The Hidden Effects of Algorithmic Recommendations

Alex Albright[*]

July 31, 2023[†]

Abstract

Algorithms inform human decisions in many high-stakes settings. They provide decision-makers with predictions concerning the probability of an event. However, there is typically an additional step involved: decision-makers are recommended particular decisions based on the predictions. I isolate the causal effects of these algorithmic recommendations by leveraging a setting in which the recommendations given to bail judges changed, but the algorithm's predictions given to judges did not. Recommendations significantly impacted decisions: lenient recommendations increased lenient bail decisions by over 50% for marginal cases. I explore possible mechanisms behind this effect and provide evidence that recommendations can change the costs of errors to decision-makers. Judges may be more lenient when their choices are consistent with recommendations because the recommendation can shield them from political backlash. Finally, I show that variation in adherence to recommendations complicates how algorithm-based systems affect racial disparities. Judges are more likely to deviate from lenient recommendations for Black defendants than for white defendants with identical algorithmic risk scores.

[*] Federal Reserve Bank of Minneapolis, Opportunity and Inclusive Growth Institute
(Email: alex@albrightalex.com & Website: albrightalex.com)

I thank Ed Glaeser, Larry Katz, and Winnie Van Dijk for consistent support and helpful feedback on this project. I also thank Megan Stevenson, Jennifer Doleac, Andrew Goodman-Bacon, Alicia Modestino, Elior Cohen, Amanda Michaud, Mandy Pallais, Crystal Yang, Alma Cohen, Emma Harrington, Hannah Shaffer, Natalia Emanuel, Jon Roth, Ashesh Rambachan, Liz Engle, Justin Bloesch, Lauren Taylor, Karen Shen, Namrata Narain, Louis Kaplow, Jason Furman, Christina Ciocca Eller, Adam Soliman, Evan Soltas, Emma Pierson, Kadeem Noray, Brian Wheaton, John Tebes, Roland Fryer, Philip Marx, LJ Ristovska, Tianwang Liu, Sarah Armitage, and presentation audiences at Harvard, LSU, Macalester, Williams, St. Olaf, ECINEQ, WEC Jr., Wiser, and SOLE for constructive discussions. I am grateful to Daniel Sturtevant, Tara Blair, Christy May, and Kathy Schiflett for sharing the data used in this paper and their patience in explaining institutional details about Kentucky Pretrial Services. I thank Amisha Kambath for excellent research assistance. This research has been supported by a Stone PhD Fellowship from the Harvard Inequality & Social Policy Program and a Considine Fellowship from the Olin Center at Harvard Law School. The views expressed in this paper are my own and do not necessarily represent those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

[†] This paper subsumes and replaces an earlier draft called "If You Give a Judge a Risk Score" (2019).

1 Introduction

Algorithms are used in allocating affordable housing, making loan decisions, setting bail in the justice system, and many other high-stakes settings. However, in most contexts, humans – not algorithms – still make the final decisions. Therefore, understanding how algorithms change human decision-making is crucial for policy.

How algorithms change human decisions depends on how they are communicated to decision-makers. Algorithms provide decision-makers with predictions about the probability of an event (e.g., loan default, pretrial misconduct). However, there is typically another step in this process that is often overlooked: decision-makers are given recommended decisions based on the algorithm's predictions ("algorithmic recommendations"). Predictions and recommendations are distinct, but research and press coverage about how algorithms shape outcomes usually conflate the two. Therefore, attempts to estimate the effects of algorithms can muddle the effects of providing a new prediction technology with the effects of setting normative recommendations. In this paper, I disentangle these two to isolate the hidden effects of algorithmic recommendations on human decisions.

It is an empirical challenge to disentangle the effects of algorithmic recommendations from the effects of algorithmic predictions because the two are usually implemented simultaneously. I make progress on this front by leveraging a setting in which algorithmic predictions given to decision-makers stayed the same, but the use of algorithmic recommendations changed. I highlight the importance of algorithmic recommendations by demonstrating their independent causal effects on high-stakes human decisions.

My empirical setting covers bail decisions in Kentucky from 2011 to 2013. Judges making bail decisions received an algorithm's predicted risk of pretrial misconduct for each case. Before June 2011, judges observed risk scores but were not given recommendations. However, in June 2011, judges began receiving recommendations, but only for cases with risk scores below a sharp cut-off. Judges were recommended to not set money bail for cases with low or moderate risk scores. Not setting money bail is a more lenient decision than setting money bail because the latter requires defendants post money for release from jail, while the former does not. Therefore, I call the recommendation introduced in 2011 a "lenient bail" recommendation.

Cases with the highest moderate risk scores received lenient recommendations, but similar cases with the lowest high risk scores did not. To isolate the effect of the lenient recommendation, I use a differences-in-differences approach in which the control group is cases with scores slightly above the critical threshold and the treated group is cases with scores slightly below the threshold. This method focuses on the effect of the lenient recommendation for marginal moderate risk cases, those that are close to the moderate-high threshold.

I find that algorithmic recommendations meaningfully change human decisions. The lenient recommendation increased lenient decisions by 50%-70% (or 9-13 percentage points) for marginal cases. These is no evidence of pre-trends, and results are nearly identical regardless of which controls (if any) are included in the specifications. Across the full low and moderate risk distribution, the recommendation increased lenient decisions by a similar magnitude, between 10 and 15 percentage points. While the algorithm's predictions given to judges stayed the same, recommendations impacted decisions, making judges more lenient than they were before. Algorithmic recommendations, therefore, have first-order impacts on decision-making and outcomes.

Why do these recommendations change decision-maker behavior? In the bail context, judges make decisions based on perceived probabilities of pretrial misconduct (a bad outcome) as well as their perceived costs of pretrial detention and misconduct. I investigate two possible mechanisms: (1) recommendations nudge decision-makers to put more weight on the algorithm's predictions, and (2) recommendations change private costs of errors to decision-makers (i.e., making a lenient decision that results in misconduct). If the predictions mechanism is correct, then judge predictions should change for cases with all types of risk scores. But if the error cost mechanism is correct, then judge error costs should change only for low and moderate risk cases (those covered by the recommendation). Therefore, the two mechanisms generate distinct testable predictions: the predictions mechanism lowers lenient bail rates for high risk cases, while the error cost mechanism does not change lenient bail rates for high risk cases. I find evidence in favor of the error costs mechanism. Intuitively, judges perceive lower costs to lenience if they are covered by a lenient recommendation. If their lenience results in misconduct, part of the blame goes to the people who set the recommendation, rather than all of the blame going to the judge. Thus, algorithmic recommendations can change decision-maker pay-offs – there is not free disposal of the recommendation.

Finally, I explore heterogeneity in the effects of algorithmic recommendations and find that lenient recommendations do not benefit all groups equally. Observed Black-white gaps in lenient bail were almost three times as large as they would have been if judges had perfectly complied with algorithmic recommendations: white defendants were 9.3 percentage points more likely to receive lenient bail than Black defendants, instead of

3.3 percentage points more likely. This is not fully explained by differences in risk score distributions. I show that judges deviate from lenient recommendations more frequently for Black defendants than for white defendants with identical underlying algorithmic risk scores. Moreover, I estimate differences-in-differences specifications separately for cases with white and Black defendants and find that Black defendants benefit less from a lenient recommendation. This is driven primarily by differences across judges: judges who make decisions for populations with more Black defendants are less likely to respond to the lenient recommendation. This is consistent with judges thinking recommendations provide them less political cover in more racially heterogeneous places. These results demonstrate how differential adherence to recommendations complicates how algorithmic systems impact racial disparities.

Algorithms have been shown to outperform human decision-makers in a variety of settings (Berk 2017; Mullainathan and Obermeyer 2022; Kleinberg et al. 2017; Cowgill 2018a). However, because human decision-makers still retain discretion in most settings, it is necessary to study how algorithms and humans interact. For this reason, recent work has compared outcomes in the absence of algorithms with outcomes when algorithms are used at the discretion of humans (Sloan, Naufal, and Caspers 2018; Doleac and Stevenson 2018; Garrett and Monahan 2018; DeMichele et al. 2018; Cowgill and Tucker 2019; Agarwal et al. 2023). How the algorithms' predictions are communicated to human decision-makers varies across these settings; the existing evidence does not distinguish between the effects of new predictions and new algorithmic recommendations. I contribute to this literature by disentangling the effects of predictions from the effects of recommendations.

The 2011 reform in Kentucky that I leverage for identification was first studied by Stevenson (2018). Her results highlighted the distinction between how algorithms change outcomes in theory and in practice. I diverge from and build on her important work by illuminating the distinction between algorithm predictions and recommendations. I use administrative court reports to identity a time period in which the algorithm predictions stayed constant but the recommendations changed. I also use internal documents to calculate the underlying risk scores for cases, which allows me to isolate the causal effects of algorithmic recommendations by focusing on cases near the critical threshold.

The paper that best formalizes the distinction between algorithm predictions and recommendations is McLaughlin and Spiess (2022). McLaughlin and Spiess (2022) develop a model in which algorithmic recommendations may change preferences, rather than just beliefs. The theory is in their paper is similar to how I describe recommendations changing the costs of errors, rather than just changing predictions. McLaughlin and Spiess (2022) show why recommendations matter independent of predictions in theory, while my paper provides direct empirical evidence of their independent effects in practice.

Humans have been shown to use their own discretion in implementing algorithmic tools. Ethnographic work demonstrates a "decoupling" between how algorithms are expected to be used and how they are actually used in practice (Christin 2017). Decision-makers frequently overrule algorithm recommendations (Hoffman, Kahn, and Li 2017; Gruber et al. 2020; Pruss 2023), and research has shown decision-makers may respond to algorithms differently according to the age or socioeconomic status of the people about whom they are making decisions (Skeem, Scurich, and Monahan 2019; Doleac and Stevenson 2018). I contribute to this literature by showing that judges frequently deviate from algorithmic recommendations and their deviation rates vary with their defendant population. Judges who work in places with more Black defendants deviate from lenient recommendations more often.

This paper contributes to a broad literature on algorithms and group inequality. The interactions between algorithms and racial inequality, in particular, have been studied in popular writing (Angwin et al. 2016; Alexander 2018), computer science (Lum and Isaac 2016; Corbett-Davies et al. 2017), and, increasingly, economics (Cowgill and Tucker 2019; Kleinberg et al. 2018; Kleinberg and Mullainathan 2019). My paper diverges from the majority of related work because I focus on the role of algorithmic recommendations rather than the underlying prediction technology. However, my results still align with a number of previous papers using criminal justice system data. First, like Doleac and Stevenson (2018), I show judges may be more lenient for white defendants than for Black defendants with the same algorithmic score. Second, like Cowgill (2018b), I demonstrate that outcomes for Black and white defendants are not equally responsive to crucial thresholds in the algorithmic score distributions.

Finally, the results in this paper help inform policy issues related to algorithms and human decisions. In 2019, Obermeyer et al. (2019) found that a healthcare algorithm was "less likely to refer black people than white people who were equally sick to [programs] that aim to improve care." The company responsible for the algorithm's predictions and recommendations replied that the algorithm's recommendation is "just one of many data elements intended to be used to select patients for clinical engagement programs." The company argued that doctors' choices may not exhibit the same racial bias because doctors consider more than just algorithmic recommendations when making decisions. However, my results show that algorithmic recommendations have strong causal effects on human decisions. Thus, discovered bias in an algorithmic recommendation is likely to impact final human decisions.

The remainder of the paper proceeds as follows. Section 2 provides background on different algorithmic decision-making environments and highlights the importance of algorithmic recommendations. Section 3 describes my empirical setting and the relevant administrative data. Section 4 estimates the causal effects of algorithmic recommendations. Section 5 provides evidence on the mechanism behind these effects. Section 6 demonstrates how heterogeneous adherence to recommendations can have unintended effects on racial gaps. Section 7 concludes.

2 Algorithmic Systems and Bail Decision-Making

2.1 A Spectrum of Algorithmic Decision-Making Systems

How do algorithms change decisions? The answer depends on the differences between the status quo decision-making system and the new algorithm-based decision-making system. Algorithm-based decision-making systems vary widely. They include systems in which algorithm-based rules automatically make decisions as well as systems in which humans are given an algorithm's predictions with no direction on how to use them. There is no singular algorithmic decision-making system – there is a spectrum of them.[1]

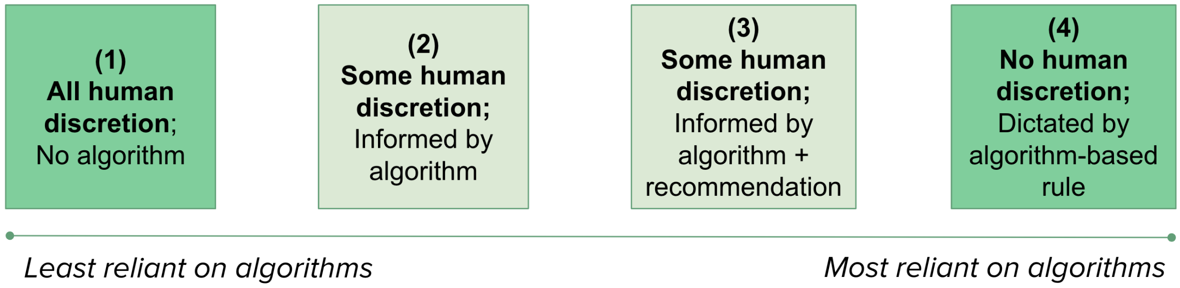
When studying the effects of algorithms, researchers often contrast a world with human discretion and no algorithms to one with algorithms and no human discretion. While prior research has shown that algorithms can outperform human decision-makers (Berk 2017; Mullainathan and Obermeyer 2022; Kleinberg et al. 2017; Cowgill 2018a), this is often not the relevant comparison. Humans still usually make final decisions even when algorithms are present.

An evolving literature compares outcomes in the absence of algorithms to outcomes when algorithms are used at the discretion of human decision-makers (Sloan, Naufal, and Caspers 2018; Stevenson 2018; Doleac and Stevenson 2018; Garrett and Monahan 2018; DeMichele et al. 2018; Cowgill and Tucker 2019). How algorithms are integrated into human decision-making varies across these settings. Some settings provide decision-makers

[1] Congress's Algorithmic Accountability Act defines an "automated decision-making system" as "a computational process, including one derived from machine learning, statistics, or other data processing or artificial intelligence techniques, that makes a decision or facilitates human decision making." This definition includes the spectrum of possible decision-making environments I discuss in this section.

with an algorithm's predictions only, while others provide algorithmic recommendations that suggest explicit decisions based on the predictions. In this paper, I show that algorithmic recommendations are an important feature of algorithm-based decision-making systems. They have independent effects on decisions and merit their own focus in policy discussions and research.

Figure 1: Spectrum of Algorithm-Based Decision-Making Systems


(1)
All human discretion;
No algorithm
(2)
Some human discretion;
Informed by algorithm
(3)
Some human discretion;
Informed by algorithm + recommendation
(4)
No human discretion;
Dictated by algorithm-based rule
Least reliant on algorithms
Most reliant on algorithms

Notes: This figure illustrates a theoretical spectrum of algorithmic systems. There are four settings illustrated. Going from left to right, they are ordered from least to most reliant on algorithms.

In Figure 1, I illustrate a spectrum of algorithm-based decision-making systems to make explicit the differences across potential settings. I list four settings, ordered from least to most reliant on algorithms from left to right. In dark green, on the ends, are the two extremes, (1), all human discretion and no algorithms, and (4), no human discretion with outcomes dictated by an algorithm-based rule. In the middle, in light green, are the two intermediate settings, in which humans have the final say, but the context differs. In (2), human decisions are informed by an algorithm (its predictions), but there are no algorithmic recommendations. In (3), there are also algorithmic recommendations.

Using the spectrum shown in Figure 1, I can situate the previously referenced literature spatially. Research showing that algorithms alone can outperform human decision-makers contrasts (1) with (4), while research showing how outcomes change when humans are given algorithms but have discretion contrasts (1) with either (2) or (3). My paper contributes to this ecosystem of papers by estimating the distinction between (2) and (3), highlighting the understudied and hidden effects of algorithmic recommendations implicit in previous research.

2.2 Algorithms and Recommendations in Bail Decision-Making

Algorithms in the criminal justice system are prevalent and varied. They are used in pretrial risk assessment, sentencing, prison management, and parole. In a survey of state practices, the Electronic Privacy Information Center (2020) found dozens of different algorithms used in criminal justice systems across the country – every state uses one in some capacity. In general, these algorithms make predictions of types of risk according to individual-level and case-level characteristics. For example, the Public Safety Assessment, which is used in over 40 counties, calculates pretrial misconduct risk by adding up integer weights based on 9 risk factors (Laura and John Arnold Foundation 2018). The tool derives these weights by regressing misconduct measures on a slate of case-level characteristics in a dataset of 750,000 observations (Laura and John Arnold Foundation 2018). Meanwhile, the more complicated COMPAS algorithm has hundreds of inputs and is a black-box machine learning model.

In the pretrial setting, algorithms are used to help make bail decisions. After arrest, judges make decisions about how to set bail for the arrested person. The bail decision stipulates the conditions the arrested person must meet to be released from jail. The legal objective of bail is to set the lowest possible bail to ensure court appearance and public safety. In this context, algorithms are designed to predict the risk of pretrial misconduct (failing to appear in court or re-arrest).

While algorithms can vary greatly in how they make these misconduct predictions, they share a commonly stated goal. The goal is to provide a "data-driven way to advance pretrial release." In other words, the goal is to reduce judges' prediction errors and therefore allow more people to be released without compromising on misconduct.[2] Prior research on risk assessments in bail settings highlights their potential in this regard. Kleinberg et al. (2017) find that jail populations could be reduced by 42%, with no change in crime rates, by using algorithms in bail setting. A strictly preferred combination of jailing and crime rates is possible simply through better (algorithm-based) decision-making. However, algorithms' predictions give information about the ranking of individual cases by risk; they do not give information about which jailing rate judges should pick.

Bail is not a good in fixed supply that judges simply allocate. Rather, judges pick the rate of bail setting (i.e., what percentage of the population receives money bail). This is in stark contrast to other high-stakes environments, in which algorithms play a role only in allocation. For instance, in the coordinated entry system, people are scored (according to

[2] In economists' language, algorithms are meant to help more efficiently allocate detention.

housing need or housing readiness) and then ordered on a list by their score. The available housing is then allocated down the list by score until the housing runs out. The supply of housing in that context is fixed; the algorithm has no capacity to change that margin. In contrast, in the bail system, changing decision-making environments can change allocation (who gets which bail decisions) and the overall rate of bail settings (what percentage of defendants receive money bail).

2.3 A Theory of Bail Decisions across Algorithmic Environments

To fix ideas, I outline a simple canonical framework to demonstrate how judges set bail. I then complicate the framework by introducing algorithms and algorithmic recommendations.

No algorithms or algorithmic recommendations: Each judge j chooses whether to release or detain defendants. The judge chooses what to do on the basis of which action is less costly. Detention has cost c_{jd}, which is specific to judge j because judges perceive different costs of detention. The cost is not specific to defendant i. If defendant i is released, there is some probability of pretrial misconduct (failure to appear or re-arrest). Judge j perceives this probability for defendant i as $P_{ji}(m)$ and perceives the cost of misconduct as c_{jm}. Judge j will then release defendant i if and only if $c_{jd} > P_{ji}(m)c_{jm}$ (i.e., the cost of detention to the judge is higher than the cost of release to the judge).

Therefore, for the full set of N cases for which judge j makes decisions, I can illustrate a judge's decision rule by ordering cases based on the judge's perceived probabilities of misconduct, $P_{ji}(m)$. In Figure 2a, I index and order cases so that $P_{ji}(m) < P_{j2}(m) < ... < P_{jN-1}(m) < P_{jN}(m)$. Thus, the judge's decision in case i hinges on which side of the critical threshold $\frac{c_{jd}}{c_{jm}}$ perceived risk for case i is on. All cases with risk lower than the threshold – illustrated as to the left of the dotted vertical line – will be released, while the rest will be detained.

Adding algorithms: Now, I complicate the set-up by introducing algorithms into the judge's decision-making process. There is some algorithm that predicts probabilities of misconduct, $P_{ai}(m)$, based on case-level characteristics for each case i. However, these probabilities are not provided to the judges to inform decisions. Rather, the probabilities are mapped to scores $s_{ai}(m)$, which provide relative risk information rather than absolute risk information. I make this assumption to fit with common practice in the real world. (For instance, one case may have a risk score of 8, while another has a risk score of 3. It is not clear what probabilities of misconduct these entail; however, it is clear that 8 is larger than 3, and thus the case with that score is riskier.) Judge j then updates their risk prediction based on this new score information, $s_{ai}(m)$. The new risk prediction is $P'_{ai}(m)$, and the fraction of cases such that $P_{ji}(m) < \frac{c_{jd}}{c_{jm}}$ is the same as the fraction of cases such that $P'_{ji}(m) < \frac{c_{jd}}{c_{jm}}$. In words, the release rate for each judge is the same because the algorithmic scores can change the ordering of predicted probabilities, but they do not change absolute perceptions of risk.[3] Figure 2b depicts how algorithms may change the ordering of cases but retain the same release rate.

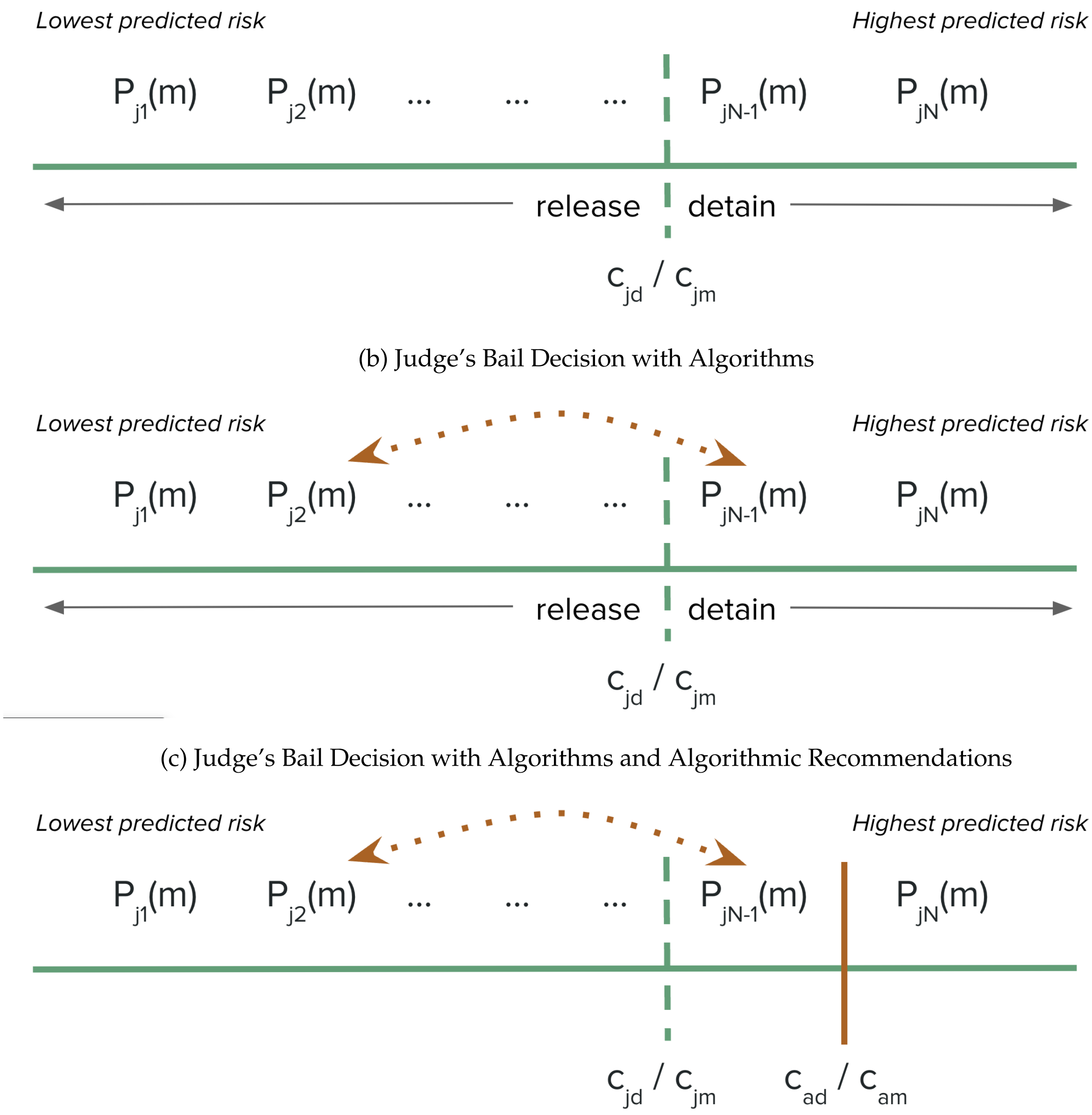
Adding algorithmic recommendations: Now, I introduce algorithmic recommendations as well. Like judges, an algorithm designer decides if they want defendants released or detained. However, the predicted probabilities of misconduct by an algorithm may differ from those made by judge j. Also, importantly, the designer can have different costs of detention and misconduct than judge j. I label the algorithm designer's detention and misconduct costs as c_{ad} and c_{am}, respectively. The designer wants to release defendant i if and only if $c_{ad} > P_{ai}(m)c_{am}$ (i.e., the cost of detention to the designer is higher than the cost of release to the designer in this case). Figure 3a illustrates how the bail decision would be made if it were fully delegated to the algorithm designer.

However, the decision is not made by the algorithm designer. Instead, the algorithm designer provides judges with scores $s_{ai}(m)$ and a threshold T^*, which I show in Figure 3b. Just as the scores are a mapping from $P_{ai}(m)$, the threshold is a mapping from $\frac{c_{ad}}{c_{am}}$, which is the ratio of the cost of detention to the cost of misconduct according to the algorithm designer. While the mapping is unknown, the judge gets information about $\frac{c_{ad}}{c_{am}}$ because the fraction of cases with $s_{ai}(m) < T^*$ is the same as the fraction of cases with $P_{ai}(m) < \frac{c_{ad}}{c_{am}}$.

The recommendation, therefore, communicates information about the preferred trade-offs of the algorithm designer. If $\frac{c_{jd}}{c_{jm}} < \frac{c_{ad}}{c_{am}}$, the algorithmic recommendation suggests a more lenient threshold than the judge's natural threshold. If not, the algorithmic recommendation suggests a more stringent threshold than that of the judge. Figure 2c demonstrates how algorithmic recommendations introduce new information that could update judge thresholds for bail decisions.

[3] Using Kleinberg et al. (2017)'s algorithm to decrease misconduct rates by 24% with no effect on jailing rates would be theoretically achieved in a similar way. Relative risk rankings are changed, and so the composition of who is released changes, but the overall number of people released does not change.

Figure 2: Bail Decision-Making



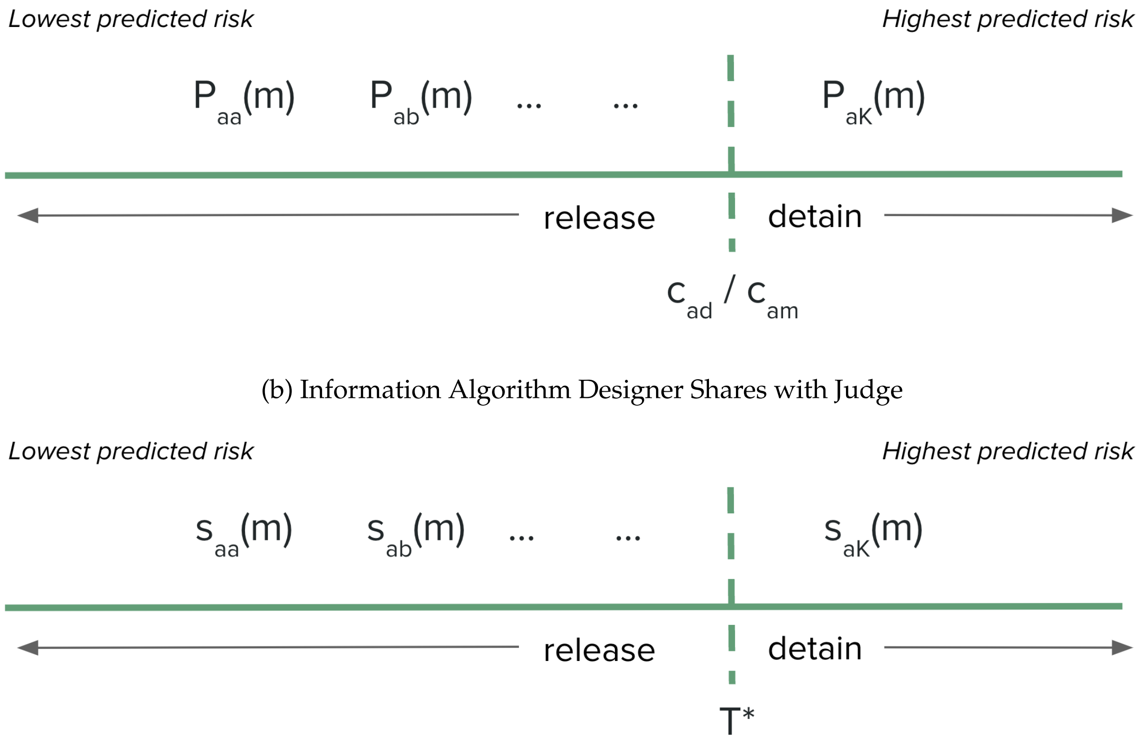
(a) Judge's Bail Decision

(b) Judge's Bail Decision with Algorithms

(c) Judge's Bail Decision with Algorithms and Algorithmic Recommendations

Notes: Figure 2a illustrates judge j's bail decision decision. The threshold $\frac{c_{jd}}{c_{jm}}$ is illustrated with a dotted line. Cases with perceived risk $P_{ji}(m)$ below that threshold will be released, while cases with perceived risk higher than it will be detained. When algorithms are introduced in Figure 2b, this can change the ordering of cases based on perceived risks (illustrated by the orange dotted arrow); however, the critical threshold and the number of cases on either side stays the same. When algorithmic recommendations are introduced in Figure 2c, new information is introduced about a recommended threshold for bail decisions, illustrated by the orange vertical line.

Figure 3: How an Algorithm Designer Communicates Recommendations


(a) Algorithm Designer's Preferred Bail Decision

(b) Information Algorithm Designer Shares with Judge

Notes: Figure 3a illustrates how the algorithm designer prefers to make bail decisions. The threshold $\frac{c_{ad}}{c_{am}}$ is illustrated with a dotted line. Cases with algorithm-predicted risk $P_{ai}(m)$ below that threshold will be released, while cases with perceived risk higher than it will be detained. Figure 3b depicts the information that is communicated to judges. The scores $s_{ai}(m)$ are a mapping from the algorithm-predicted risk probabilities $P_{ai}(m)$, and the threshold T^* is an equivalently transformed version of the threshold $\frac{c_{ad}}{c_{am}}$.

2.4 Why Might Recommendations Matter?

The algorithmic recommendations introduced into judges' bail decision-making might matter for a few different reasons. First, if there are administrative costs to deviate, then judges might adhere to the recommendations to avoid high administrative costs. Second, recommendations might nudge judges to further weight the algorithm's predictions in forming their predicted probabilities of misconduct. Third, recommendations might change misconduct costs in the event of a bad outcome. Specifically, in the bail setting case, a bad outcome is when someone is released but then commits misconduct (a type II error.)

Recommendations can change misconduct costs in two ways. First, if the recommendation is detention (harsh bail) and the judge releases the defendant (judge is lenient), then the judge is sticking their neck out more than they would have in the absence of a recommendation. If a defendant commits misconduct, the judge could face higher costs in the form of increased scrutiny and political backlash (loss of a future election). This theory aligns with recent events related to the Waukesha Christmas parade attack. The Milwaukee DA faced calls for removal after setting low bail for a person who later committed a violent crime, killing six people. Part of the political backlash was because the bail decision was "not consistent with… the risk assessment of the defendant prior to the setting of bail" (Fung 2021). Similar anecdotal evidence exists in other contexts. For instance, medical professionals have expressed hesitation to deviate from algorithmic recommendations because of concerns around increased liability. As one school therapist put it, "You have this thing telling you someone is high risk, and you're just going to let them go?" (Khullar 2023).

The second way recommendations can change costs of errors is that lenient recommendations may make lenient decisions less risky. The algorithm designer who sets the recommendation provides reputational cover to the judges. If someone commits misconduct, judges can point out that the lenient decision was in accordance with recommendations out of their control. Judges have made statements in court to this effect. For instance, in New York City, where there have been recent attempts at bail reform, judges "routinely stated that they only ordered people to be released … because the law forced them to" (Covert 2022). Making such statements is a way to signal that lawmakers, not the judges, should be responsible for subsequent pretrial misconduct outcomes.

In Section 5, I revisit this discussion of mechanisms to test which mechanism is responsible for the effect of recommendations on decisions.

3 Empirical Setting: Bail Decisions in Kentucky

I study bail decisions in Kentucky because this setting provides a unique opportunity to estimate the independent effects of algorithmic recommendations. Between March 2011 and May 2013, there was an algorithm used to predict pretrial misconduct risk. Judges were given these predictions to inform their bail decisions. However, from June 2011 onward, algorithmic recommendation were given to judges as well, but only for some cases. I leverage the variation over time and cases to estimate the effects of the recommendations on decisions.

Before the introduction of algorithmic recommendations, bail decisions were made as follows. After a defendant was booked into jail, a pretrial services officer (an administrative court employee) interviewed the defendant to collect information and calculate a risk score. Within 24 hours of booking, the officer presented information about the defendant, their risk score, and the alleged incident to a judge. After receiving this information, the judge made a bail decision in a few minutes.

Judges' bail decisions determine conditions for people's pretrial release from jail. These conditions are frequently financial in nature and require defendants to post some amount of money to be released from jail. Judges can choose not to require money for release, which would be a more lenient decision. Throughout this paper, I discuss judges setting "lenient bail," which means not requiring money for release, or "harsh bail," which means requiring money for release or requiring detention outright.[4]

The Kentucky setting has a few features worth highlighting that contrast with other US bail settings. First, the bail decisions are usually made in phone conversations between pretrial officers and judges, rather than during in-person bail hearings, which are common in much of the US.[5] The fact that the decision is made over the phone means that the defendant is not present at the bail setting. Second, police have full authority to charge in Kentucky, which means there is no prosecutorial review before the judge makes a bail decision. Thus, judges' bail decisions do not follow any prosecutor actions.[6]

[4] The bail decision differs from my theoretical framework in Section 2, because in practice, judges do not choose to release or detain people. Rather, they choose whether to require money for release. Making the decision a binary choice to set money bail or not cleanly maps onto the 2011 Kentucky policy recommendation.

[5] Kentucky has been using calls for pretrial services since 1976. They are especially efficient in areas of the state where people are very spread out where in-person bail hearings would therefore mean significant time costs.

[6] See Appendix A.1 for more background on Kentucky bail setting.

3.1 The Kentucky Pretrial Risk Assessment

The algorithm used to predict misconduct during the study period was the Kentucky Pretrial Risk Assessment (KPRA). The KPRA was created in-house, fitting a regression model to predict pretrial misconduct using the existing Kentucky administrative data. The KPRA was not a complex black-box machine learning tool. Rather, it was a checklist tool that added up points based on "yes" or "no" answers to a series of questions. Points were added up and then converted to score levels of "low," "moderate," or "high." Totals of 0-5, 6-13, and 14-24 corresponded to low, moderate, and high levels, respectively. Judges were informed of risk levels rather than the underlying number of points.

The factors in the KPRA are mostly criminal history elements (e.g., prior failure to appear, pending case), but there is also information about the current charge (e.g., whether the charge is a felony of class A, B, or C) and the defendant's personal history (e.g., verified local address, means of support). See Appendix A.2 for more details about how the score is calculated.

3.2 House Bill 463 Introduced Algorithmic Recommendations

In response to large increases in the incarcerated population between 2000 and 2010, Kentucky House Bill 463 (HB463) went into effect on June 8, 2011. The law recommended low or moderate risk defendants be released without the requirement to post money – these cases were recommended "lenient bail." The policy change did not change how the risk scores and levels were calculated; it introduced recommendations for how to use them.

If judges wanted to override the recommendation, they could do so easily by providing a reason. In practice, this was as simple as saying a few words (e.g., "flight risk") to the pretrial officer on the phone. The policy change did not set a recommendation for high risk defendants. Therefore, the policy introduced a recommendation (lenient bail) for some defendants (people with low or moderate risk scores) but not others (people with high risk scores).

3.3 A Challenge to Studying HB463

One challenge to studying HB463 is that it was a large bill of about 150 pages and 110 sections.[7] The bill introduced more policy changes beyond introducing algorithmic recommendations to bail decision making. Therefore, a key empirical concern is incorrectly attributing estimated effects to the recommendations when they are instead due to concurrent policy changes.

Qualitative review of the bill, paired with interviews with practitioners, pinpointed a change to policing in the bill that is a potential empirical concern. According to a memo from the Louisville chief of police, the bill amended existing law "by requiring law enforcement officers to issue citations instead of making physical arrests" for many misdemeanor offenses.[8] In other words, some misdemeanor offenses may have no longer resulted in arrest after HB463.

To address this possible change in the composition of cases, I omit cases that were supposed to result in citation after HB463, according to the bill's language. Therefore, my sample of cases excludes this group that could have been simultaneously impacted by policing policy change. I use "Standard Operating Procedures" documentation (SOP 10.1) from the Louisville Metro Police Department to identity the relevant cases.

3.4 Kentucky Administrative Court Data

To study algorithmic recommendations, I use administrative court data from Kentucky's Administrative Office of the Courts, which covers all criminal cases with felony- or misdemeanor-level charges. I focus on pretrial interviews and initial bail decisions that cover single cases. I limit the sample to initial bail decisions made by district judges between March 18, 2011 and June 30, 2013 because that is the time period during which Pretrial Services used a consistent algorithm to predict pretrial misconduct (the KPRA) but algorithmic recommendations changed. As described in Section 3.3, I also subset to cases with offenses that are arrestable both before and after HB463.

The final dataset consists of around 154,000 initial bail decisions across approximately 122,000 unique defendants and 433 unique judges. Each row is an initial bail decision

[7] The full bill can be downloaded from this webpage: https://apps.legislature.ky.gov/record/11rs/hb463.html.

[8] However, there are exceptions to this requirement "which still allow officer discretion to make a physical arrest for certain offenses." The referenced memo is from Robert C. White on June 2, 2011, "Re: SOP 10.1, Enforcement - Revised General Order #11-013."

with information about defendants (demographics, risk score characteristics, risk scores), associated bail judges, and information on the associated set of charges.[9]

3.5 New Algorithmic Recommendations Were More Lenient Than The Status Quo

Figure 4 demonstrates the distribution of risk scores across all cases in the administrative data. Green bars on the left are cases with low risk scores, gray bars in the middle are cases with moderate risk scores, and orange bars on the right are cases with high risk scores. The distribution skews low risk. In fact, 90% of cases are in the low or moderate categories. Therefore, after recommendations are introduced, 90% of cases receive the lenient bail recommendation. However, before recommendations were introduced, only 32% of cases received lenient bail. So, it is clear that the new recommendations set a much lower threshold for lenient bail than existed beforehand. If the state wanted to set a threshold to align with the pre-existing level of bail setting, the lenient recommendation would have kicked in for cases with scores below 4 rather than below 14.

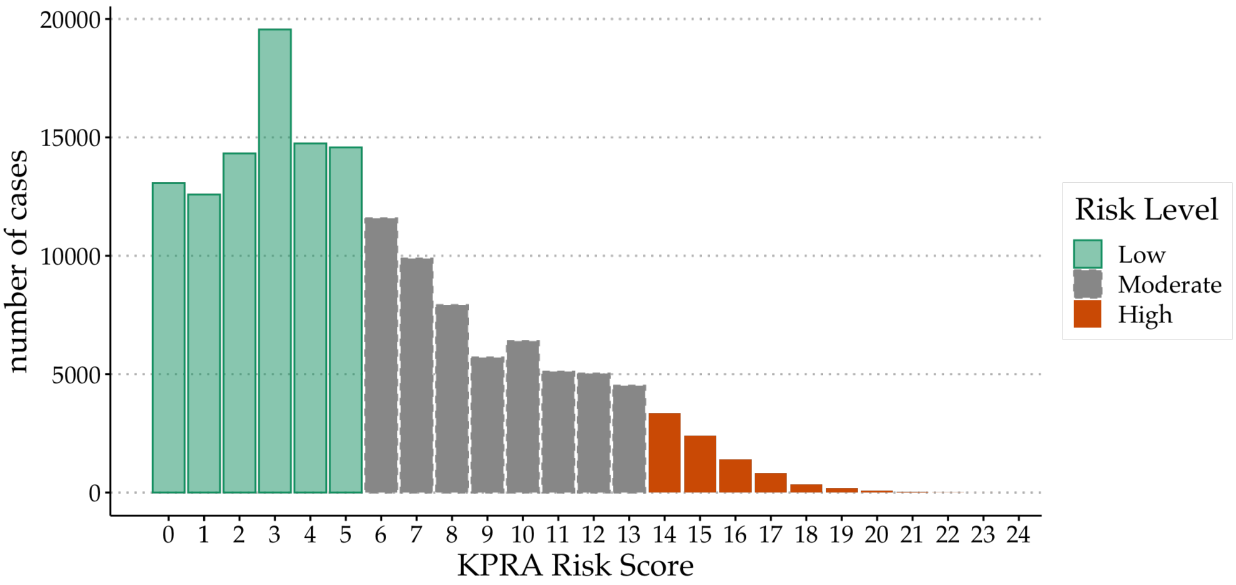
The chosen recommendation threshold was a normative decision on the part of the state rather than a natural consequence of any underlying risk scoring system. As explained in Section 2.4, many different decision thresholds are consistent with the same underlying risk rankings. The threshold of 14 suggested the state wanted judges to set lenient bail more frequently than they did under the status quo. However, if the state had chosen a threshold of 2, that would have suggested the state wanted judges to set lenient bail less frequently. Both thresholds are relative to the same underlying algorithm for predictions, but they have very different policy implications. While many researchers focus on how algorithms can change decisions in a "locally optimal" (or an "allocative") sense (Kleinberg et al. 2017), the threshold choice previews how algorithmic recommendations may have their own independent causal effects.

4 Algorithmic Recommendations Change Judge Decisions

Throughout my study period, risk scores were available to judges when they were setting bail. In June 2011, the risk scores did not change, but judges were given explicit recom-

[9] The raw administrative data do not include the KPRA scores themselves, but I use data on the underlying risk score components, along with the known weights from Table A.1, to construct them.

Figure 4: The Risk Score Distribution



Notes: This figure demonstrates the number of cases across the full risk score distribution. The color of the bars changes from green (with a solid outline) to gray (with a dashed outline) between 5 and 6 to demonstrate the threshold between low and moderate risk cases. The color of the bars again changes from gray (with a dashed outline) to orange (with no outline) between 13 and 14 to demonstrate the threshold between moderate and high risk cases.

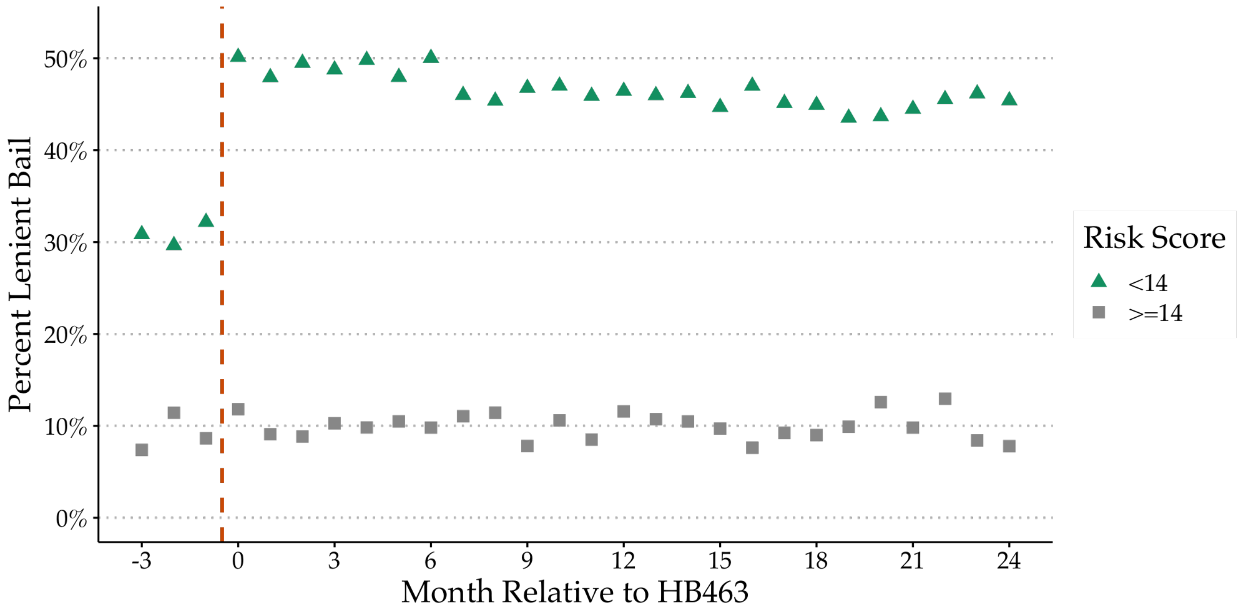
mendations on how to set bail based on the scores. The new recommendation was to set lenient bail (no money bail) for cases with low or moderate risk scores.

Does the new algorithmic recommendation change human decisions, even though the algorithm's predictions stay the same? I answer this question by using a differences-in-differences specification, in which high risk cases are the control group and low/moderate risk cases are the treatment group. Intuitively, this is because high risk cases do not experience a change in recommendations, but low/moderate risk cases do.

Figure 5 illustrates the rate of lenient bail for low/moderate and high risk cases over time. There are only a few pre-policy time periods because the method of calculating the risk score changed in March 2011. There is an obvious jump up in lenient bail for low/moderate cases once the recommendations go into effect. The underlying assumption of using a differences-in-differences approach is parallel pre-trends, which is supported by the visual evidence.

To formally estimate causal effects and test for pre-trends, I estimate a standard specification of the form

Figure 5: Lenient Bail Rates by Risk Level over Time



Notes: This figure shows the rate of lenient bail over months by risk score groups. Months are indexed relative to the introduction of algorithmic recommendations. Cases with risk scores below 14 are shown as green triangles, while cases with risk scores at or above 14 are shown as gray squares. The orange dotted line shows when HB463 went into effect.

$$lenient_{itj} = \sum_{m \neq -1} [\beta_m \times I(score_i < 14)] + X_{itj} + \epsilon_{itj}, \tag{1}$$

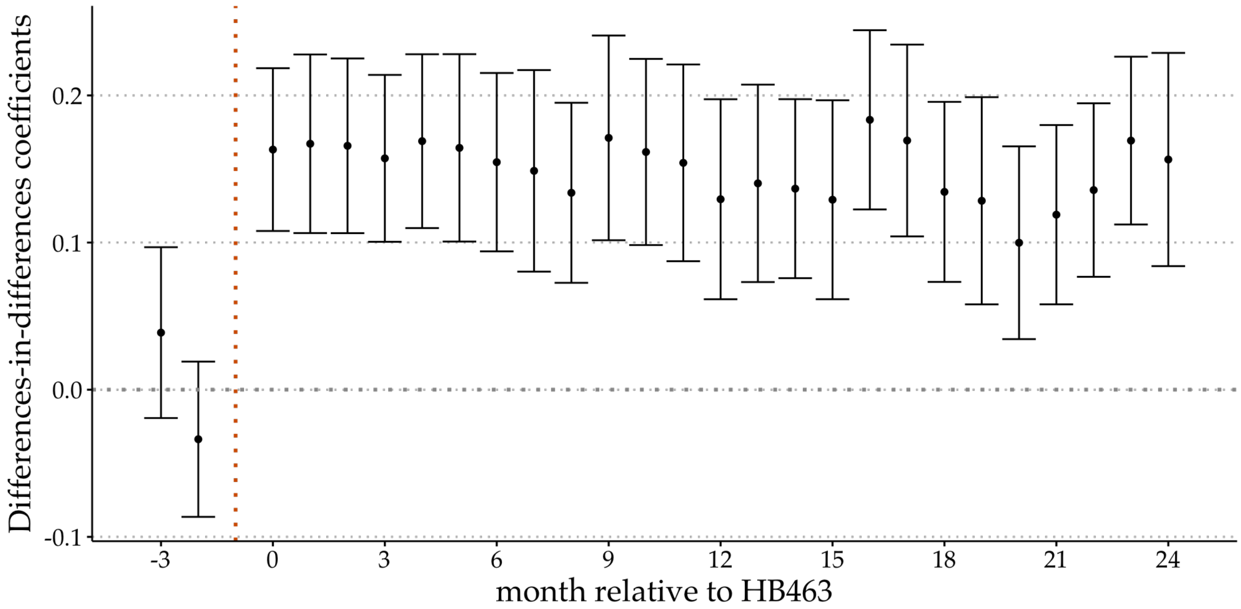
where $lenient_{itj}$ is an indicator for if the bail for case i at time t decided by judge j is lenient (no money bail), and $I(score_i < 14)$ is an indicator for if the risk score for case i is below 14. Distinct coefficients are estimated for each month m relative to HB463 adoption, and m=-1 is the omitted group. I include the same vector of controls X_{itj}, which includes controls for day of week, month-year, exact risk score, top charge level and class, defendant demographics (race, gender), judge, county, and other risk score components, listed in Table A.1.[10] I cluster standard errors by judge.

Figure 6 shows the dynamics by plotting the values of β_m. The coefficients before recommendation introduction are close to 0 and do not demonstrate any evidence of pre-trends. Thus, the differences-in-differences approach is convincing in this setting. Figure A.1 in the Appendix demonstrates that results look similar across different sets of controls. To estimate a useful summary coefficient, I estimate pooled differences-in-differences coefficients. Pooling time periods shows that the algorithmic recommendation increased

[10] I include all risk score components except for verified address and support.

lenient bail by 15 percentage points following the policy change off of a baseline of 31%. Therefore, the recommendations increased lenient bail by about 50%.

Figure 6: Dynamic Differences-in-Differences Estimates



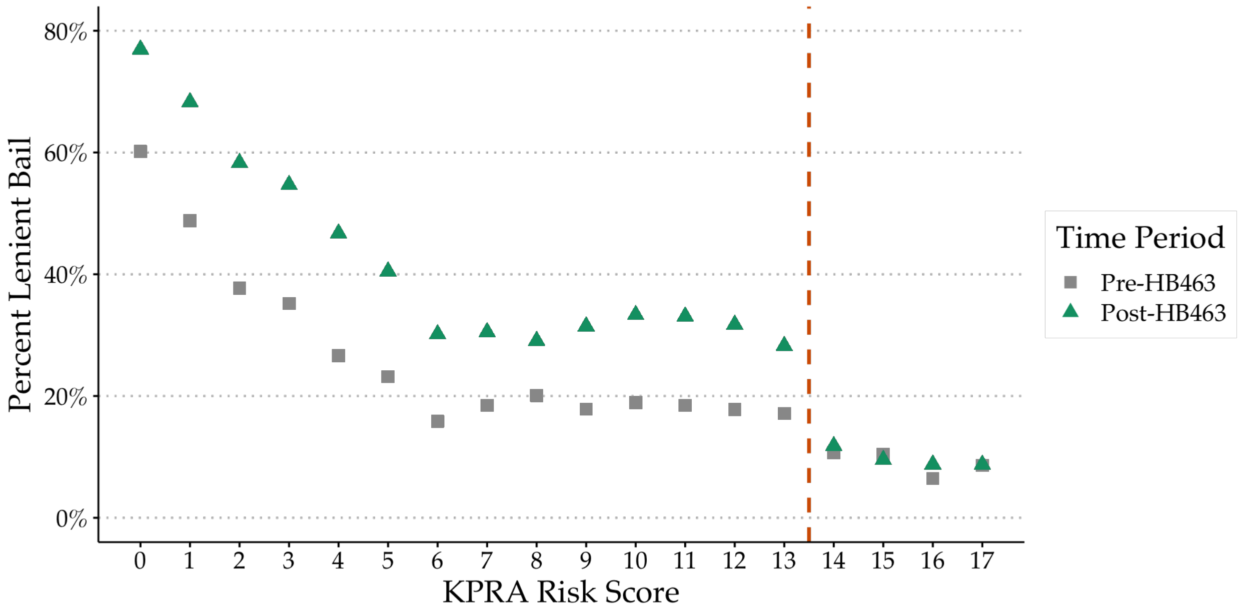
Notes: This figure shows the difference-in-differences coefficients for months relative to recommendation introduction. The orange dashed line denotes the omitted period of the month before recommendation introduction.

I can also explore effects for marginal cases – that is, cases near the recommendation threshold. Recall that after June 2011, the lenient bail recommendation kicks in for cases below a risk score of 14. Therefore, cases that are very similar in terms of risk score receive different recommendations based on which side of the threshold they fall on. Visually, I can demonstrate how lenient bail rates change across the distribution after recommendations are introduced. Figure 7 shows the percentage of cases that receive lenient bail before and after HB463 for each risk score from 0-17.[11] The pre-policy and post-policy points are similar for those cases that do not receive a recommendation, but the post-policy rates are 10-20 percentage points higher for the low and moderate risk level cases.[12]

[11] I plot rates for the scores 0-17 instead of the full distribution of 0-24 to focus on risk scores with a sufficient number of observations before and after HB463. Figure 4 shows that there are few observations at the high end of the risk distribution: the number of observations is very small for scores above 17, especially in the pre-HB463 period, because there are only 2 months of pre-period data. For instance, there are only 22 cases pre-HB463 with a score of 18.

[12] Figure 7 also demonstrates that moderate risk scores receive similar lenience across the range of scores, but there is a clear downward trend in lenient bail for the low risk scores as the scores get higher. This is likely because it is obvious when someone is very low risk (no or almost no criminal history) as opposed to when they are on the boundary of low-moderate risk. This trend suggests that judges may not differentiate much between different levels of risk once a certain level has been passed.

Figure 7: Percent Lenient Bail across Risk Scores and Time Periods



Notes: This figure demonstrates the percentage of cases that receive lenient bail across the risk score distribution, both before and after HB463. The orange dashed line marks the threshold between moderate and high risk. Before HB463, there were no bail recommendations. After HB463, cases with scores left of the orange line received a lenient bail recommendation, but those with scores right of the orange line did not. The gray rectangles show the rates before HB463, while the green triangles show the rates after HB463.

To formally compare marginal effects to full effects, I estimate pooled differences-in-differences coefficients for different groups of cases based on risk scores. I estimate coefficients for a variety of "bandwidths" around the relevant 13/14 threshold – 10-17, 11-16, 12-15, and 13-14 – and I estimate specifications analogous to Specification 1 but pool time periods into "before" and "after."[13] I also display the pooled coefficient for the full range of risk scores. Table 1 presents the results.

Table 1 shows that the recommendation's causal effect is qualitatively similar across the bandwidths. The effect for the full population, as mentioned before, is 15.3 percentage points off of a baseline of 31.0% (a 49.4% increase). For scores of 10-17, the effect is 13.3 percentage points off of a baseline of 18.9% (a 70% increase). For scores of 11-16, the effect is 12 percentage points off of a baseline of 18.5% (a 65% increase). For scores of 12-15, the effect is 11.5 percentage points off of a baseline of 17.8% (a 65% increase). For scores of 13-14, the effect is slightly smaller at 9.2 percentage points off of a baseline of 17.1% (a

[13] When I subset to smaller risk score bandwidths, I am amending my differences-in-differences approach to make the control and treatment groups more similar in terms of risk scores. Since the risk scores are discrete, this is conceptually distinct from regression discontinuity or differences-in-discontinuity approaches, which require continuous running variables.

Table 1: Differences-in-Discontinuities Results across Bandwidths

	Dependent variable: I(lenient bail)				
Bandwidth	All	10-17	11-16	12-15	13-14
I(score<14) x Post	0.153^{***} (0.020)	0.133^{***} (0.021)	0.120^{***} (0.021)	0.115^{***} (0.023)	0.092^{***} (0.032)
Pre-Mean Score<14	0.310	0.182	0.179	0.175	0.171
Time/Score FEs	Y	Y	Y	Y	Y
Charge/judge/county/demo controls	Y	Y	Y	Y	Y
Risk component controls	Y	Y	Y	Y	Y
Observations	153,597	29,005	21,785	15,273	7,854
R^2	0.268	0.183	0.181	0.173	0.173
Adjusted R^2	0.265	0.168	0.161	0.146	0.121

Notes: This table displays estimated differences-in-differences coefficients in specifications in which the dependent variable is lenient bail. The table shows results across different bandwidths – all risk scores, as well as risk score ranges of 10-17, 11-16, 12-15, and 13-14 – using the same controls: fixed effects for month-year, day of week, exact risk score, top charge level and class, judge, county, defendant gender, and defendant race. Specifications also control for all the characteristics that factor into risk score, listed in Table A.1. Standard errors are always clustered at the judge-level. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

54% increase). Therefore, the effect of the recommendation is a 55%-70% increase in the lenient bail rate for marginal cases. This effect is quantitatively meaningful and is evident in the raw data as well (Figure 7). Table A.3 in the Appendix also shows that the results are nearly identical regardless of the exact set of controls included in the specifications.

5 Recommendations Can Change Costs of Decision Errors

In Section 2, I outlined three possible mechanisms through which algorithmic recommendations may impact human decisions. First, judges may adhere to recommendations to avoid any administrative costs of deviations. However, in this institutional context, there were minimal administrative costs. To override the recommendation, judges could just say a few words (e.g., "flight risk") to pretrial officers during the bail decision. Since this channel is not relevant in this context, I focus on distinguishing between the remaining two mechanisms.

One possibility is that recommendations nudge judges to further weight the algorithm's predictions when forming their own predicted probabilities of misconduct. This would reshuffle how judges order cases in terms of risk but would not effect the threshold at which they set their lenient bail threshold.

The other possibility is that recommendations change the costs of errors to judges. The lenient recommendation could make lenient decisions less risky for judges in low and moderate risk cases because it provides reputational cover in the event of misconduct. The legislature sets the recommendation and therefore is responsible for some of the misconduct costs when recommendations are followed.

If the algorithmic recommendations change how judges weight risk scores, then bail decisions should change for cases with all types of risk scores. But, if they change costs of errors, decisions should change only for low and moderate risk cases (because high risk cases are not impacted). Therefore, the two possible mechanisms generate distinct testable predictions. The changing predictions mechanism should lower lenient bail rates for high risk cases, while the error cost mechanism should not change lenient bail rates for high risk cases.

To test these diverging predictions, I estimate the effect of the recommendations on high risk cases. I use a regression discontinuity in time approach, leveraging the sharp change in recommendations on June 8, 2011. The key identifying assumption is that other covariates vary smoothly at the policy change threshold (Hausman and Rapson 2018).

As mentioned in Section 3.4, the House bill that implemented the algorithmic recommendations also made some offenses ineligible for physical arrests.[14] To deal with this change, I limit the sample to only cases that are arrestable both before and after the policy change. Without this step, the regression discontinuity in time approach would be invalid because there is a discontinuous increase in the share of felony cases at the time of the June 2011 change (in accordance with some misdemeanors yielding citations instead of arrests). If I include only arrestable cases, defendant and cases covariates do not experience discontinuous changes at the policy threshold (see Figure A.2).

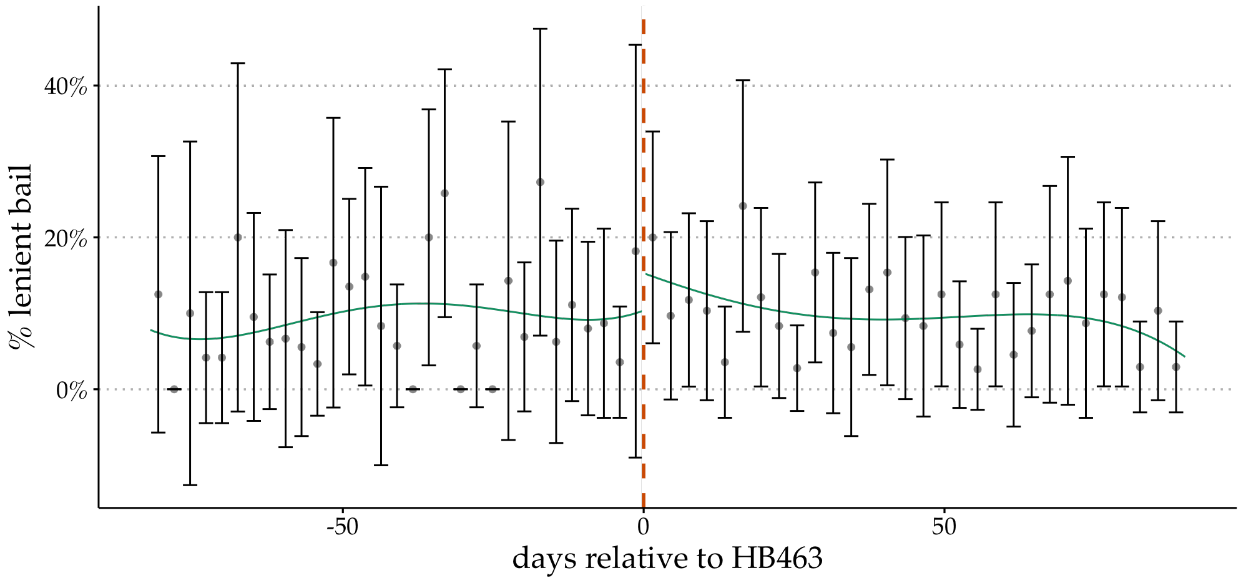
I also check if there is any discontinuity around HB463 that indicates manipulation of the running variable (days). This tests whether criminal justice actors strategically bunched bail decisions before or after the policy change. I test this with the McCrary test (McCrary 2008). Specifically, I use manipulation testing procedures using local polynomial density estimators from Cattaneo, Jansson, and Ma (2020). The p-value from this test is 0.43, which indicates no significant difference between the trends in number of cases before and after

[14] According to a memo from the Louisville chief of police, the bill amended "KRS 431.015 by requiring law enforcement officers to issue citations instead of making physical arrests" for many misdemeanor offenses; however, there are "exceptions to this requirement within KRS 431.015 and several other statutes, which still allow officer discretion to make a physical arrest for certain offenses."

the policy.[15]

Now that I have established that there is no evidence of manipulation across the threshold and that covariates move smoothly across the threshold, I estimate the regression discontinuity of interest. I use nonparametric methods from Calonico, Cattaneo, and Titiunik (2015) and the mean square error optimal bandwidth from Calonico, Cattaneo, and Titiunik (2015). I plot the resulting model in Figure 8. The p-value for the regression discontinuity coefficient is around 0.3, indicating it is not statistically significant at conventional levels. Regardless of the specific choice of parameters, the coefficients remain statistically insignificant.[16] Mapping this back to the testable predictions shows that this evidence is consistent with the recommendations changing the costs of errors to judges.

Figure 8: Regression Discontinuity in Time Plot



Notes: This figure graphically plots the regression discontinuity in time. The dots are the average rates of lenient bail for each bin of observations. The optimal bin size used is 25 observations. The error bands show 95% confidence intervals for each bin. The dotted orange line marks the first day after HB463. The green lines display the flexible model fitted both before and after HB463.

[15] I can also test this graphically by constructing density plots based on methods from Cattaneo, Jansson, and Ma (2020), Cattaneo, Jansson, and Ma (2022), and Cattaneo, Jansson, and Ma (Forthcoming). Figure A.3 in the Appendix groups the data into bins and plots the relative frequency of observations in each bin as well as the averages and confidence intervals of those bins. The figure shows that the confidence intervals meaningfully overlap, which again indicates there is no significant difference around the threshold.

[16] I test the sensitivity of these results across a range of bandwidths, kernels, and confidence interval options. I present all estimated coefficients and their 95% confidence intervals in Figure A.4.

6 Heterogeneous Effects of Algorithmic Recomendations and Racial Disparities

In this paper, I have shown that algorithmic recommendations meaningfully impact human decision-making: lenient recommendations increase lenient bail by 50%-70%. I also showed results consistent with a theory that recommendations change the costs of errors to human decision-makers. In this section, I investigate how the impacts of algorithmic recommendations feed into the impacts of algorithm-based decision-making on racial disparities.

In Kentucky, the algorithm for predicting misconduct generates higher risk scores for Black defendants than white defendants. Figure 9 shows the risk score distributions for both racial groups.[17] After HB463 all low and moderate risk cases were recommended lenient bail. If judges had followed the recommendations 100% of the time (no deviations), Black defendants would have been 3.3 percentage points less likely than white defendants to receive lenient bail (91.5% vs. 94.8%).[18] However, after HB463, the true observed racial gap in lenient bail is 9.3 percentage points (36.7% vs. 46%).[19] This means that the observed racial gap in lenient bail is almost three times as large as it would have been had the algorithmic recommendation mechanically set bail. Therefore, the differences in lenient recommendation rates across race explain a minority of the observed racial gap.

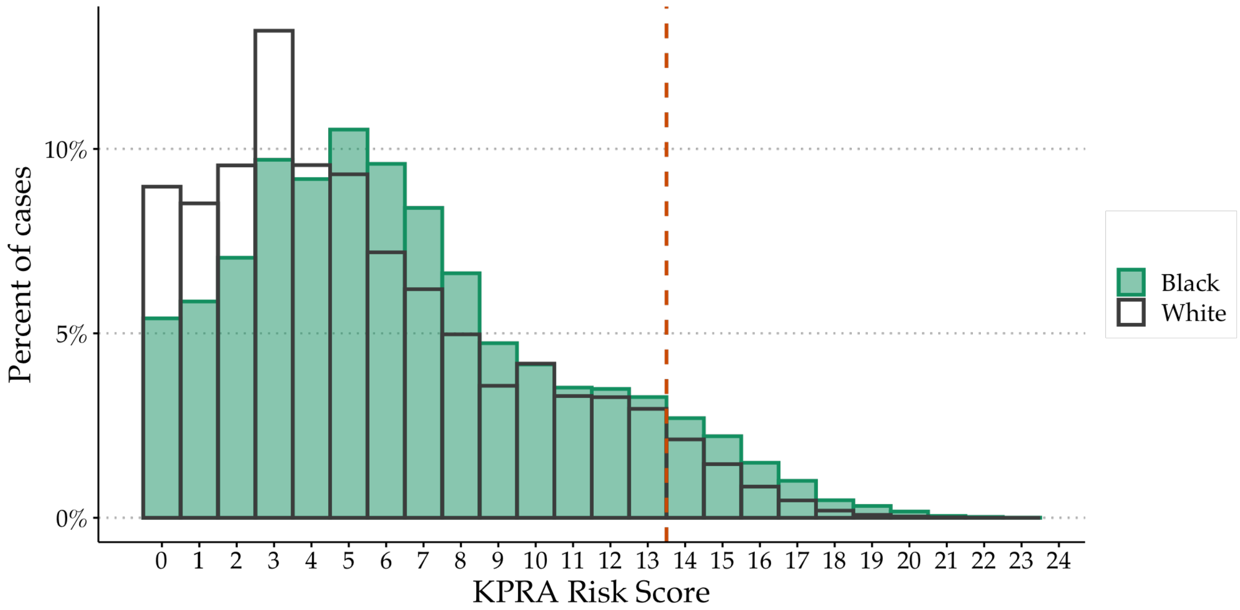
Is the difference between the simulated racial gap and the observed racial gap explained by differences in underlying score distributions? In other words, do Black defendants receive lenient bail less often because their risk scores are higher within the low/moderate score range? I use Figure 10 to answer this question visually. Figure 10 uses data after recommendations are introduced to demonstrate the average rate of lenient bail for Black and white defendants for each specific risk score. For cases in which there is no recommendation in effect (cases with risk scores over 13), lenient bail rates are similar for Black and white defendants with identical risk scores. However, for cases in which the lenient bail recommendation is in effect (cases with risk scores under 14), Black defendants are less likely than white defendants with identical risk scores to receive lenient bail. In

[17] The largest contributing factors to those differences are higher failure to appear rates, prior felony rates, and prior violent conviction rates for Black defendants (all of which negatively impact risk score). See Table A.4 for a table comparing all risk score components by racial group.

[18] Note that owing to the different distribution of risk scores by race, any cut-off based on those scores will yield a racial disparity based on perfect compliance, but the implied gap varies according to the chosen cut-off.

[19] Figure A.5 demonstrates how lenient bail rates between white and Black defendants changed over time based on risk score group.

Figure 9: Risk Score Distributions by Defendant Race



Notes: This figure illustrates the distribution of cases across the risk score distribution for Black and white defendants after HB463. Solid green bars represent cases with Black defendants, while unfilled gray bars represent cases with white defendants. The orange dashed line shows the threshold between moderate and high risk scores.

other words, judges are more likely to deviate from the lenient recommendation for Black defendants than for white defendants with identical risk scores.

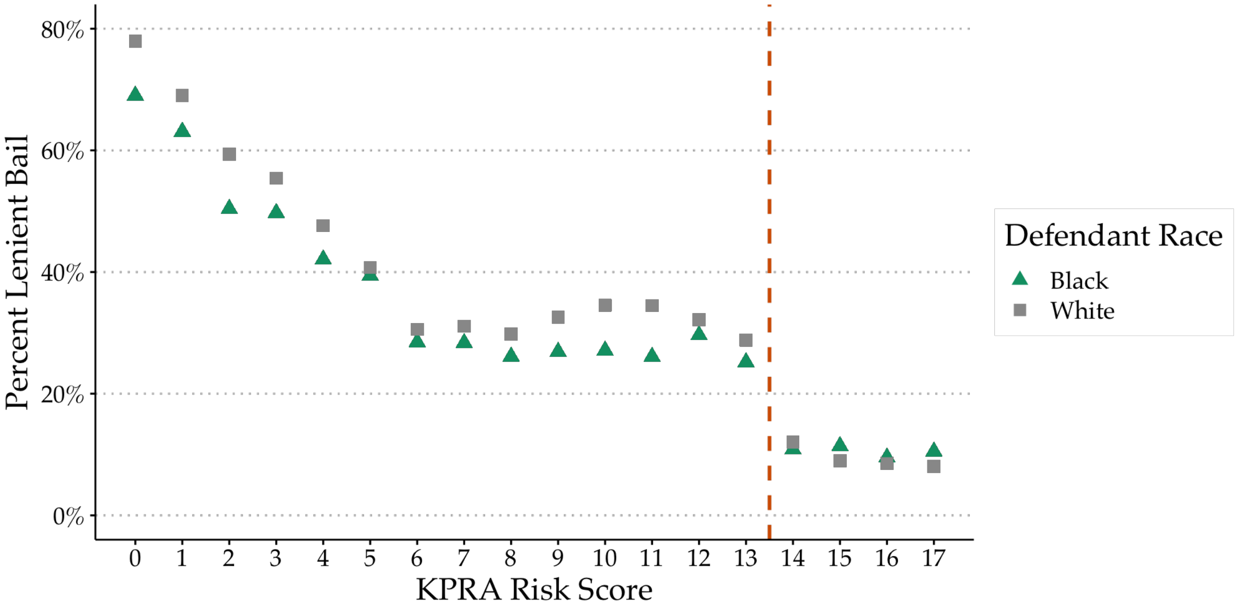
The visual evidence in Figure 10 suggests the algorithmic recommendations have larger effects for white defendants than for Black defendants. I can further validate this result by estimating differences-in-differences coefficients, following Specification 1, separately for each racial group. Figure 11 illustrates the dynamics separately for each group. The estimates for white defendants (in gray squares) are consistently higher than the estimates for Black defendants (in green triangles).[20]

To obtain a useful summary coefficient, I estimate pooled differences-in-differences coefficients for both racial groups. I can also estimate a triple differences specification of the form

$$lenient_{itj} = \beta_1[I(score_i < 14) \times Post_t] + \beta_2[I(score_i < 14) \times Black_i] + \\ \beta_3[Post_t \times Black_i] + \beta_4[I(score_i < 14) \times Post_t \times Black_i] + X_{itj} + \epsilon_{itj}, \quad (2)$$

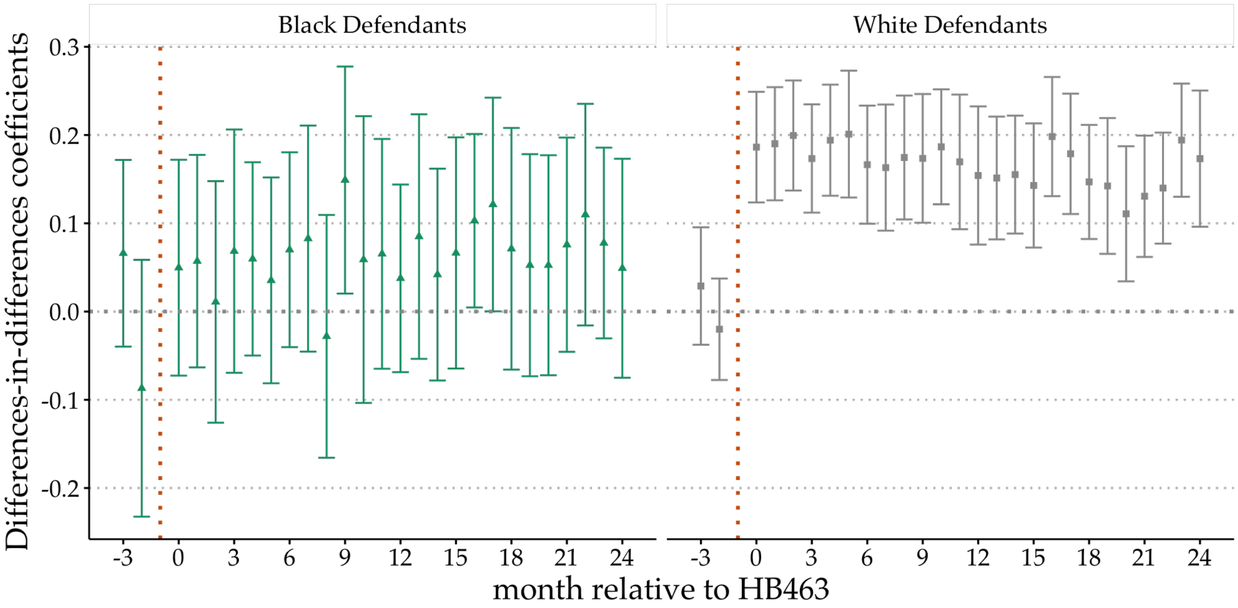
[20] Figure A.6 in the Appendix demonstrates that results are similar across different sets of controls.

Figure 10: Lenient Bail by Defendant Race and Risk Score



Notes: This figure shows the rates of lenient bail over the risk score distribution for Black and white defendants after algorithmic recommendations are in place. The gray squares are rates for cases with white defendants, while the green triangles are rates for cases with Black defendants. The orange dashed line marks the thresholds between moderate and high risk scores.

Figure 11: Dynamic Differences-in-Differences Estimates by Racial Group



Notes: This figure shows the difference-in-differences coefficients for months relative to recommendation introduction. The orange dashed line denotes the omitted period of the month before recommendation introduction. Results from the specification for only Black defendants are displayed on the left, with point estimates shown as green triangles. Results from the specification for only white defendants are displayed on the right, with point estimates shown as gray squares.

where $lenient_{itj}$ is an indicator for if the bail for case i at time t decided by judge j is lenient (no money bail), $I(score_i < 14)$ is an indicator for if the risk score for case i is below 14, and $Black_i$ is an indicator for if the defendant in case i is Black. The coefficient of interest is β_4, since this explores the heterogeneity of the effect of recommendations by defendant race. I include the same vector of controls X_{itj} that was in Specification 1, and I cluster standard errors by judge.

I pool time periods in columns 1 and 2 in Table 2, which show that the algorithmic recommendation increased lenient bail by 16.8 percentage points for white defendants and 8.8 percentage points for Black defendants. Therefore, the recommendations had about half of the effect on judge leniency for Black defendants as they did for white defendants. Column 3 of Table 2 shows the triple differences coefficient when the specification is estimated for the full sample of defendants. The results are essentially identical to those from columns (1) and (2): the difference between the effects is estimated to be around 8.3 percentage points, and that difference is statistically significant at the 5% level.

Judges work in counties, meaning that different judges make decisions for very different defendant populations. In the vast majority of Kentucky counties, less than 10% of defendants are Black, but in a few, over 30% of defendants are Black (see Figure A.7). The uneven spatial variation in Black defendant percentages across counties means that different responses across judges could generate racial disparities at the state level.[21] Are responses to recommendations different across defendants for a single judge, or are they different across judges with different defendant populations?[22]

To test this question, I re-estimate Specification 2 but allow for judge fixed effects that vary by time period ($I(Post_t)$) and risk score ($I(score < 14)$). Doing so generates four different fixed effects for each judge. I show the results of this exercise in column 2 of Table 3. The coefficient of interest (that on the interaction of $I(score < 14) \times Post \times Black$) shrinks by 75% and is no longer statistically significant. Judges usually stay working in the same county over time, so the results are similar if I instead allow for time and score varying fixed effects by county; I show these results in column 3. The evidence in Table 3 demonstrates that different responses across judges drive most of the differences in effects by defendant race. In other words, judges who see more Black defendants respond less to lenient recommendations than judges who see more white defendants do.

[21] Figure A.8 shows spatial variation in racial composition of defendant populations across Kentucky.

[22] Stevenson (2018) found that judges who see more white defendants become more lenient after HB463. Once she allowed for time-varying fixed effects for different circuit courts, there were no remaining changes to racial disparities.

Table 2: Differences-in-Differences Results by Race

	Dependent variable: I(lenient bail)		
	DD *(White Defendants)*	DD *(Black Defendants)*	DDD
	(1)	(2)	(3)
I(score<14) x Post	0.168^{***}	0.088^{***}	0.167^{***}
	(0.020)	(0.034)	(0.020)
I(score<14) x Black			0.032
			(0.029)
Post x Black			0.009
			(0.031)
I(score<14) x Post x Black			-0.083^{**}
			(0.033)
Avg Dep Var (Pre-HB463)	0.312	0.297	0.309
Time/Score FEs	Y	Y	Y
Charge/judge/county/demo controls	Y	Y	Y
Risk component controls	Y	Y	Y
Observations	128,863	24,349	153,212
R^2	0.273	0.245	0.269
Adjusted R^2	0.270	0.229	0.266

Notes: This table displays the estimated coefficients in a differences-in-differences model for the sample of white defendants in column (1) and those for the sample of Black defendants in column (2). Column (3) shows the results from the triple differences specification, Specification 2, for the full sample. The coefficient of interest is that on the triple interaction of "I(score<14) x Post x Black." All specifications use the same controls: fixed effects for month-year, day of week, exact risk score, top charge level and class, judge, county, defendant gender, and defendant race. Specifications also control for the same characteristics which factor into risk score from Specification 1. Standard errors are always clustered at the judge-level. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table 3: Triple Differences Results with Time and Score Varying Fixed Effects

	Dependent variable: I(lenient bail)		
	DDD	DDD	DDD
	(1)	(2)	(3)
I(score<14) x Post	0.167^{***}		
	(0.020)		
I(score<14) x Black	0.032	−0.006	−0.010
	(0.030)	(0.035)	(0.028)
Post x Black	0.009	0.006	0.008
	(0.031)	(0.029)	(0.024)
I(score<14) x Post x Black	-0.083^{**}	−0.020	−0.026
	(0.034)	(0.034)	(0.029)
Extra FEs	NA	judge x under14 x post	county x under14 x post
Observations	153,230	153,230	153,230
R^2	0.269	0.279	0.275
Adjusted R^2	0.266	0.264	0.267

Notes: All three columns show results from the triple differences specification, Specification 2. The coefficient of interest is that on the triple interaction of "I(score<14) x Post x Black." All specifications use the same base controls: fixed effects for month-year, day of week, exact risk score, top charge level and class, judge, county, defendant gender, and defendant race. Specifications also control for the same characteristics that factor into risk score from Specification 1. Column (2) allows for score-time-varying judge fixed effects, and column (3) allows for score-time-varying county fixed effects. Standard errors are always clustered at the judge-level. *p<0.1; **p<0.05; ***p<0.01.

I can further demonstrate the link between defendant population and judge responsiveness. I subset to cases with risk scores below 14 so that time period is the only component that impacts recommendation exposure. I then limit the sample to cases decided by judges with 50 or more bail decisions both before and after HB463 (114 judges). I estimate the following specification:

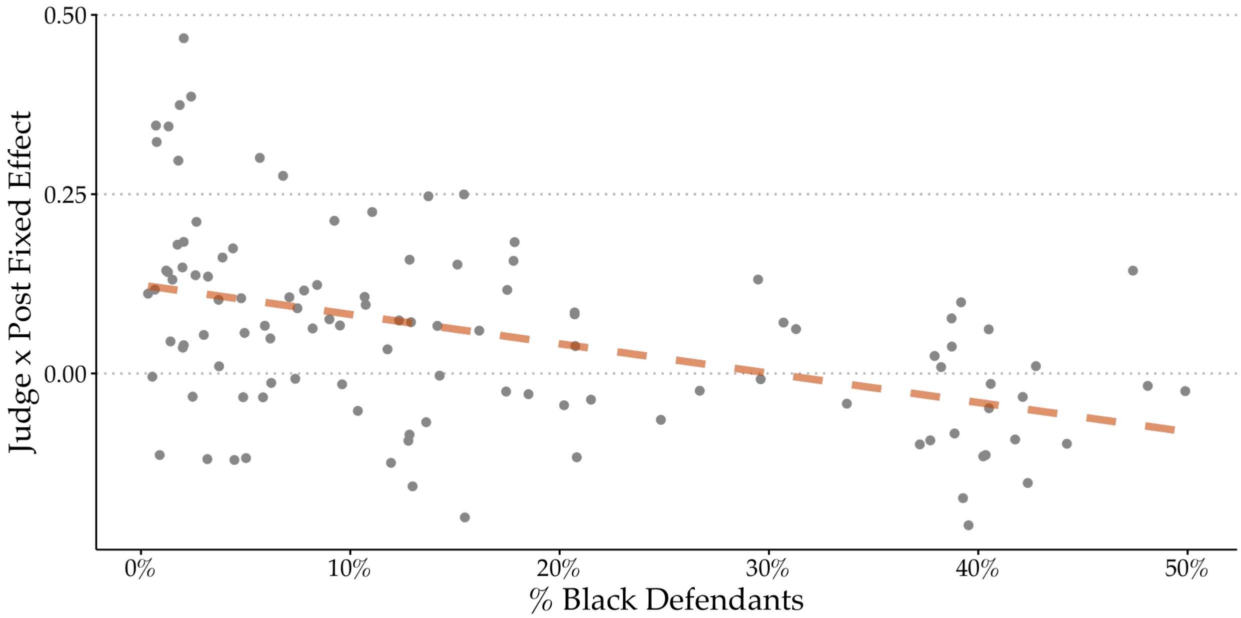
$$lenient_{itj} = \beta_3[Post_t \times Black_i] + X_{itj} + \epsilon_{itj}, \tag{3}$$

where I include judge fixed effects interacted with $Post_t$ in X_{itj}. I collect all 113 judge fixed effects for the post-HB463 period (since 1 of the 114 judges is necessarily omitted) and plot those fixed effects against the calculated defendant composition for that judge (namely, the percentage of defendants that judge sees who are Black) in Figure 12. Figure 12 clearly demonstrates that on average, the judges who see more white defendants respond more to the recommendation than judges who see more Black defendants do. The orange dashed line demonstrates the linear regression line generated when judge fixed effects in the post-period are regressed on the percentages of Black defendants. It has a slope of -0.41, meaning judges who see 10 percentage points more Black defendants respond to the recommendation 4.1 percentage points less. This effect is an economically meaningful 30% drop down from the pooled baseline effect of 15 percentage points.[23]

Why do judges who see more Black defendants respond less to lenient recommendations? There are many reasons why different judges may respond differently to policy reforms. For one, judges with more experience might be less likely to respond to policy changes. If judges who work in counties with more Black defendants are more experienced (perhaps because these are larger counties and thus more competitive elections for judgeships), this could generate the observed relationship. Second, judges who have made decisions that are associated with higher pretrial misconduct could respond less, since they face a higher expected cost of release in changing their threshold. If judges who experience higher failure to appear or new criminal activity in their bail decisions work in the counties with more Black defendants, this could generate the observed effect. One can generate similar hypotheses around judge demographics (race, gender, experience, etc.) and other

[23] The estimated relationship between judge population and judge responsiveness is quantitatively similar across different methodological choices. For instance, the relationship is similar regardless of whether I estimate the post-period judge fixed effects in differences-in-differences or differences-in-differences-in-differences specifications. It is also similar regardless of whether I estimate the specification with no control variables, only case information control variables, or all possible control variables. Figure A.9 in the Appendix illustrates this similarity: the estimated coefficients range from -0.41, which is the smallest in magnitude, to -0.62, which is the largest in magnitude.

Figure 12: Judge Recommendation Responses and Defendant Populations



Notes: Each point in this scatterplot demonstrates a judge's post-HB463 fixed effect (y-axis) and the percentage Black defendants seen by the judge (x-axis). The judge post-HB463 fixed effects come from estimating Specification 3 and extracting the coefficient on the interaction of a specific judge fixed effect and $Post_t$. I use the sample of judges who make at least 50 bail decisions before and after HB463, which yields 114 judges total. Since fixed effects are relative to one omitted judge, there are 113 different points in this plot. The dashed orange line is the estimated linear regression line, when judge fixed effects in the post-period are regressed on the percentages of Black defendants.

county-level characteristics (crime, population, etc.).

To explore whether other factors explain the relationship between judge-post coefficients and racial composition of the defendant population, I collect variables from interest from a variety of sources. First, I use public data on the 2010 general election results in Kentucky to collect information on whether a given judge was contested in the 2010 election, whether anyone in the district (across all divisions) was contested, and the total number of voters in the judge's election. Second, I collect demographic data on judges (gender, race, and numbers of years of experience) from publicly available online information and qualitative conversations with clerks and Administrative Office of the Courts staff. I successfully collected relevant demographics and elections data for 94 of the 114 judges of interest. Lastly, at the judge level, I calculate the judge's failure to appear and rearrest rates before HB63.

I also combine this collection of judge-level variables with county-level variables. Since judges may make decisions for more than one county, I focus on the county where each judge made the most of their decisions. I then collect data on 2010 county population, whether the county is "rural" (defined as having a county population of fewer than 50,000 people in 2010), and county-level crime rates from the 2010 UCR (total crime rate, total index crime rate, property crime rate, and violent crime rate). Using my collected data on judge demographics, election factors, pretrial misconduct rates, and county-level variables, I run a horse race with the share of Black defendants and all these factors to predict the judge-post fixed effects estimated by the prior model. Table 4 shows that the range of collected judge-level and county-level covariates do not drown out the strong relationship between defendant demographics and judge recommendation responsiveness.

These results connect to the results on recommendations changing error costs. If judges thought recommendations give them less political cover in more racially heterogeneous places, then this belief would generate the empirical patterns observed in the data. This suggested mechanism is conceptually similar to results from Feigenberg and Miller (2021), which show a relationship between local punishment severity and racial heterogeneity. In the cross-section, they find that punishment severity peaks where the Black share of defendants is around 40%.[24] Both my results and those of Feigenberg and Miller (2021) are consistent with public scrutiny of lenient criminal justice reforms being lower in places that are more racially homogeneous.

[24] In Kentucky, the largest shares of Black defendants map onto the peak area of Feigenberg and Miller (2021)'s U-shape relationship between racial heterogeneity and severity. My results are consistent with a model in which recommendations unevenly shield judges from the costs of lenient errors.

Table 4: Explaining Judge Responsiveness to Lenient Recommendations

	Dependent Variable = Judge x Post FE					
	(1)	(2)	(3)	(4)	(5)	(6)
Share Black Defendants	−0.374***	−0.384***	−0.377**	−0.323**	−0.307*	−0.374**
	(0.081)	(0.085)	(0.144)	(0.149)	(0.169)	(0.178)
Black Judge		0.064	0.076	0.064	0.058	0.065
		(0.067)	(0.070)	(0.070)	(0.073)	(0.073)
Woman Judge		−0.025	−0.017	−0.020	−0.019	−0.020
		(0.024)	(0.026)	(0.026)	(0.028)	(0.027)
Years as Judge		−0.002	−0.003	−0.002	−0.002	−0.001
		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Contested in 2010			−0.038	−0.033	−0.034	−0.025
			(0.033)	(0.033)	(0.035)	(0.035)
Any Contest in District in 2010			0.011	0.002	0.003	−0.014
			(0.035)	(0.036)	(0.036)	(0.037)
log(Election Voters)			−0.003	−0.004	0.001	−0.008
			(0.022)	(0.025)	(0.042)	(0.042)
Rearrest Rate Pre-HB463				0.582	0.578	0.634*
				(0.371)	(0.377)	(0.381)
FTA Rate Pre-HB463				−0.180	−0.182	−0.201
				(0.374)	(0.378)	(0.374)
log(County Population)					−0.011	−0.032
					(0.042)	(0.044)
Rural County					−0.017	−0.029
					(0.043)	(0.045)
Total Crime Rate						−0.00004
						(0.00004)
Total Index Crime Rate						0.010
						(0.007)
Property Crime Rate						−0.010
						(0.007)
Violent Crime Rate						−0.009
						(0.007)
Constant	0.116***	0.140***	0.184	0.146	0.223	0.522
	(0.018)	(0.025)	(0.200)	(0.218)	(0.299)	(0.333)
N	94	94	94	94	94	94
R^2	0.188	0.217	0.230	0.252	0.253	0.313
Adjusted R^2	0.179	0.181	0.168	0.172	0.153	0.181

Notes: This table shows the estimated coefficients from regressing post-HB463 judge fixed effects on judge- and county-level characteristics. Judge-level characteristics include share of Black defendants seen by the judge, whether the judge is Black, whether the judge is a woman, number of years of experience as a judge (as of 2011), whether the judge was contested in their 2010 election, whether the judge was elected in a district where any judge faced a contest in 2010, number of voters in the judge's election, the judge's pre-HB463 rearrest rate, and the judge's pre-HB463 failure to appear rate. County-level characteristics refer to characteristics for the county where each judge made the most decisions in the data (because judges can make decisions in multiple counties). County-level characteristics include county population, whether the county is rural (under 50,000 people), total crime rate, total index crime rate, property crime rate, and violent crime rate. *p<0.1; **p<0.05; ***p<0.01.

My results complicate notions of designing algorithms to reduce group inequality. If adherence to algorithmic recommendations varies across decision-makers, these tools can widen group inequality even when they are designed to alleviate it.

7 Conclusion

As algorithms continue to be integrated into high-stakes decisions, studying their impact on human decisions becomes increasingly necessary. While algorithms' predictions are often translated into recommendations for decision-makers, there is little research on the independent effects of such recommendations. In this paper, I highlight the importance of algorithmic recommendations by demonstrating their independent causal effects on high-stakes decisions.

I demonstrate three key results in this paper. First, recommendations meaningfully impact human decision-making. Judges make lenient bail decisions 50% more often when given a lenient recommendation, and these effects are even larger for marginal cases (55%-70%). Second, I provide evidence that recommendations matter for decision-making because they can change the costs of errors to decision-makers. Namely, judges may be more lenient when their choices are consistent with recommendations because the recommendation can shield them from political backlash. Third, I show that lenient recommendations can have heterogeneous effects: judges deviate from recommendations more for Black defendants than for white defendants with identical algorithmic scores.

Importantly, my results show that algorithmic recommendations have first-order effects on decisions. They change more than just the allocation of decisions (who gets which decision); they change the overall composition of decisions (how many decisions are lenient). If decision-makers and algorithmic designers are in conflict about the costs of type II errors, then recommendations are one way to design systems to better align decision-maker incentives with social planner objectives (e.g., less money bail) (McLaughlin and Spiess 2022). Moreover, the interaction between the political economy of decision-making and algorithm-based systems can have unintended effects on racial gaps.

References

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology." *Working Paper 31422, National Bureau of Economic Research.*

Alexander, Michelle. 2018. "The Newest Jim Crow." *New York Times.* https://www.nytimes.com/2018/11/08/opinion/sunday/criminal-justice-reforms-race-technology.html.

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

Austin, James, Roger Ocker, and Avi Bhati. 2010. "Kentucky Pretrial Risk Assessment Instrument Validation." *Bureau of Justice Statistics, Grant,* no. 2009-DB.

Berk, Richard. 2017. "An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism." *Journal of Experimental Criminology* 13 (2): 193–216.

Calonico, Sebastian, Matias D Cattaneo, and Rocio Titiunik. 2015. "Rdrobust: An r Package for Robust Nonparametric Inference in Regression-Discontinuity Designs." *R Journal* 7 (1): 38–51.

Cattaneo, Matias D, Michael Jansson, and Xinwei Ma. Forthcoming. "Local Regression Distribution Estimators." *Journal of Econometrics,* Forthcoming.

———. 2020. "Simple Local Polynomial Density Estimators." *Journal of the American Statistical Association* 115 (531): 1449–55.

———. 2022. "Lpdensity: Local Polynomial Density Estimation and Inference." *Journal of Statistical Software* 101 (2): 1–25.

Christin, Angèle. 2017. "Algorithms in Practice: Comparing Web Journalism and Criminal Justice." *Big Data & Society* 4 (2): 1–14.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. "Algorithmic Decision Making and the Cost of Fairness." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 797–806. Association for Computing Machines.

Covert, Bryce. 2022. "Why New York Jail Populations Are Returning to Pre-Pandemic Levels." *The Appeal.*

Cowgill, Bo. 2018a. "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening." *Research Paper, Columbia Business School.*

———. 2018b. "The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities." *Research Paper, Columbia Business School.*

Cowgill, Bo, and Catherine E Tucker. 2019. "Economics, Fairness and Algorithmic Bias." *Research Paper, Columbia Business School*.

DeMichele, Matthew, Peter Baumgartner, Kelle Barrick, Megan Comfort, Samuel Scaggs, and Shilpi Misra. 2018. "What Do Criminal Justice Professionals Think about Risk Assessment at Pretrial?" *Research Paper, RTI International*.

Doleac, Jennifer, and Megan Stevenson. 2018. "The Roadblock to Reform." *Research Report, American Constitution Society*.

Electronic Privacy Information Center. 2020. "Liberty at Risk: Pre-trial Risk Assessment Tools in the U.S." *Research Report, Electronic Privacy Information Center*.

Feigenberg, Benjamin, and Conrad Miller. 2021. "Racial Divisions and Criminal Justice: Evidence from Southern State Courts." *American Economic Journal: Economic Policy* 13 (2): 207–40.

Fung, Katherine. 2021. "Darrell Brooks Should Not Have Been Released on Low Bail, Milwaukee DA Admits." *Newsweek*. https://www.newsweek.com/darrell-brooks-should-not-have-been-released-low-bail-milwaukee-da-admits-1652059.

Garrett, Brandon L, and John Monahan. 2018. "Judging Risk." *Working Paper, Duke University*.

Gruber, Jonathan, Benjamin R Handel, Samuel H Kina, and Jonathan T Kolstad. 2020. "Managing Intelligence: Skilled Experts and AI in Markets for Complex Products." *Working Paper 27038, National Bureau of Economic Research*.

Hausman, Catherine, and David S Rapson. 2018. "Regression Discontinuity in Time: Considerations for Empirical Applications." *Annual Review of Resource Economics* 10: 533–52.

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li. 2017. "Discretion in Hiring." *Quarterly Journal of Economics* 133 (2): 765–800.

Khullar, Dhruv. 2023. "Can A.I. Treat Mental Illness?" *New Yorker*. https://www.newyorker.com/magazine/2023/03/06/can-ai-treat-mental-illness.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133 (1): 237–93.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. 2018. "Discrimination in the Age of Algorithms." *Journal of Legal Analysis* 10: 113–74.

Kleinberg, Jon, and Sendhil Mullainathan. 2019. "Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability." Working Paper 25854, National Bureau of Economic Research.

Laura and John Arnold Foundation. 2018. "Pretrial Justice." https://www.

arnoldfoundation.org/initiative/criminal-justice/pretrial-justice/.

Lum, Kristian, and William Isaac. 2016. “To Predict and Serve?” *Significance* 13 (5): 14–19.

McCrary, Justin. 2008. “Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test.” *Journal of Econometrics* 142 (2): 698–714.

McLaughlin, Bryce, and Jann Spiess. 2022. “Algorithmic Assistance with Recommendation-Dependent Preferences.” *arXiv Preprint, arXiv:2208.07626*.

Mullainathan, Sendhil, and Ziad Obermeyer. 2022. “Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care.” *Quarterly Journal of Economics* 137 (2): 679–727.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” *Science* 366 (6464): 447–53.

Pruss, Dasha. 2023. “Ghosting the Machine: Judicial Resistance to a Recidivism Risk Assessment Instrument.” *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 312–23.

Skeem, Jennifer L, Nicholas Scurich, and John Monahan. 2019. “Impact of Risk Assessment on Judges’ Fairness in Sentencing Relatively Poor Defendants.” *Research Paper No. 2019-02, Virginia Public Law and Legal Theory Research Paper*, no. 2019-02.

Sloan, CarlyWill, George Naufal, and Heather Caspers. 2018. “The Effect of Risk Assessment Scores on Judicial Behavior and Defendant Outcomes.” *Discussion Paper No. 11948, IZA Institute of Labor Economics*.

Stevenson, Megan. 2018. “Assessing Risk Assessment in Action.” *Minnesota Law Review* 103: 303–83.

Appendix

A.1 Kentucky Pretrial Services Institutional Details

Kentucky has one pretrial services agency, which serves all 120 counties in the state, meaning that data management and collection is unified and well-organized.[25]

Kentucky is well-known for its pretrial services for a few reasons. For one, it was the first state to ban commercial bail bonds in 1976. (It was one of four states with this ban as of 2018.) While pretrial employees are housed in individual counties, they do not work for the individual counties. As of January 2019, there were about 251 employees in Pretrial Services in Kentucky. Approximately 202 employees are pretrial officers and/or supervisors, and 49 are risk assessment specialists and/or coordinators. Kentucky was also the first jurisdiction to pilot the Public Safety Assessment (PSA) risk assessment, which is now used in dozens of jurisdictions across the US.

What information do judges have during bail decisions? Because bail decisions in Kentucky occur over the phone, I cannot directly observe the relevant conversations. However, in 2019, there were eight examples of judge calls available on the Kentucky pretrial website. I listened them. These calls included the following information: name, age, risk score information, list of charges, and incident description. The incident description quotes information from the police report.

Note that while demographic information on race or gender may not be explicit in the call, these details are implicitly included. Gender is revealed through use of pronouns (e.g., "he" and "she") when the pretrial officer discusses the defendant. Meanwhile, names (especially in combination with the county) can signal information about race. Moreover, race and ethnicity were on judge forms about cases during my time period of interest, meaning they could be explicitly observed if judges looked at said forms in their decision-making. (However, these details have since been removed from judge forms.)

In Kentucky, if the defendant has not posted bail within 24 hours of the initial decision, the pretrial officer informs the court, and the judge can change the bail decision to increase the chance that they can be released pretrial. If the defendant remains detained pretrial, the next time bail could be reconsidered is usually first appearance.

[25] Unlike in other states, Kentucky pretrial services is part of the judicial branch; it is a state entity that works for the courts (and is state-funded). Information in the following paragraphs is sourced from an interview with Tara Blair, former executive officer of Kentucky Pretrial Services.

A.2 Kentucky Risk Assessment Institutional Details

Kentucky has used a few different risk assessment scoring tools over the years. The first scoring tool was a six-question tool developed by the Vera Institute. In 2006, Kentucky moved to the Kentucky Pretrial Risk Assessment (KPRA) tool. In July 2013, Kentucky started using the Public Safety Assessment (PSA) tool, which was developed by the Laura and John Arnold Foundation.

Although Kentucky used the KPRA tool from 2006 to 2013, the algorithm changed slightly on March 18, 2011 (Austin, Ocker, and Bhati 2010). Because of these changes, I use data after March 18, 2011, but before adoption of the PSA tool to focus on a time period in which there were no changes to the algorithm.

The KPRA is a checklist-style instrument. Table A.1 documents how to calculate the score for the post-March 18 version of the tool. There were 12 risk score factors, which were "yes" or "no" questions. Each "yes" or "no" answer was associated with a set number of points. Pretrial officers calculated the total of the 12 numbers associated with the relevant questions to generate the final risk score, which was between 0 (lowest) and 24 (highest). Pretrial officers then converted the risk scores into risk levels, which they provided to judges. Scores of 0-5 were categorized as "low risk," scores of 6-13 were categorized as "moderate risk," and scores of 14-24 were categorized as "high risk."

Table A.2 documents how the risk score was calculated before March 18. Relative to the post-March 18 method, this one featured one additional question (Item 0, which is about references), and the weights for 7 question responses were different. In addition, the way risk scores were converted to levels was slightly different: scores of 0-5 were categorized as "low risk," scores of 6-12 were categorized as "moderate risk," and scores of 13-24 were categorized as "high risk."

Table A.1: Kentucky Pretrial Risk Assessment Factors (After March 18, 2011)

Factor #	**Risk Score Question**	**"Yes" Points**	**"No" Points**
1	Does the defendant have a verified local address and has the defendant lived in the area for the past twelve months?	0	2
2	Does the defendant have a verified sufficient means of support?	0	1
3	Is the defendant's current charge a Class A, B, or C Felony?	1	0
4	Is the defendant charged with a new offense while there is a pending case?	7	0
5	Does the defendant have an active warrant(s) for Failure to Appear prior to disposition? If no, does the defendant have a prior FTA for felony or misdemeanor?	2	0
6	Does the defendant have a prior FTA on his or her record for a criminal traffic violation?	1	0
7	Does the defendant have prior misdemeanor convictions?	2	0
8	Does the defendant have prior felony convictions?	1	0
9	Does the defendant have prior violent crime convictions?	1	0
10	Does the defendant have a history of drug/alcohol abuse?	2	0
11	Does the defendant have a prior conviction for felony escape?	3	0
12	Is the defendant currently on probation/parole from a felony conviction?	1	0

Notes: This table shows the weights associated with risk score factors in the KPRA after March 18, 2011. To calculate total risk score, pretrial officers added up the points associated with each answer. Item 1 was a "yes" if at least five people (reached via the defendant's cell phone) were able to verify the defendant's local address and confirm they had lived in the area for the past twelve months. Item 2 was a "yes" if a defendant was one or more of the following: employed full-time, the primary caregiver of a child or disabled relative, a Social Security/disability recipient, employed part-time or a part-time student, a full-time student, retired, or living with someone who supported them. Item 11 was a "yes" if the defendant had 3 or more drug- or alcohol-related convictions in the last 5 years (a longer period was considered if the defendant had been incarcerated at some point).

Table A.2: Kentucky Pretrial Risk Assessment Factors (Before March 18, 2011)

Factor #	**Risk Score Question**	**"Yes" Points**	**"No" Points**
0	Did a reference verify that he or she would be willing to attend court with the defendant or sign a surety bond?	0	1
1	Does the defendant have a verified local address and has the defendant lived in the area for the past twelve months?	0	1
2	Does the defendant have a verified sufficient means of support?	0	1
3	Is the defendant's current charge a Class A, B, or C Felony?	1	0
4	Is the defendant charged with a new offense while there is a pending case?	5	0
5	Does the defendant have an active warrant(s) for Failure to Appear prior to disposition? If no, does the defendant have a prior FTA for felony or misdemeanor?	4	0
6	Does the defendant have a prior FTA on his or her record for a criminal traffic violation?	1	0
7	Does the defendant have prior misdemeanor convictions?	1	0
8	Does the defendant have prior felony convictions?	1	0
9	Does the defendant have prior violent crime convictions?	2	0
10	Does the defendant have a history of drug/alcohol abuse?	2	0
11	Does the defendant have a prior conviction for felony escape?	1	0
12	Is the defendant currently on probation/parole from a felony conviction?	2	0

Notes: This table shows the weights associated with risk score factors in the KPRA before March 18, 2011. To calculate total risk score, pretrial officers added up the points associated with each answer. Item 1 was a "yes" if at least five people (reached via the defendant's cell phone) were able to verify the defendant's local address and confirm they had lived in the area for the past twelve months. Item 2 was a "yes" if a defendant was one or more of the following: employed full-time, the primary caregiver of a child or disabled relative, a Social Security/disability recipient, employed part-time employee or a part-time student, a full-time student, retired, or living with someone who supported them. Item 11 was a "yes" if the defendant had 3 or more drug- or alcohol-related convictions in the last 5 years (a longer period was considered if the defendant had been incarcerated at some point).

A.3 Additional Tables and Figures

Table A.3: Differences-in-Differences Results across Bandwidths and Specifications

	Dependent variable: I(lenient bail)		
	Bandwidth: All Risk Scores		
	(1)	(2)	(3)
I(score<14) x Post	0.153***	0.152***	0.163***
	(0.020)	(0.020)	(0.019)
Pre-Mean Score<14	0.310	0.310	0.310
Time/Score FEs	Y	Y	Y
Charge/judge/county/demo controls	Y	Y	N
Risk component controls	Y	N	N
Observations	153,597	153,597	153,597
R^2	0.268	0.263	0.132
Adjusted R^2	0.265	0.260	0.132

	Dependent variable: I(lenient bail)					
	Bandwidth: 10-17			Bandwidth: 11-16		
	(4)	(5)	(6)	(7)	(8)	(9)
I(score<14) x Post	0.133***	0.127***	0.131***	0.120***	0.115***	0.128***
	(0.020)	(0.020)	(0.021)	(0.021)	(0.021)	(0.021)
Pre-Mean Score<14	0.189	0.189	0.189	0.185	0.185	0.185
Time/Score FEs	Y	Y	Y	Y	Y	Y
Charge/judge/county/demo controls	Y	Y	N	Y	Y	N
Risk component controls	Y	N	N	Y	N	N
Observations	29,006	29,006	29,006	21,785	21,785	21,785
R^2	0.183	0.173	0.055	0.181	0.174	0.057
Adjusted R^2	0.168	0.159	0.053	0.161	0.155	0.055

	Dependent variable: I(lenient bail)					
	Bandwidth: 12-15			Bandwidth: 13-14		
	(10)	(11)	(12)	(13)	(14)	(15)
I(score<14) x Post	0.115***	0.113***	0.124***	0.092***	0.091***	0.102***
	(0.023)	(0.022)	(0.023)	(0.032)	(0.032)	(0.033)
Pre-Mean Score<14	0.178	0.178	0.178	0.171	0.171	0.171
Time/Score FEs	Y	Y	Y	Y	Y	Y
Charge/judge/county/demo controls	Y	Y	N	Y	Y	N
Risk component controls	Y	N	N	Y	N	N
Observations	15,273	15,273	15,273	7,854	7,854	7,854
R^2	0.173	0.170	0.053	0.173	0.171	0.043
Adjusted R^2	0.146	0.143	0.051	0.121	0.119	0.039

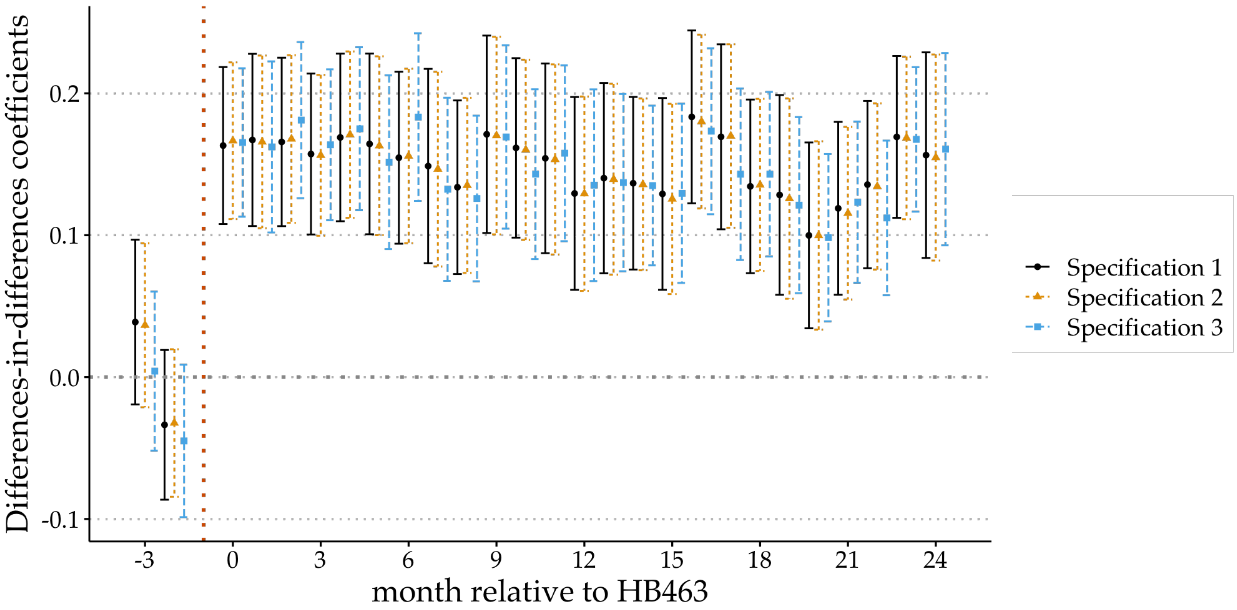
Notes: This table displays the estimated differences-in-differences coefficients in specifications in which the dependent variable is lenient bail. The table shows results across different bandwidths – 10-17, 11-16, 12-15, and 13-14 – as well as across different specifications. The first column within each bandwidth controls for fixed effects for month-year, day of week, and exact risk score. The second also includes controls for top charge, judge, county, and defendant demographics (gender, race). The third also controls for all the characteristics which factor into risk score, listed in Table A.1. Standard errors are always clustered at the judge-level. *p<0.1; **p<0.05; ***p<0.01.

Table A.4: Fraction of Defendants with Different Risk Score Factors by Defendant Race

Risk Component	**Black**	**White**	**p-value**
Verified Address	0.157	0.110	<0.001
Verified Support	0.586	0.589	0.5
Felony Charge	0.144	0.097	<0.001
Pending Charge	0.211	0.207	0.088
Failure to Appear (measure 1)	0.426	0.274	<0.001
Failure to Appear (measure 2)	0.246	0.225	<0.001
Prior Misdemeanor Conviction	0.768	0.715	<0.001
Prior Felony Conviction	0.420	0.272	<0.001
Prior Violent Conviction	0.359	0.229	<0.001
Drug/Alcohol Abuse	0.112	0.110	0.4
Prior Felony Escape Charge	0.035	0.011	<0.001
Felony Probation/Parole	0.160	0.103	<0.001

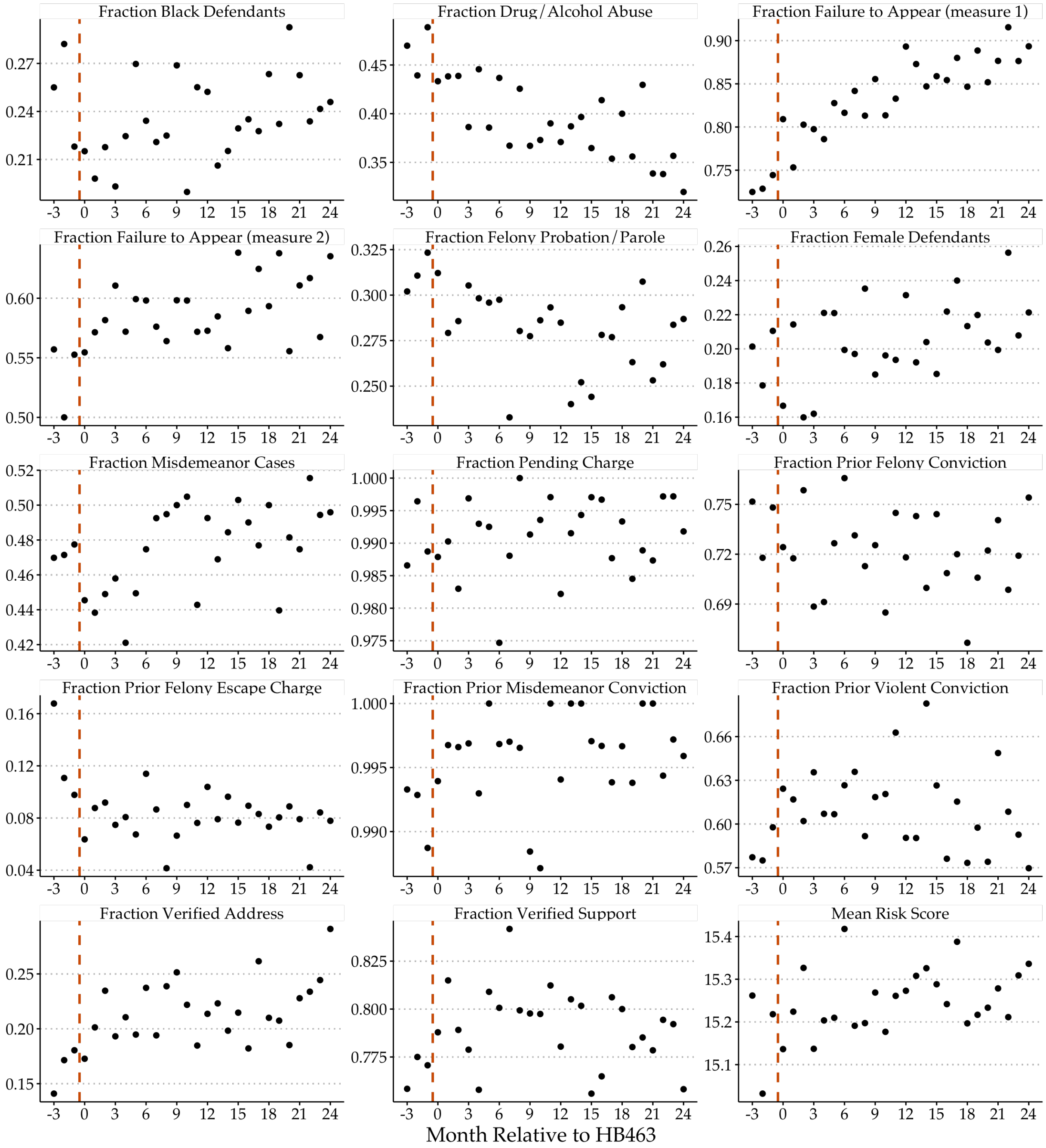
Notes: This table shows what fraction of defendants have characteristics that factor into the overall risk score. There are 24,352 Black defendants and 128,878 white defendants. The p-value tests if the fractions are statistically different in the Black and white defendant populations.

Figure A.1: Dynamic Differences-in-Differences Estimates across Specifications



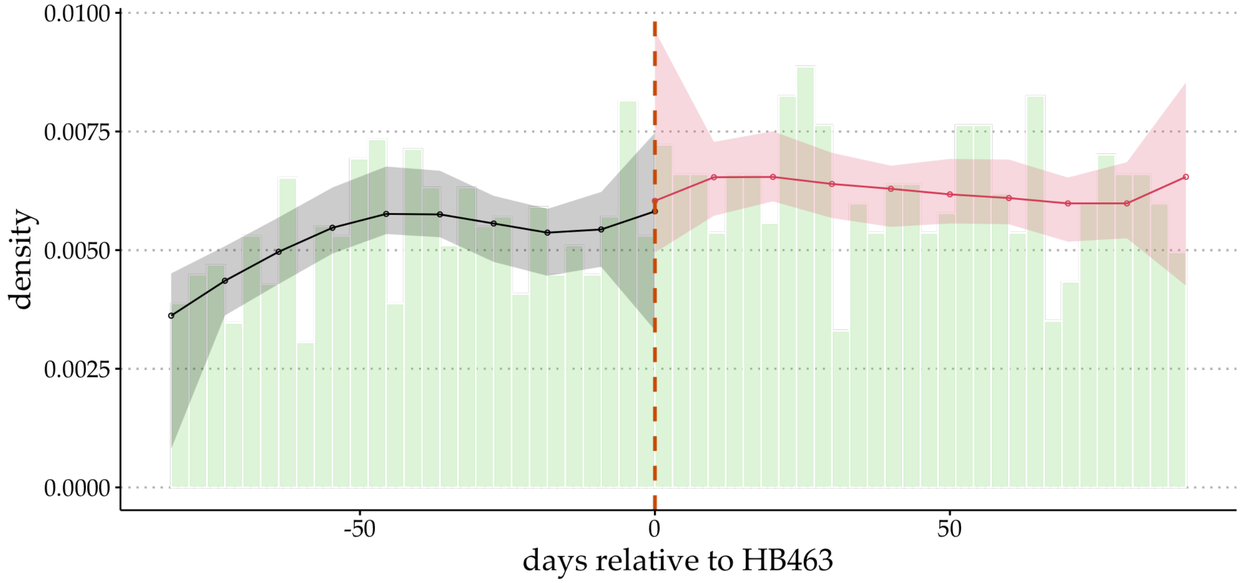
Notes: This figure shows the difference-in-differences coefficients for months relative to recommendation introduction across specifications. The orange dashed line denotes the omitted period of the month before recommendation introduction. Specification 1 (black circles and error bars) is the main specification, also shown in Figure 6, which includes controls for day of week, month-year, exact risk score, top charge level and class, defendant demographics (race, gender), judge, county, and all risk score components listed in Table A.1 except for verified address and support. Specification 2 (orange triangles and dotted error bars) includes controls for day of week, month-year, exact risk score, top charge level and class, defendant demographics (race, gender), judge, and county. Finally, Specification 3 (blue squares and dashed error bars) includes controls only for day of week, month-year, and exact risk score.

Figure A.2: Defendant and Case Covariates over Time



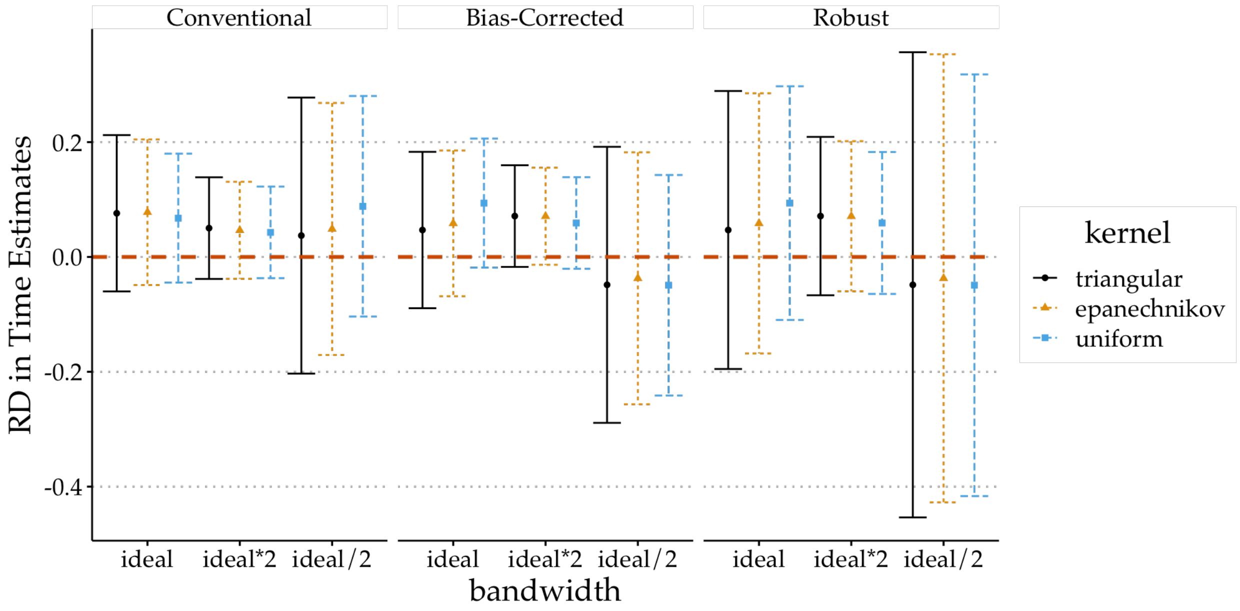
Notes: This figure shows average defendant and case covariates binned by month for the sample of high risk level cases. Months are indexed relative to the introduction of algorithmic recommendations due to HB463. The orange dotted line shows when HB463 went into effect.

Figure A.3: McCrary Density Test



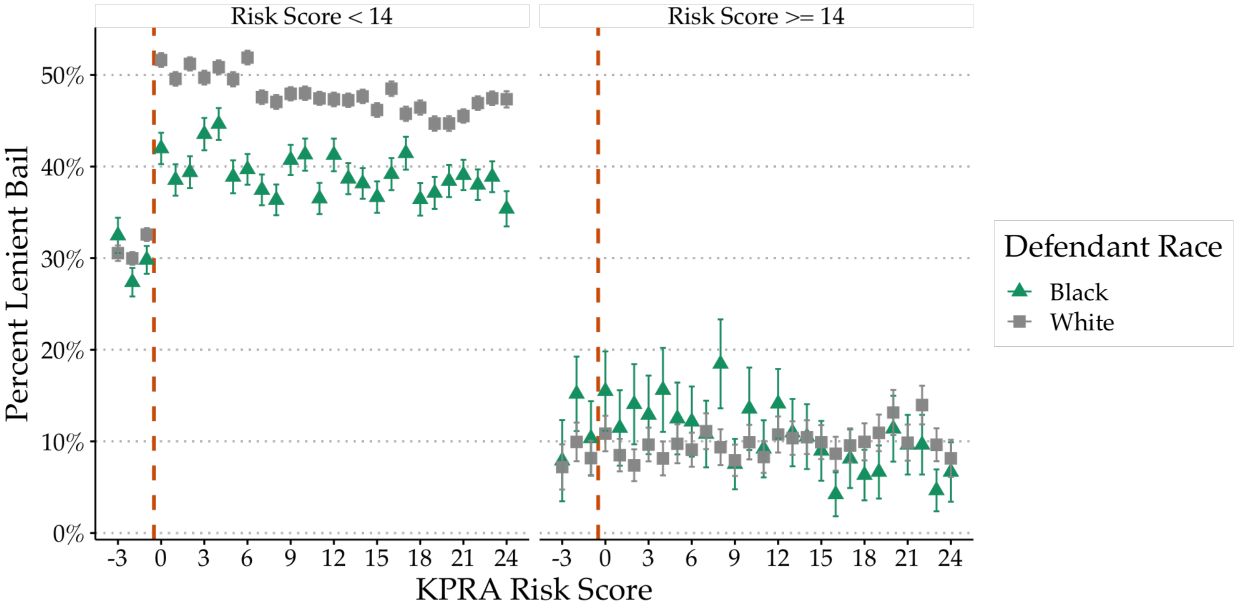
Notes: This figure shows a McCrary density test. The data is grouped into bins, and the figure plots the relative frequency of observations in each bin (in light green bars) as well as the averages (dots) and confidence intervals (shaded regions) of those bins. Black dots and regions are for days before HB463, and red dots and regions are for days after HB463.

Figure A.4: RD in Time Robustness Across Methods



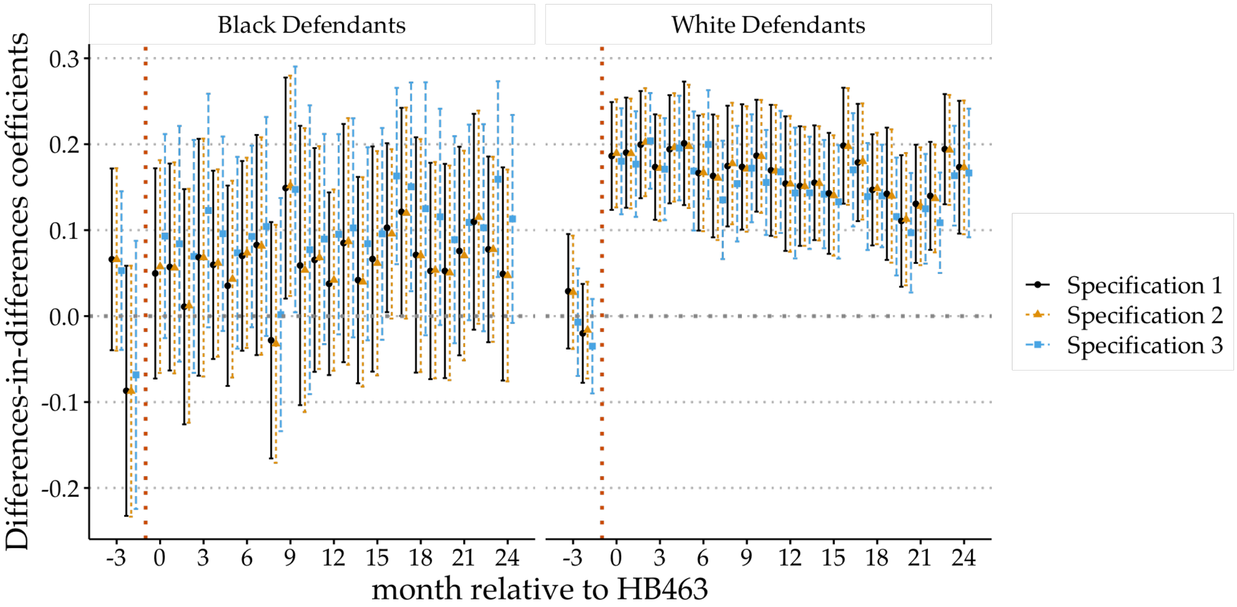
Notes: This figure plots the regression discontinuity in time coefficients across a range of bandwiths, kernels, and estimation adjustments. The plots are grouped based on whether I use conventional, bias-corrected, or robust bias-corrected confidence intervals. Within plots, the x-axis shows the bandwidth used and "ideal" refers to the mean square error optimal bandwidth. I also show results for twice that bandwidth and one-half of that bandwidth. Estimates across kernals are illustrated with distinct colors, shapes, and line types. The default estimate uses a triangular kernel, optimal bandwidth, and conventional confidence intervals.

Figure A.5: Lenient Bail Rates by Risk Group and Race over Time



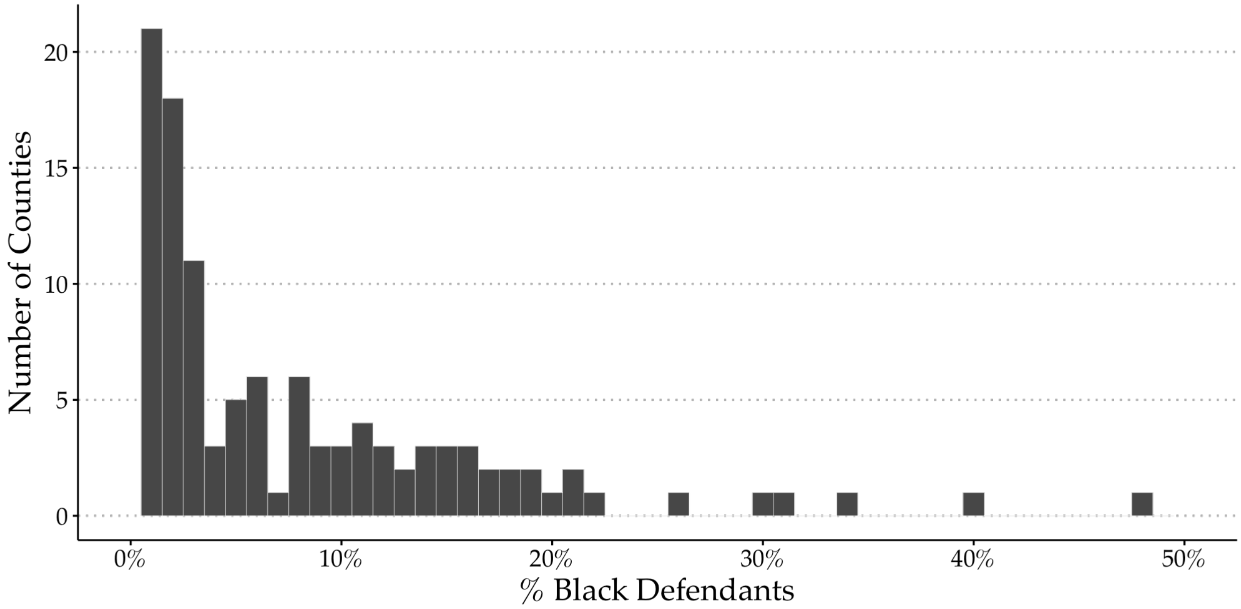
Notes: This figure shows the rate of lenient bail over time for Black and white defendants separately for cases with risk scores over or under the critical threshold. Months are indexed relative to the introduction of algorithmic recommendations due to HB463. The orange dotted line shows when HB463 went into effect. Cases with Black defendants are shown as green triangles, while cases with white defendants are shown as gray squares. Rates are shown for cases with risk scores under 14 on the left and for cases with risk scores at or over 14 on the right.

Figure A.6: Dynamic Differences-in-Differences Estimates across Specifications, by Race



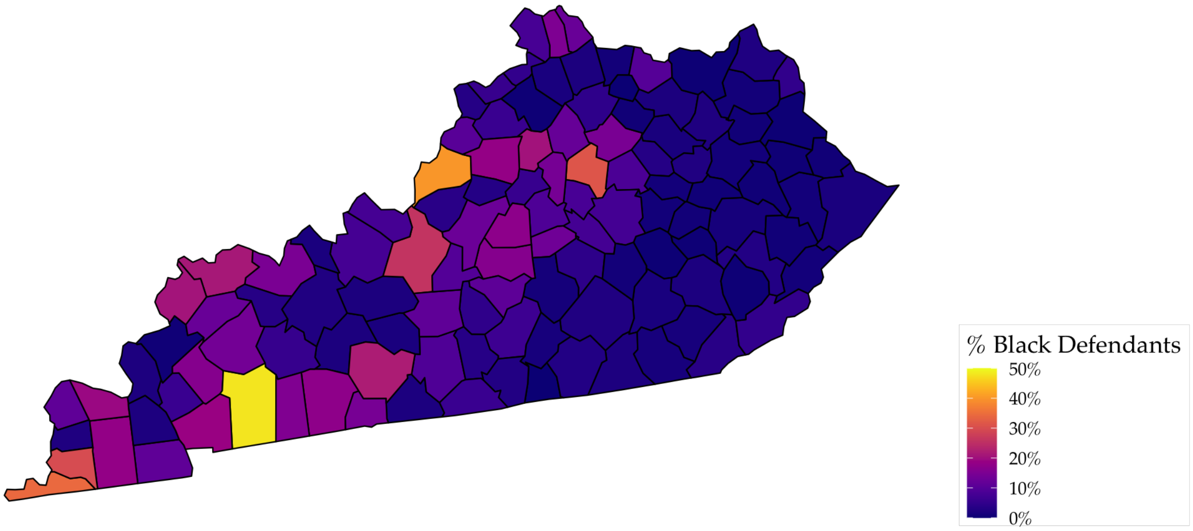
Notes: This figure shows the difference-in-differences coefficients for months relative to recommendation introduction across specifications for both white and Black defendants. The orange dashed line denotes the omitted period of the month before recommendation introduction. Specification 1 (black circles and error bars) is the main specification, also shown in Figure 6, which includes controls for: day of week, month-year, exact risk score, top charge level and class, defendant demographics (race, gender), judge, county, and all risk score components listed in Table A.1 except for verified address and support. Specification 2 (orange triangles and dotted error bars) includes controls for day of week, month-year, exact risk score, top charge level and class, defendant demographics (race, gender), judge, county. Finally, Specification 3 (blue squares and dashed error bars) includes controls only for day of week, month-year, and exact risk score.

Figure A.7: Histogram of Counties by Defendant Racial Composition



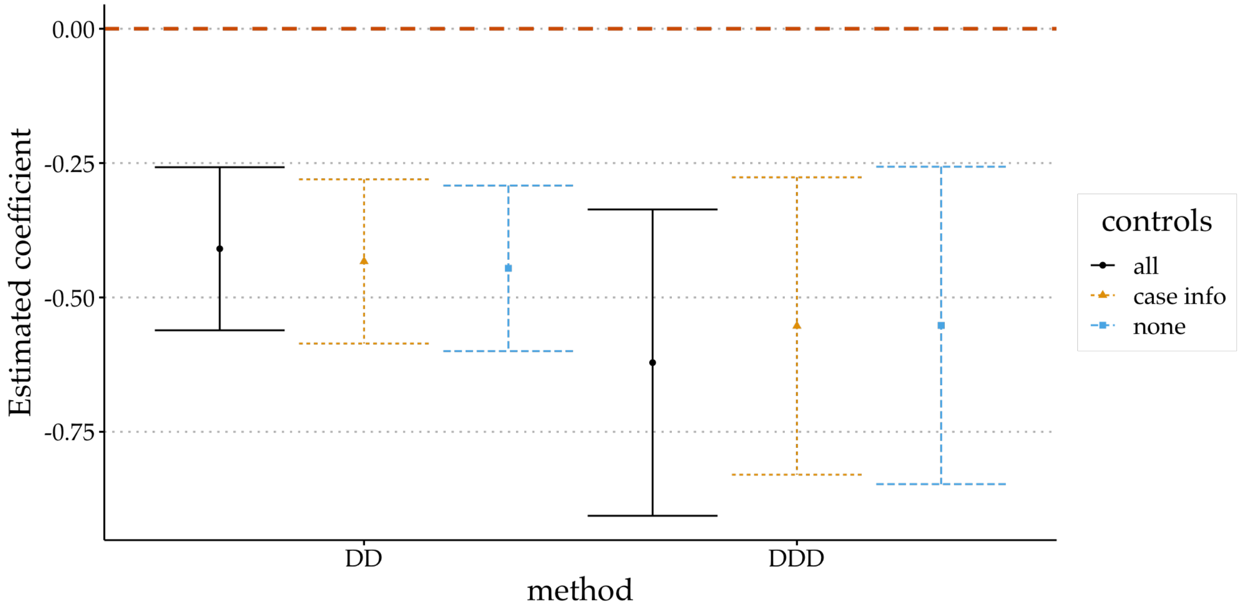
Notes: This figure is a histogram that displays the number of counties with different percentages of Black defendants. The binwidths are each 1 percentage point.

Figure A.8: Choropleth of Counties by Defendant Racial Composition



Notes: This figure displays the percentage of defendants who are Black for each of Kentucky's 120 counties.

Figure A.9: Robustness of Relationship between Racial Composition of Judge Population and Judge Response to Recommendation



Notes: This figure plots the estimated coefficients from regressing judge fixed effects in the post-period on percentages of Black defendants. I use the sample of judges who make at least 50 bail decisions before and after HB463, which yields 114 judges total. In the case of the three estimates on the left, fixed effects are extracted from differences-in-differences approaches ("DD") interacting a post-period indicator and an indicator for a defendant being Black (using only the subset of data with risk scores under 14). In the case of the three estimates illustrated on the right, fixed effects are extracted from triple differences specifications ("DDD") interacting a post-period indicator, an indicator for a defendant being Black, and an indicator for the risk score being under 14. Estimates across different sets of control variables are illustrated with distinct colors, shapes, and line types.