theory, judges may thus face additional incentives to favor of political strategy ahead of algorithmic suggestions.

The Austin (2014) report attempts to characterize the use of COMPAS scores anecdotally based on informal observation and the judge's statements. He writes, "At of 2014 the COMPAS risk assessment tool is not being relied on" by the primary pre-trial judge, and that "COMPAS risk information is not being used by the court to make pretrial release or bail release decisions [...] even though the questions asked by the Judge are similar to those covered by COMPAS." Of course, the influence of COMPAS may be greater than is evident from anecdotal observation and self-reports.

The above evidence suggests unwillingness for judges to defer. On the other hand, a survey of judges in Virginia – which adopted risk assessment in some form as early as 2002 – found that "most judges are familiar with and embrace risk assessment as a major consideration in sentencing property and drug offenders," (Garrett and Monahan, 2018). Even if these judges were not consciously inclined to trust the risk scores, lab studies show judges influenced by irrelevant anchors (Englich et al., 2006). Some research suggests that decision-making becomes simplified when

[9] http://www.judicialselection.us/judicial_selection/index.cfm?state=FL

sequential because of exhaustion (Muraven and Baumeister, 2000).

In the case of Broward County, a single pretrial judge was responsible for nearly all bail decisions, and takes only a few minutes per case (Austin, 2014). When the primary pre-trial judge in Broward County moved into a new position, the *Florida Sun Sentinal* estimated that he had decided nearly 200,000 pre-trial bail and detention cases in approximately eight years in his job,[10] or about one hundred cases per day for three minutes each.[11]

This setting –featuring high-volumes, fast decision-making and the possibility of exhaustion – may be ripe for an anchor to influence a judge. In the analysis below, I find that judges are most responsive to the "Violent Recidivism" score, and particularly the low/medium cutoff.

Figures 1 and 2 contain examples of COMPAS reports given to judges, including the visualization of the scores. In both examples this figure is displayed in bright red, on the top of the first page – the first result of the assessment. The low/medium threshold is visually represented as a red bar crossing the halfway mark across the page. The label I find judges are second most responsive ("General Recidivism Score") is listed second, also in bright red and similarly penetrating the vertical midline. The visual salience of these items – in the context of judges who spend a few minutes per case – may explain part of their influence.

3 Data

The data used in this study was acquired by *ProPublica* (Larson et al., 2016), and was obtained mostly through public records requests. *ProPublica* requested all arrests that were scored by COMPAS over a two-year period of 2013-2014. They then matched these risk scores (using name and date of birth) to publicly available criminal records from the Florida Department of Corrections website, the Broward County Clerk's Office Website and the Broward County's Sherriff's Office.

Table 1 contains descriptive statistics about the defendants and their charges. Below, I describe the salient features for this analysis, including the original source of the records and choices about the data's preparation.

Primary Outcome Variables. The outcomes variables used in this study are the defendant's pretrial length of stay in jail, set by a judge and measured in days. This number is zero for defendants who make bail the same day, and some regressions examine the choice to send the defendant to jail at all.

Recidivism is defined as a new arrest within two years. This was based on Northpointe's practitioners guide, which says that its recidivism score is meant to predict 'a new misdemeanor or felony offense within two years of the COMPAS administration date." It was also based on a U.S. Sentencing Commission study of 25,000 federal prisoners' recidivism rates (Hunt and Dumville, 2016). This report shows that most recidivists who commit a new crime after release do so within the first two years.

[10] http://www.sun-sentinel.com/local/broward/fl-bond-court-judge-changes-20160606-story.html, last accessed September 4, 2018.

[11] Assuming eight years of service, 50 weeks per year and five days per week and seven hours per day. An average of three minutes is similar to the informal observations in Austin's 2014 report.

For violent recidivism, *ProPublica* used a definition of violent crime from the FBI's Uniform Crime Reporting (UCR) Program.[12] This includes murder, manslaughter, forcible rape, aggravated assault and robbery.

Running Variable and Thresholds. As discussed above, the COMPAS algorithms return a continuous variable. This variable is included in the supplemental database withinf *ProPublica*'s data release.[13] However, since the continuous variable was *not* used by *ProPublica*'s own analysis, it was left out of the pre-assembled datasets used by most researchers to replicate and extend the *ProPublica* analysis. However, they can be easily merged into this data.

Importantly, the continuous scores have no inherent meaning and are not interpretable. Northpointe's *Practitioners Guide to COMPAS*[14] clarifies: "It is important to note that [...] scores can only be interpreted in a relative sense." The order of the scores is supposed to be meaningful, but the numbers themselves cannot be interpreted as probabilities, number of future crimes or anything else.

To facilitate use by judges, these scores are bucketed. Northpointe first buckets the scores "by ranking the scale scores [...] then dividing these scores into ten equal sized groups." They then apply further bucket the deciles into low (deciles 1-4), medium (5-7) and high (8-10).

These divisions suggest that approximately 40% of the defendant population falls into low, and 30% fall into medium or high. This division may have been true on the original training dataset used to develop COMPAS, but is *not* true of the arrested population in Broward County, where there is greater mass on low-risk.

The origin of these thresholds in different population strengthens the claim of exogeneity. These thresholds were created based on another population, and applied to to our setting without tailoring – striking the actual distribution of Broward defendants in arbitrary and unanticipated places. Figure 3 plots a histogram of scores. Their distribution is smooth and resembles a normal distribution.

Other Covariates. ProPublica's data also contains information on each defendant's race, gender, age (in years), current charge, degree of current charge, number of prior arrests, and number of juvenile felony, misdemeanor, and other arrests.

4 Estimation Strategy

The primary identification strategy used in this paper is the regression discontinuity design (Lee and Lemieux, 2010). The unit of analysis in paper is the individual defendant, and standard errors are clustered by the defendant in all regressions.

Although Northpointe's documentation makes it clear that the labels are based on thresholds, they do not disclosed the exact value of the thresholds. In addition, the thresholds vary slightly

[12] https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/violent-crime, accessed September 4, 2018.

[13] https://github.com/propublica/compas-analysis

[14] http://www.northpointeinc.com/files/technical_documents/FieldGuide2_081412.pdf, accessed September 6, 2018.

throughout the Broward sample period. As a result, the exact threshold levels must be inferred. I have used a maximum likelihood estimator to infer the thresholds at the monthly level, and have used this in a fuzzy regression discontinuity design. As Figure 3 shows, the breaks are clear. Although the is not sharp in a literal sense, the *F*- statistics in the IV results below indicate a very strong change in labels from crossing these thresholds.

One potential threat to identification would be if the scores were manipulated, creating high densities of defendants immediately below the thresholds. This is possibly one area where the complexity and opacity of COMPAS is a benefit – this manipulation would require pretrial services officers and defendants to understand how the 130+ survey answers are turned into results.

Tests of the density around the thresholds (McCrary, 2008; Cattaneo et al., 2017) suggest it has not been manipulated. Similarly, tests of the characteristics of defendants around the thresholds show balance on observables.

I examine discontinuities using two approaches. First, I examine the discontinuities surrounding the deciles. As Figures 1 and 2 show, contains not only low/medium/high labels, but also 1-10 deciles. These deciles have also been bucketed around cutoffs described above in Section 3. Here, I pool the multiple discontinuities into a single normalized RD (a "stacked" or "normalization-and-pooled" approach). As Cattaneo et al. (2016) show, the "pooled" estimand has a complicated interpretation: It gives higher weights to values of the cutoff that are most likely to occur (i.e., cutoffs around which there are more observations). I use this approach here only to examine whether judges on average change detention decisions around the decile thresholds over the support of the data.

In addition, I estimate local average treatment effects for the low/medium thresholds for general recidivism and violent recidivism.

4.1 Interpretation of Estimand

As I discuss in Section 2, expecting judges to comply with algorithms may require bold assumptions: That judges and algorithms frequently disagree, and that judges are willing to defer to algorithms in disagreements.

Given these requirements, it may be understandable that Stevenson (2017a) finds small and eroding average effects in her study of Kentucky.

By contrast, my approaches to this research are focused on *marginal effects*. My approach identifies defendants about whom judges may be persuadable – those on the margin of a low/medium or medium/high – and isolates the impact of the pretrial algorithms on these cases. This research design does not measure the most important counterfactual for this literature: What would a judge would have done in the absence of any algorithmic guidance at all?

However, the analysis may offer some evidence of how and when judges may react to cues coming from algorithms. For example: If judges' decisions are particularly responsive to black defendants around the thresholds – but not to white defendants – this suggests that the presence or absence of scores are disproportionately influential on black defendants.

The marginal defendant in a regression discontinuity may be different than the marginal de-

fendant in a randomized controlled trial of providing algorithmic advice. However, if judges' independent instincts frequently overlap with the algorithm – even in a directional sense – then the marginal defendants in an experiment would appear only around the edges of the labels (i.e., the same marginal defendants as the regression discontinuity).

The regression discontinuity design features another benefit. It is substantially more blind to the research subjects (judges, defendants and attorneys), than either a pre/post- analysis, a difference-in-differences strategy based on staggered adoption, or a randomized controlled trial.

In these alternative designs, changes in defendants' labels are public knowledge. The research subjects may know they are being randomized into a study. Introduction of the algorithm may introduces larger shocks to the decision process, beyond randomly re-labeling defendants. The introduction of new information may require time to adjust or absorb. These issues may complicate the exclusion restriction and SUTVA requirements for causal inference.

By contrast, the regression design is a more clandestine strategy for measuring the effects of algorithmic labels in criminal courts. Because of how defendants are presented to judges ("low, medium or high"), judges and defendants may be plausibly ignorant of which candidates are above or below the underlying threshold.

Finally, the regression discontinuity allows treatment effects to be measured repeatedly over time. Every new month populates candidates around the threshold, thus permitting frequent regression discontinuity estimates. Thus the evolution of judicial response to scores can thus be measured at multiple points, which may not be possible using before/after sources of variation.

Although this paper currently aggregates over all of 2013-2014, future research may be able to quantify the evolution of judicial reaction to the scores and incorporate variation both from score discontinuities and before/after variation.

5 Results

I begin my discussion of results with the pooled regressions in Table 3. Averages increases across the pooled thresholds tend to increase pre-trial detention by approximately one week, both for the violence- and general- recidivism scores. Columns 1 and 3 include a linear slope in the normalized running variable and dummy variables for the regions surrounding each threshold. Columns 2 and 4 add an additional, segment-specific slope. The estimates are generally in the same qualitative amount – approximately one extra week of jail for the average cross of the thresholds (over the support of the data).

Next, I examine the the Low/Medium thresholds in Table 4. This table contains results from two identification approaches. In Columns 1 and 3, I use the entire sample and use the threshold as a binary instrument for receiving a medium score. The regressions include linear trends on either side of the threshold. In Columns 2 and 4, I calculate the optimal bandwidth from Imbens and Kalyanaraman (2012) and run a local regression in the neighborhood around the thresholds.

Panel A of Table 4 examines the effect on overall days in jail. Both specifications create larger effects than the pooled regressions, which suggests that the judges view the low/medium dis-

tinction as more impactful than the average decile increase. Crossing the "general recidivism" low/medium threshold causes an increase in detention of around two weeks, while crossing the same threshold for violent recidivism has a treatment effect of almost a month of additional jail-time.[15]

Panel A of Table 4 uses the same econometric approaches to examine whether a subject was detained for at least one day. The coefficients mostly rule out large effects – suggest that the effects are mostly driven by the length of pretrial detention, conditional on greater than one day. These estimates are local to the low/medium thresholds, where most may already be in jail for at least some period of time.

Table 5 examines heterogeneous effects by race. Because black and white defendants makeup 85% of the sample, I focus on this comparison using the entire sample and threshold instrument (as in Columns 1 and 3 in Table 4). For both the "general recidivism" and "violent recidivism" thresholds, the effect of black defendants passing over the threshold has a much greater magnitude. This is particularly true for the violence threshold, where black defendants crossing the threshold receive extra penalty of two months. The equivalent penalty for white defendants is not statistically significant from zero.

Finally, Table 6 examines the effect on recidivism within two years from the release. The results suggest that an additional day in pretrial detention increases two-year recidivism by a small but statistically significant amount (about one percentage point, from a baseline of 48%). The effect on violente recidivism is comparable (slightly less than 1%). These results suggest that falling slightly above the low/medium threshold not only sends defendants to jail for longer, but also causes the defendant to be arrested again in the future. I discuss this finding and its implications in the Conclusion (Section 6).

6 Discussion and Conclusion

These results suggest that – at least in the setting of Broward County Criminal court – the algorithmic guidances does affect pretrial bail decisions. The effects of crossing a threshold differ by race. What this implies about judicial bias is not clear. The race-related effects in this paper may plausibly be driven by a variety of taste-based or statistical mechanisms of discrimination. Future work should shed more light on this question.

The magitudes of effects in this paper are economically significant. While jailed, a defendant cannot maintain regular employment and may be fired. This is one reason that other research examining the causal impact of pre-trial detention finds decreased formal sector employment (Dobbie et al., 2018). Jailed defendants are also ineligible to claim unemployment insurance benefits and EITC benefits for wages earned while in jail. Several defendant advocate groups have argued that the negative effects of pre-trial detention begin as early as three days – half of the smallest effect discovered in this paper – and have therefore proposed regulations focused on this time period.[16]

[15] I separately analyzed the Low/High threshold. This threshold included in the pooled average from Table 3. However, because the data do not contain as much density in this area, estimates of Low/High specifically are much noisier.

[16] For example, the Pretrial Justice Institute has an advocacy project called "3DaysCount."

The findings about recidivism leave substantial room for interpretation about the mechanism. As previously noted, the measure of recidivism in this data – as in much of the literature – is being arrested for a future crime. One possible explanation for the this result is that marginal defendants – those just above the thresholds – engage in more criminal behavior after their release. As other researchers have noted, the time in jail may increase criminal proclivity by exposing the defendant to other possible criminals.

Or, the extra time could make it harder for a defendant to return to a life without crime. This may be particularly true if the extra pre-trial detention is accompanied by a conviction. The data for this paper do not include the ultimate outcome for the charges, although future versions of the paper may utilize this outcome data. However, several recent papers (Gupta et al., 2016; Stevenson, 2017b; Leslie and Pope, 2017; Didwania, 2018) examine the causal effect of pretrial detention using judge leniency instruments.

They all find that higher pre-trial assessments increase the likelihood of adverse case outcomes for the defendant in the trial. Such a conviction may damage the defendant's future employability, making extra-legal activity more tempting in the future. These results are consistent with pre-trial detention weakening defendants' bargaining position during plea negotiations, and criminal convictions harming defendants' prospects in the legal job market.

The recidivism results may also come about through police and prosecutorial mechanisms. The defendant does not need to behave more criminally in order to be accused of more crimes. The results in this paper could also come about through greater police monitoring or enforcement by prosecutors. This may be particularly likely given the earlier results linking pre-trial detention and convictions.

Convictions are public information that police, prosecutors and juges can view. Suppose law enforcement engages in greater monitoring and enforcement for citizens with more prior convictions. This would not be surprising given public comments by law enforcement and patterns in observational data. Then the extra arrests and convictions coming from COMPAS' threshold would create greater monitoring and enforcement against the affected individuals. This may lead to higher arrests – even if the defendants' underlying criminal behavior has not changed. This paper cannot currently speak to the mechanism of the recidivism result, and thus this is an area for future research.

References

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks," *ProPublica, May*, 2016, *23*.

Arnold, David, Will Dobbie, and Crystal S Yang, "Racial bias in bail decisions," *The Quarterly Journal of Economics*, 2017.

Austin, James, "Evaluation of Broward County Jail Population: Current Trends and Recommended Options," 2014.

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

Bhuller, Manudeep, Gordon B Dahl, Katrine V Løken, and Magne Mogstad, "Incarceration, recidivism and employment," Technical Report, National Bureau of Economic Research 2016.

Cattaneo, Matias D, Michael Jansson, and Xinwei Ma, "Simple local polynomial density estimators," *University of Michigan, Working Paper*, 2017.

_ , Rocío Titiunik, Gonzalo Vazquez-Bare, and Luke Keele, "Interpreting regression discontinuity designs with multiple cutoffs," *The Journal of Politics*, 2016, *78* (4), 1229–1248.

Chen, M Keith and Jesse M Shapiro, "Do harsher prison conditions reduce recidivism? A discontinuity-based approach," *American Law and Economics Review*, 2007, *9* (1), 1–29.

Christin, Angèle, "Models in Practice," *Points: Data and Society*, 2016.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq, "Algorithmic decision making and the cost of fairness," in "Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" ACM 2017, pp. 797–806.

Cowgill, Bo and Eric Zitzewitz, "Corporate prediction markets: Evidence from google, ford, and firm x," *The Review of Economic Studies*, 2015, *82* (4), 1309–1341.

Didwania, Stephanie Holmes, "The Immediate Consequences of Pretrial Detention: Evidence from Federal Criminal Cases," 2018.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err.," *Journal of Experimental Psychology: General*, 2015, *144* (1), 114.

_ , _ , and _ , "Overcoming Algorithm Aversion: People Will Use Algorithms If They Can (Even Slightly) Modify Them," *Available at SSRN 2616787*, 2015.

Dobbie, Will, Jacob Goldin, and Crystal S Yang, "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, 2018, *108* (2), 201–40.

Dressel, Julia and Hany Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science advances*, 2018, *4* (1), eaao5580.

Durlauf, Steven N and Daniel S Nagin, "Imprisonment and crime," *Criminology & Public Policy*, 2011, *10* (1), 13–54.

Englich, Birte, Thomas Mussweiler, and Fritz Strack, "Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making," *Personality and Social Psychology Bulletin*, 2006, *32* (2), 188–200.

Garrett, Brandon L and John Monahan, "Judging Risk," 2018.

Gillen, Benjamin J, Charles R Plott, and Matthew Shum, "A Pari-Mutuel-Like Mechanism for Information Aggregation: A Field Test inside Intel," *Journal of Political Economy*, 2017, *125* (4), 1075–1099.

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

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li, "Discretion in hiring," Technical Report, National Bureau of Economic Research 2015.

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

Imbens, Guido and Karthik Kalyanaraman, "Optimal bandwidth choice for the regression discontinuity estimator," *The Review of Economic Studies*, 2012, *79* (3), 933–959.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human decisions and machine predictions," Technical Report, National Bureau of Economic Research 2017.

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

Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin, "How we analyzed the COMPAS recidivism algorithm," *ProPublica (5 2016)*, 2016.

Lee, David S and Thomas Lemieux, "Regression discontinuity designs in economics," *Journal of economic literature*, 2010, *48* (2), 281–355.

Leslie, Emily and Nolan G Pope, "The Unintended Impact of Pretrial Detention on Case Outcomes: Evidence from New York City Arraignments," *The Journal of Law and Economics*, 2017, *60* (3), 529–557.

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

Logg, Jennifer Marie, "Theory of Machine: When Do People Rely on Algorithms?," 2017.

McCrary, Justin, "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of econometrics*, 2008, *142* (2), 698–714.

Muraven, Mark and Roy F Baumeister, "Self-regulation and depletion of limited resources: Does self-control resemble a muscle?," *Psychological bulletin*, 2000, *126* (2), 247.

Rose, Evan and Yotam Shem-Tov, "Does Incarceration Increase Crime?"

Siemroth, Christoph, "The informational content of prices when policy makers react to financial markets," *Browser Download This Paper*, 2015.

Stevenson, Megan, "Assessing Risk Assessment in Action," *George Mason Law & Economics Research Paper*, 2017, (17-36), 4.

_ , "Distortion of justice: How the inability to pay bail affects case outcomes," 2017.

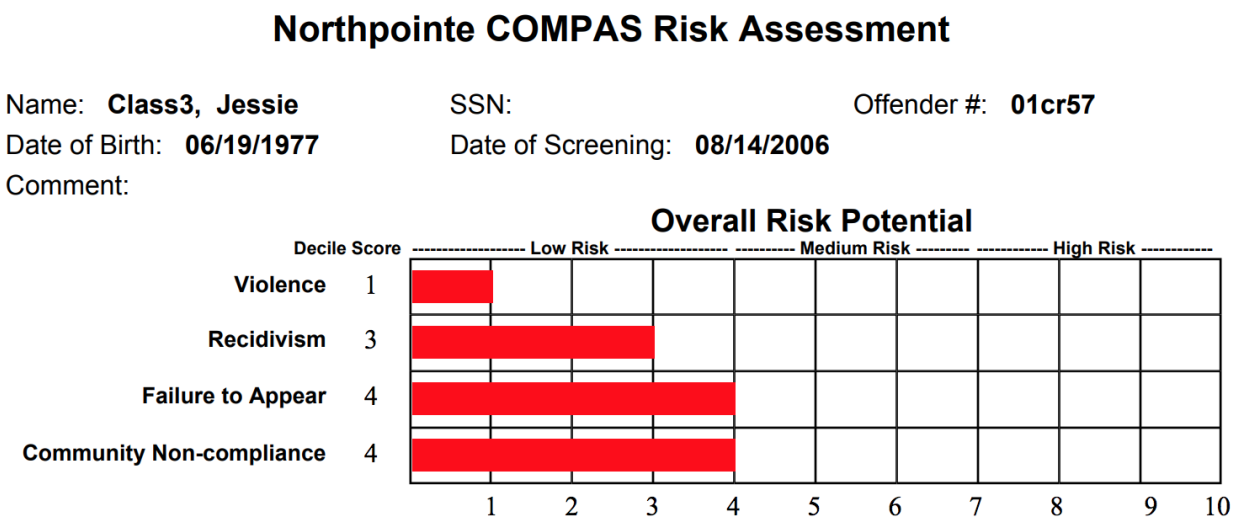
Tan, Sarah, Julius Adebayo, Kori Inkpen, and Ece Kamar, "Investigating Human+ Machine Complementarity for Recidivism Predictions," *arXiv preprint arXiv:1808.09123*, 2018.

Yeomans, Michael, A Shah, Sendhil Mullainathan, and Jon Kleinberg, "Making sense of recommendations," 2017.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in "Proceedings of the 26th International Conference on World Wide Web" International World Wide Web Conferences Steering Committee 2017, pp. 1171–1180.

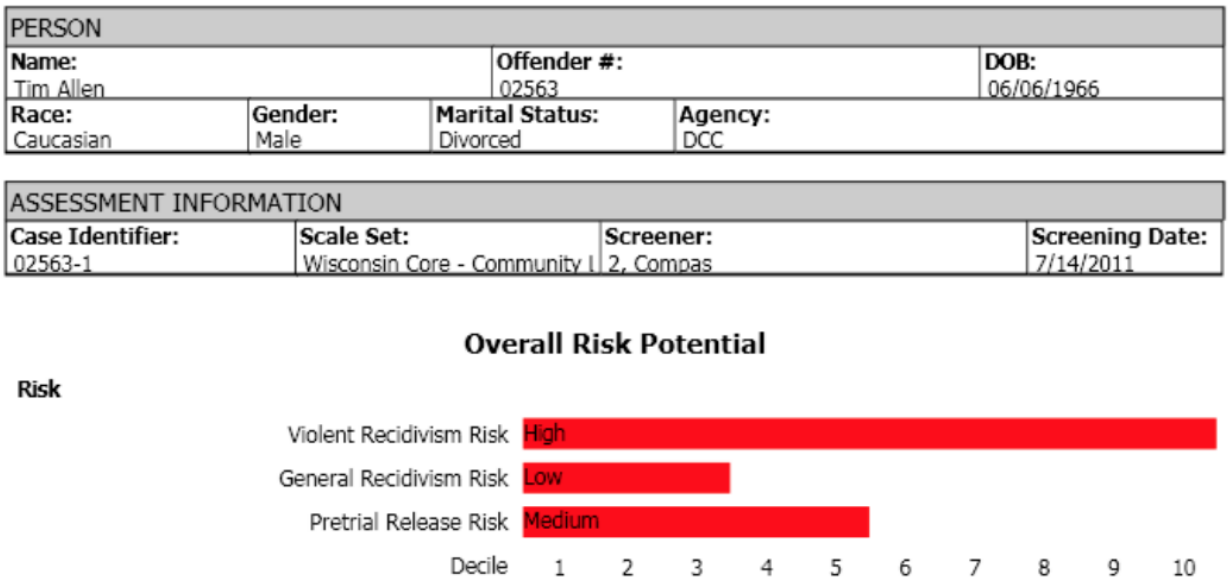
Tables and Figures

Figure 1: Sample of COMPAS Output: Michigan



Notes: The above is a screenshot of a sample report shown to judges in Michigan criminal court. An example of a complete report can be viewed at: https://www.michigan.gov/documents/corrections/Timothy_Brenne_Ph.D._Jessie_Report_297502_7.pdf

Figure 2: Sample of COMPAS Output: Wisconsin



PERSON

Name:	Offender #:	DOB:
Tim Allen	02563	06/06/1966

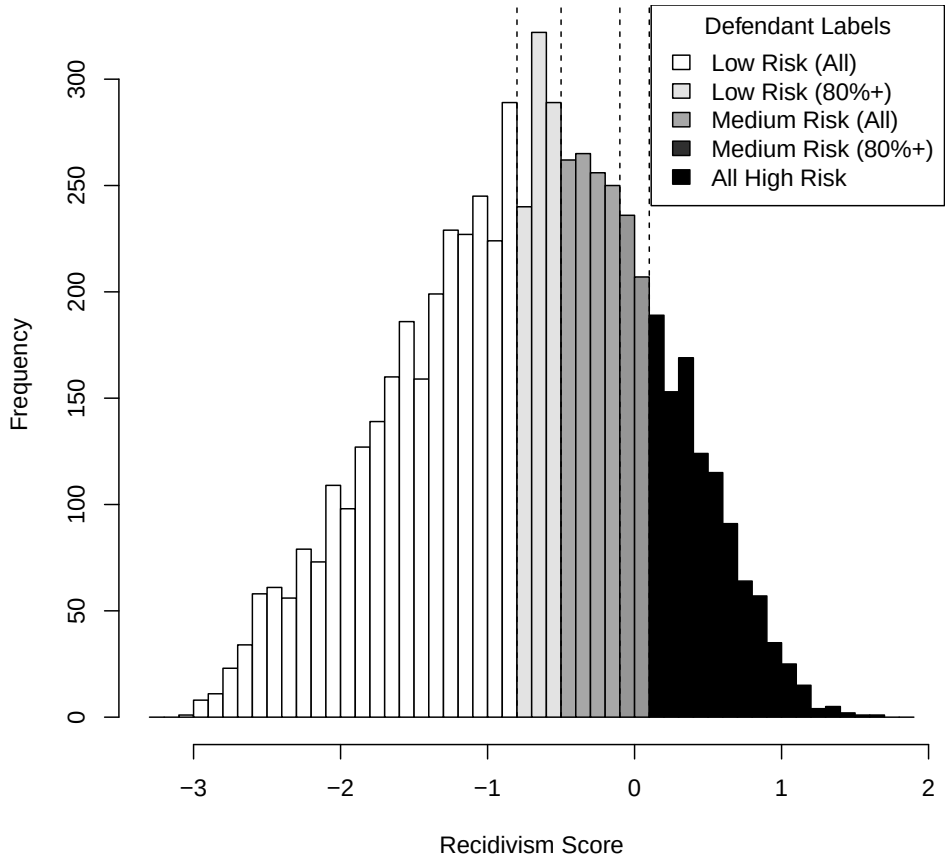
Race:	Gender:	Marital Status:	Agency:
Caucasian	Male	Divorced	DCC

ASSESSMENT INFORMATION

Case Identifier:	Scale Set:	Screener:	Screening Date:
02563-1	Wisconsin Core - Community I	2, Compas	7/14/2011

Notes: The above is a screenshot of a sample report shown to judges in Wisconsin criminal court. An example of a complete report can be viewed at:https://docs.legis.wisconsin.gov/misc/lc/study/2016/1495/030_august_31_2016_meeting_10_00_a_m_room_412_east_state_capitol/doc_responses

Figure 3: Histogram of Recidivism Scores and Defendant Labels



Notes: COMPAS scores have no interpretation without Northpointe's labeling system of low, medium and high.

Table 1: Descriptive Statistics: Defendants

	Mean	SD
Male	0.810	0.393
Black	0.514	0.500
White	0.341	0.474
Hispanic	0.082	0.275
Asian	0.005	0.071
Native American	0.002	0.042
Other Race	0.056	0.229
Age (Years)	32.478	11.707
Unmarried	0.885	0.319
Priors	3.246	4.744
Juniville Priors (All)	0.261	0.928
Recidivates	0.484	0.500
Violently Recidivates	0.112	0.316
Observations	6172	

Notes: This table contains descriptive statistics about the defendants in the sample of this paper. All defendants were charged and arrested in Broward County, Florida in 2013-2014. For descriptive statistics about the charges, see Table 2.

Table 2: Descriptive Statistics: Crimes

	Mean	SD
Felony	0.643	0.479
Battery or Assault	0.270	0.444
Theft, Robbery or Burglary	0.153	0.360
No Charge	0.127	0.333
Posession	0.126	0.332
Driving or Vehicle-related	0.103	0.304
Cocaine-related	0.077	0.267
DUI	0.047	0.212
Cannabis-related	0.046	0.210
Opioid-related	0.030	0.171
Weapon-related	0.029	0.168
Domestic-related	0.027	0.162
Childen or Minors-related	0.019	0.136
Disorderly or Lews Conduct	0.011	0.105
Sex-related	0.001	0.036
Murder or Manslaughter	0.000	0.022
Observations	6172	

Notes: This table contains descriptive statistics about the criminal charges in the sample of this paper. All defendants were charged and arrested in Broward County, Florida in 2013-2014. For descriptive statistics about the defendeats, see Table 1.

Table 3: Pooled Regression Discontinuity Estimates on Pre-Trial Detention Length

	Jail Days	Jail Days	Jail Days	Jail Days
Violent Recidivism Thresholds (Pooled Deciles)	4.896*	6.174*		
	(2.640)	(3.605)		
General Recidivism Thresholds (Pooled Deciles)			7.393**	6.477*
			(3.024)	(3.818)
F-stat	38709.0	35096.4	2257.9	1994.7
Controls	No	Yes	No	Yes
Observations	5469	5469	5284	5284

Notes: This table contains pooled fuzzy-RD results from all decile thresholds. I pool the multiple discontinuities into a single normalized RD (a "stacked" or "normalization-and-pooled" approach). Columns 1 and 3 include a linear slope in the normalized running variable and dummy variables for the regions surrounding each threshold. Columns 2 and 4 add an additional, segment-specific slope. Standard errors are clustered by defendant.

Table 4: Regression Discontinuity Estimates on Pre-Trial Detention: Low/Medium Thresholds

Panel A: Outcome = Length of Stay (days)

	Jail Days	Jail Days	Jail Days	Jail Days
Low/Medium Recidivism Threshold	13.94**	14.39**		
	(6.337)	(7.311)		
Low/Medium Violence Threshold			37.56***	43.24***
			(12.26)	(13.03)
F-stat	356.1	274.7	190.5	171.6
Sample	Global	Local	Global	Local
Observations	6172	6068	6172	4904

Panel B: Outcome = Length of Stay > 0

	Jailed	Jailed	Jailed	Jailed
Low/Medium Recidivism Threshold	-0.0463	-0.708		
	(0.0635)	(0.584)		
Low/Medium Violence Threshold			0.0657	0.177
			(0.106)	(0.181)
F-stat	356.1	5.810	190.5	139.3
Sample	Global	Local	Global	Local
Observations	6172	5024	6172	3650

Notes: These regressions present fuzzy regression discontinuity results about the Low/Medium threshold. In Columns 1 and 3, I use the entire sample and use the threshold as a binary instrument for receiving a medium score. The regressions include linear trends on either side of the threshold. In Columns 2 and 4, I calculate the optimal bandwidth from Imbens and Kalyanaraman (2012) and run a local regression in the neighborhood around the thresholds. Standard errors are clustered by defendant.

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 5: Regression Discontinuity Estimates by Race

	Jail Days	Jail Days	Jail Days	Jail Days
Low/Medium Recidivism Threshold	7.511	24.51		
	(11.56)	(15.29)		
Low/Medium Violence Threshold			-19.64	61.76***
			(20.40)	(16.14)
F-stat	138.8	51.66	51.40	102.5
Racial Sub-Sample	White	Black	White	Black
Observations	2103	3175	2103	3175

Notes: This table presents fuzzy regression discontinuity results about the Low/Medium threshold. All regressions use the entire sample and use the threshold as a binary instrument for receiving a medium score. The regressions include linear trends on either side of the threshold. Standard errors are clustered by defendant.

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 6: Recidivism Outcomes

	Recidivate	Violent Recidivate
Length of Stay	0.00991**	0.00492**
	(0.00410)	(0.00236)
F-stat	15.62	15.62
Observations	6172	6172

Notes: This table present fuzzy regression discontinuity results about the Low/Medium threshold. All regressions use the entire sample and use the threshold as a binary instrument for the length of the sentence. The regressions include linear trends on either side of the threshold. Standard errors are clustered by defendant.

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.