The Hidden Effects of Algorithmic Recommendations

Alex Albright*

March 26, 2024

Abstract

Algorithms are intended to improve human decisions with data-driven predictions. However, algorithms provide more than just predictions to decision-makers — they often provide explicit recommendations. In this paper, I demonstrate these algorithmic recommendations have significant independent effects on human decisions. I leverage a natural experiment in which algorithmic recommendations were given to bail judges in some cases but not others. Lenient recommendations increased lenient bail decisions by 50% for marginal cases. The results are consistent with algorithmic recommendations making visible mistakes, such as violent rearrest, less costly to judges by providing them reputational cover. Algorithms can change human decisions by shifting incentives, in addition to directly providing new prediction information. Finally, I show there is variation across decision-makers in adherence to recommendations, which generated unintended effects across racial groups. Lenient recommendations increased Black-white gaps in lenient bail for 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 Larry Katz, Winnie Van Dijk, Ed Glaeser, Megan Stevenson, Jennifer Doleac, Alicia Modestino, Elior Cohen, Abbie Wozniak, Andrew Goodman-Bacon, Mike Mueller-Smith, Crystal Yang, Alma Cohen, Mandy Pallais, Louis Kaplow, Adam Soliman, and presentation audiences at Harvard, LSU, Macalester, Williams, St. Olaf, Minneapolis Fed, Clemson, University of Minnesota, FGV EESP, Insper, PUC-Rio, SDSU, WEC Jr. (Northeastern), Wiser (Kansas City Fed), SOLE, and ASSAs for constructive feedback and comments. 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 James Holt and Sara Brandel for their invaluable editorial and MOU support, respectively. Amisha Kambath provided 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.

1 Introduction

Predictive algorithms are used in many high-stakes decisions. Algorithms predicting default are used in granting loans, algorithms predicting self-harm are used in mental health treatment, and algorithms predicting rearrest are used in criminal justice. Despite their prevalence, it is still the norm that humans (loan officers, therapists, judges) – not the algorithms – make the final decisions that govern outcomes. Therefore, understanding how algorithms change outcomes in these systems requires understanding how algorithms change human decisions.

The conventional wisdom is that algorithms impact human decisions because they provide decision-makers with data-driven predictions, but they can do more than that. They often give explicit recommendations: the loan algorithm can recommend rejection, the mental health algorithm can recommend hospitalization, and the pretrial algorithm can recommend release. These *algorithmic recommendations* are distinct from predictions; recommendations are the result of a normative mapping from predictions to actions, and many different recommendations can be consistent with identical underlying predictions.[1] Despite the distinction between predictions and recommendations, they are usually conflated under the catch-all term of "algorithm." Therefore, most attempts to estimate the effects of algorithms muddle the impact of a new prediction technology with the impact of setting normative recommendations. In this paper, I disentangle the two to isolate the hidden effects of algorithmic recommendations on human decisions.

It is an empirical challenge to isolate the effects of algorithmic recommendations for a few reasons. For one, the institutional details around how predictive algorithms are developed and used in high-stakes settings are often opaque, which can impede careful study. Moreover, even if the details are transparent, algorithmic predictions and recommendations are often introduced simultaneously, which makes isolating the two difficult. I progress on this front by leveraging a natural experiment in which algorithmic predictions given to decision-makers stayed the same, but 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. Bail decisions are important in the US criminal legal system because they set the conditions for defendants' release from jail after arrest. During my study period, judges making bail decisions

[1] As Cowgill and Stevenson (2020) write, "The choice of policy objectives is a separate issue from the utility of accurate predictions."

received information on the predicted risk of pretrial misconduct (rearrest or failure to appear in court) for each case. Before June 2011, judges were given algorithmic predicted risk but were not given any recommendations. In June 2011, judges began receiving recommendations, but only for cases with algorithmic predicted risk below a discrete cut-off. Judges were recommended not to set money bail for these lower risk cases. (Not setting money bail is a more lenient decision than setting money bail, because the latter requires defendants to post money for release from jail, while the former does not. Therefore, I call the recommendations introduced in 2011 for the lower-risk cases "lenient bail" recommendations.) The institutional details yield useful variation for causal inference: lenient bail recommendations vary across the risk score distribution (over and under the cut-off) and across time periods (before and after June 2011).

Why might these recommendations change decision-maker behavior? I develop a model of bail decision-making in which judges make decisions based on their prediction of pretrial misconduct (a bad outcome) and the perceived costs of pretrial detention and misconduct. I demonstrate how introducing algorithmic recommendations changes judge behavior under two distinct theories. The first theory is that recommendations only impact decision-maker predictions of misconduct (as conventional wisdom assumes). The second theory is that recommendations can change the *cost* of misconduct because visible mistakes (e.g., a rearrest) become less costly for judges whose decisions adhere to recommendations and more costly for judges whose decisions deviate from them. (Decision-makers are less liable for mistakes when they go along with recommendations but more liable when they go against recommendations.) The two theories generate dueling testable predictions in the Kentucky empirical setting. If recommendations only change predictions, the new recommendations should have had no effects on bail setting (because algorithmic predictions were already available to judges). But if recommendations change costs, then lenient bail should have increased for lower-risk cases (because they became newly covered by lenient recommendations).

To test these predictions, I estimate the causal effects of algorithmic recommendations. I leverage the fact that only some cases received lenient recommendations in differences-in-differences and differences-in-discontinuities designs. In the differences-in-differences approach, cases with low or moderate risk scores are the treated group because they experienced a change in recommendations at the policy date, while cases with high risk scores are the control group because they experienced no such change. The differences-in-differences design estimates the causal effect of the recommendations for the entire distribution of cases in the low and moderate risk groups.

The differences-in-discontinuities design leverages the fact that the lenient recommendation kicks in at a sharp cut-off in the risk score distribution. After June 2011, cases with the highest moderate risk scores received lenient recommendations, but similarly scoring cases with the lowest high risk scores did not. If the lenient bail recommendation were the only factor that changed discontinuously over the threshold during the post-period, then estimating a simple regression discontinuity would identify the desired lenient recommendation effect. However, other relevant factors changed discontinuously at that threshold as well. These confounding factors around the threshold in the post-period were also present in the pre-period. Therefore, I use a differences-in-discontinuities approach to difference out the regression discontinuity in the pre-period from the regression discontinuity in the post-period and recover the effect of the lenient recommendation in isolation. This method, in contrast to the differences-in-differences approach, estimates the effect of the lenient recommendation for marginal cases, those that are close to the critical moderate-high threshold.

I find that algorithmic recommendations have clear independent effects on human decisions. While the algorithmic risk predictions available to judges stayed the same, algorithmic recommendations significantly changed judges' bail setting rates. Lenient recommendations increased lenient decisions by 50% for low and moderate risk cases. There is no evidence of pre-trends in the differences-in-differences approach, 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. The relative effects are similar at around 50% for the marginal moderate risk cases. The recommendation effects remain economically meaningful and are significant even when conservatively adjusted to test the import of potential concerns to identification. The results consistently show that introducing algorithmic recommendations changes how prediction technologies impact human decisions.

These empirical results are consistent with the theory that algorithmic recommendations change decision-maker costs (and therefore their incentives). In particular, lenient choices may become less costly when they adhere to lenient recommendations. If a lenient choice results in bad outcomes, some of the blame goes to the recommendation designer (the House legislature in the Kentucky case) rather than all of the blame going to the individual decision-maker (the judge in the Kentucky case). This potential mechanism, overlooked thus far in the literature on human-algorithm interaction, means that the communication of predictive algorithms can change human decisions by shifting incentives in addition to directly providing new prediction information.

One important reason why algorithm-based tools have been adopted in criminal legal settings is that they have the potential to standardize decisions across settings. The algorithmic recommendations in this setting are the same for all cases with a given risk score. However, I find that lenient recommendations do not benefit all groups of people with the same score equally. Estimating differences-in-differences specifications separately for cases with white and Black defendants, I find that compared with white defendants with identical risk scores, Black defendants benefit less from receiving a lenient recommendation. Consistent with prior work by Stevenson (2018), this finding 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. The results are consistent with judges perceiving recommendations as providing them less political cover in more racially heterogeneous places. These results on heterogeneous effects show that differential adherence to identical recommendations complicates how algorithm-based policies impact racial disparities, even when we hold algorithm-based predictions of risk constant.

Related literature: Algorithms can outperform human decision-makers in a variety of settings (Berk 2017; Mullainathan and Obermeyer 2022; Kleinberg, Lakkaraju, et al. 2018; Cowgill 2018a). However, because human decision-makers still retain discretion in most settings, studying how algorithms and humans interact is necessary. Recent work has compared outcomes in the absence of algorithms with outcomes when humans use algorithms at their discretion (Sloan, Naufal, and Caspers Forthcoming; Stevenson and Doleac Forthcoming; Garrett and Monahan 2018; DeMichele et al. 2018; Cowgill and Tucker 2019; Agarwal et al. 2023; Davenport 2023). How algorithmic predictions are communicated to human decision-makers varies in these settings. As a result, the existing empirical evidence does not distinguish between the effects of new predictions and new algorithmic recommendations. I contribute to the literature by highlighting the distinction between algorithmic predictions and recommendations, and empirically disentangling the effects of the two.

This paper complements previous research by McLaughlin and Spiess (2022), who formalize the distinction between algorithm predictions and recommendations. The authors develop a theoretical model in which algorithmic recommendations may directly change preferences (rather than recommendations only changing beliefs). The theoretical intuition is similar to how I describe recommendations changing the costs of decision errors.[2]

[2] Another example of a paper that investigates how algorithms may impact decision-makers beyond just

McLaughlin and Spiess (2022) develop theoretical results showing how recommendations may matter independent of predictions, while my paper provides direct empirical evidence of their independent effects in practice. Both papers highlight the importance of studying algorithmic recommendations, but they use different approaches (theoretical and empirical).

My research contributes to a wide range of evidence on how people use discretion when given algorithms. Ethnographic work demonstrates a "decoupling" between how algorithms are expected to be used and how they are used in practice (Christin 2017; Pruss 2023). Decision-makers frequently overrule algorithm recommendations (Hoffman, Kahn, and Li 2017; Gruber et al. 2020; Agarwal et al. 2023) and 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; Stevenson and Doleac Forthcoming). My paper provides evidence that human discretion changes in the presence of algorithmic recommendations, even if algorithmic predictions are unchanged.

Predictive algorithms generate the recommendations I study in this paper. However, algorithmic recommendations are closely related to simpler prediction tools used in the criminal justice system for many decades: sentencing guidelines. Sentencing guidelines were a tool used by legislatures to "[impose] structure on judicial discretion"; a similar idea underlies the design of algorithmic recommendations in the modern era (Bushway, Owens, and Piehl 2012). Bushway, Owens, and Piehl (2012) studied the causal effects of sentencing guidelines, holding other case characteristics constant. The authors leveraged calculation mistakes for identification and found that erroneously high or low recommendations have causal effects on sentencing. Both my paper and this previous research show that advisory legislative recommendations have causal effects on decision-makers in the criminal justice system.

The 2011 Kentucky reform that I leverage for identification was first studied by Stevenson (2018) with interrupted time series methods. Her important paper highlighted the crucial distinction between how algorithms change outcomes in theory and practice. I diverge from and build on her work in a few ways. First, conceptually, I establish and focus on the distinction between algorithmic predictions and recommendations. Second, I sharpen the empirical approach with three steps to isolate the causal effects of recommendations in the Kentucky setting. I use administrative court reports to identify when the algorithmic prediction technology stayed constant but the recommendations changed. I also use

updating their predictions is Davenport (2023). Davenport (2023) develops a model in which algorithms may impact decision-maker "image concerns," which then feed into their subsequent decisions.

internal documents to calculate the underlying risk scores for cases, which allow me to isolate the causal effects of algorithmic recommendations for cases near the critical threshold (differences-in-discontinuities). Finally, I use a variety of methods to bound my effects under different assumptions about confounding variation.

This paper also contributes to a growing 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, Ludwig, et al. 2018; Kleinberg and Mullainathan 2019; Davenport 2023; Stevenson and Doleac Forthcoming). My paper diverges from most related work because I focus on the role of algorithmic recommendations, rather than the underlying algorithm technology, in impacting racial gaps. However, my results still align with several otheres papers using criminal justice system data. First, like Stevenson and Doleac (Forthcoming), I show judges may be more lenient for white defendants than for Black defendants with the same algorithmic score. Second, I bolster Stevenson (2018)'s finding that geographically heterogeneous responses to algorithm-related policies can increase racial gaps in the justice system.

Finally, the results in this paper help inform policy issues related to algorithms and human decisions. For instance, 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 (of whether to refer someone to care) is "just one of many data elements intended to be used to select patients for clinical engagement programs." In other words, because doctors consider more than just algorithmic recommendations when making decisions, the company argued the recommendations need not impact final outcomes. However, my results demonstrate that algorithmic recommendations have independent causal effects on human decisions even when many more factors are at play. Therefore, algorithmic recommendations (like the healthcare referral recommendation) are worthy of independent study because they have demonstrable causal effects on high-stakes decisions.

Roadmap: The remainder of the paper proceeds as follows. Section 2 provides background on algorithms in different decision-making environments and highlights the role of algorithmic recommendations. Section 3 describes my empirical setting, bail decisions in Kentucky, and the natural experiment used for identification. Section 4 develops a toy model, which generates testable predictions to differentiate between dueling theories of the effects of algorithm recommendations. Section 5 describes the administrative data. Section 6 presents the main results, which demonstrate the causal effects of algorithmic recommendations using differences-in-differences and differences-in-discontinuities methods. Section 7 addresses identification concerns and bounds the estimates. Section 8 demonstrates heterogeneity in the effects of algorithmic recommendation and investigates the consequent implications for racial disparities. Section 9 concludes.

2 Background on Algorithms and Decisions

How do algorithms change decisions? The answer depends on the differences between the status quo and the new algorithm-based decision-making system. Algorithm-based decision-making systems vary widely. They include systems in which algorithm-based rules fully dictate decisions as well as systems in which humans are given some information from an algorithm with no direction on how to use the information. There is no singular algorithmic decision-making system – there is a spectrum of them.[3]

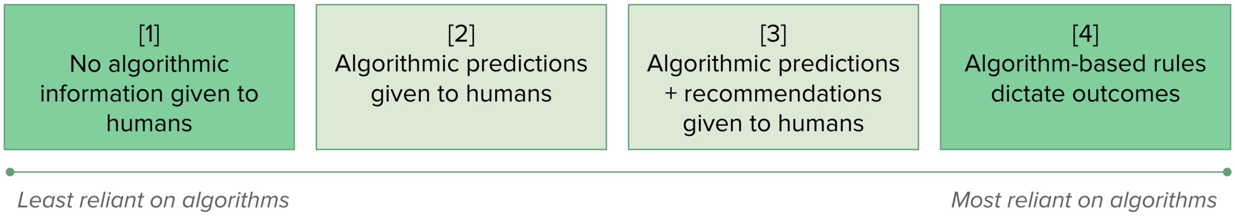
When studying the effects of algorithms, researchers often contrast a world without algorithms, where humans have complete discretion, to one with algorithms where there is no human discretion. Prior research has shown that algorithm-dictated choices can outperform human decisions in the absence of algorithms (Berk 2017; Mullainathan and Obermeyer 2022; Kleinberg, Lakkaraju, et al. 2018; Cowgill 2018a). However, this is rarely the policy-relevant comparison, because humans usually make the final decisions even when algorithms are present.

A growing literature compares outcomes in the absence of algorithms to outcomes when human decision-makers use algorithms at their discretion (Sloan, Naufal, and Caspers Forthcoming; Stevenson 2018; Stevenson and Doleac Forthcoming; Garrett and Monahan 2018; DeMichele et al. 2018; Cowgill and Tucker 2019; Davenport 2023). How algorithms are integrated into human decision-making varies across these settings. Some settings provide decision-makers with an algorithm's predictions only, while others also provide algorithmic recommendations that suggest explicit decisions. In this paper, I show that algorithmic recommendations are an important feature of algorithm-based

[3] 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" (Lum and Chowdhury 2021). This definition includes decisions strictly dictated by algorithm-based rules as well as decisions weakly informed by algorithm information. It is vague enough to include many types of algorithmic decision-making environments.

decision-making systems. They have independent effects on decisions and merit attention in policy discussions and research.

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


[1]
No algorithmic information given to humans
[2]
Algorithmic predictions given to humans
[3]
Algorithmic predictions + recommendations given to humans
[4]
Algorithm-based rules dictate outcomes
Least reliant on algorithms
Most reliant on algorithms

Notes: This figure illustrates a theoretical spectrum of algorithm-based decision-making 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 settings to make the differences across potential settings explicit. From left to right, I list four settings, from least to most reliant on algorithms. In dark green, on the ends, are the two extremes: (1) no algorithmic information given to humans and (4) algorithm-based rules dictate outcomes. In the middle, in light green, are the two intermediate settings in which humans make the final decisions, but they have some information from an algorithm. In (2), human decisions are given information on algorithmic predictions but no algorithmic recommendations. In (3), decision-makers are also given algorithmic recommendations.

Using the spectrum shown in Figure 1, I can spatially situate the previously referenced literature. 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 causally estimating the distinction between (2) and (3) to highlight the hidden effects of algorithmic recommendations implicit in previous research.

2.1 Algorithms and Bail Decisions

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. These algorithms generally predict types of risk based on individual-level and case-level characteristics. For example, the Public Safety Assessment, used in over 40 counties, calculates pretrial misconduct risk by adding up integer weights based on nine 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, which also calculates pretrial misconduct risk, has hundreds of inputs and is a black-box machine learning model (Angwin et al. 2016; Stevenson and Slobogin 2018).

In the pretrial setting, algorithms are meant to help make bail decisions. After arrest, judges decide how to set bail for the arrested person. The bail decision stipulates the conditions the arrested person must meet for release from jail. There are a few reasons why bail is an important setting for studying algorithms' effects. For one, bail decisions directly affect pretrial detention, which has downstream effects on future outcomes, such as the likelihood of conviction (Dobbie, Goldin, and Yang 2018; Cowgill 2018b). Moreover, pretrial detainees "account for two-thirds of jail inmates and 95% of the growth in the jail population over the last 20 years," and as a result, bail decisions and subsequent pretrial detention have been a substantive contributor to US mass incarceration (Stevenson and Mayson 2018). Bail is also a promising environment for study because bail decisions are made quickly (in a matter of minutes), and the legal objective is well defined (Arnold, Dobbie, and Yang 2018). The legal objective of bail is to set the lowest possible bail to ensure court appearance and public safety (American Bar Association Criminal Justice Standards Committee 2007). In this context, algorithms are designed to predict the risk of pretrial misconduct (failing to appear in court or rearrest).

While algorithms can vary greatly in how they predict misconduct, they share a common 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, allowing for the release of more people without compromising on misconduct. Prior research on risk assessments in bail settings highlights their potential in this regard. Kleinberg, Lakkaraju, et al. (2018) find that if bail decisions were delegated to a predictive algorithm, jail populations could be reduced by 42%, with no change in crime rates. 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 something 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 dimension of choice can be absent in high-stakes environments where algorithms are involved only in allocation. For instance, in the coordinated entry system, people are scored (according to housing need or 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 housing supply in that context is fixed; the algorithm cannot 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).

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, one algorithm was used to predict pretrial misconduct risk and judges could use it to inform their bail decisions. From June 2011 onward, algorithmic recommendations were also given to judges, but only for some cases. I leverage the resulting variation over time and cases to estimate the effects of the recommendations on decisions.

Bail decisions before algorithmic recommendations: 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 (the algorithmic prediction of misconduct risk). Within 24 hours of booking, the officer presented information about the defendant, their risk score, and the alleged incident to a judge. After hearing 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 and require defendants to post some money for release from jail. Judges can choose not to require money for release, which is 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 detention outright.

The Kentucky pretrial risk assessment: The algorithm used to predict misconduct during the study period was the Kentucky Pretrial Risk Assessment (KPRA). Pretrial Services created the KPRA 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 points based on "yes" or "no" answers to a series of questions. The total number of points was 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. During bail phone calls, pretrial officers told judges these 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). The factors also include 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.1 for details on risk score calculation.

Bail decisions after recommendations: In response to significant increases in the incarcerated population between 2000 and 2010, Kentucky House Bill 463 (HB463) went into effect on June 8, 2011. The law recommended release without the requirement to post money, "lenient bail," for defendants with low or moderate risk scores. The policy change did not change the calculation of the risk scores or levels; 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).

HB463 introduced recommendations for low and moderate risk cases but not for high risk cases. The risk prediction method did not change, so the same algorithmic prediction information was available to judges before and after HB463.

4 A Toy Model and Theoretical Predictions

In this toy model, I demonstrate the empirical predictions of introducing algorithm recommendations based on two distinct theories (whether they have incentive effects or prediction effects). This framework clarifies why we might or might not expect recommendations to change human decisions.

Status quo set-up: Under the status quo, judges make bail decisions using information about the case and algorithm predictions about misconduct.

The legal objective of bail is to set the lowest possible bail to ensure court appearance and public safety. To map onto the empirical setting, let judges choose whether to set money bail (or harsh bail: $b = h$) or no money bail (or lenient bail: $b = l$) for defendants. If the judge sets harsh bail, there is some probability the defendant is detained $Pr(d|b = h)$, and there is some probability the defendant is released $1 - Pr(d|b = h)$. If the defendant is detained, the judge incurs a cost $c(d|b = h)$, which is the financial cost of detaining someone in jail. If the defendant is released, they may commit misconduct with probability $Pr(m|b = h)$. If they don't commit misconduct, the judge faces no costs. If they do, the judge faces cost $c(m|b = h)$, which is the cost of misconduct, given the choice of harsh bail. In total, the judge incurs cost

$$C(b = h) = Pr(d|b = h)c(d|b = h) + (1 - Pr(d|b = h))Pr(m|b = h)c(m|b = h).$$

If the judge sets lenient bail, they incur costs based on the probabilities and costs of detention and misconduct again. However, there is no capacity for detention, so only misconduct probabilities and costs show up:

$$C(b = l) = Pr(m|b = l)c(m|b = l).$$

How are probabilities and costs determined? We assume that costs to judges are solely the reputational blowback to their decision-making. They do not face costs when the public can validate their choices as correct. When judges set harsh bail and the defendant commits misconduct, the choice is seen as correct, which means they will not face any misconduct-related consequences for being harsh. Therefore, $c(m|b = h) = 0$. Accordingly, the expression for judge costs under harsh bail simplifies to

$$C(b = h) = Pr(d|b = h)c(d|b = h).$$

On the other hand, judges face blowback for setting lenient bail for people who commit misconduct, because the choice looks like a mistake, meaning $c(m|b = l) >> 0$. Meanwhile, there is no way for anyone to assess whether harsh bail was correct when defendants are detained, because they cannot commit misconduct mechanically. Therefore, $Pr(d|b = h) \neq 0$.

How do judges predict $Pr(m|l)$? They have a vector of case information X and algorithm-based risk level information. The risk level information is a mapping from $Pr^A(m|l)$ (the algorithm's prediction of misconduct under lenient choice) to r^A. Risk levels provide relative risk information rather than absolute risk information.[4] To align with the empirical environment, we assume $r^A \in \{low, moderate, high\}$. The judge prediction is some function of observables and the algorithm's risk level: $Pr(m|l) = f(X, r^A)$.

Therefore, judges choose to set bail based on the following threshold rule:

$$b = \begin{cases} h, & \text{if } \frac{c(d|b=h)}{c(m|b=l)} < \frac{Pr(m|b=l)}{Pr(d|b=h)}, \\ l, & \text{otherwise.} \end{cases} \tag{1}$$

Adding algorithmic recommendations: Now, I complicate the status quo set-up by introducing algorithmic recommendations. Call the algorithmic recommendation R. It is based on algorithmic risk level r^A:

$$R = \begin{cases} b = l, & \text{if } r^A \in \{low, moderate\}, \\ -, & \text{otherwise.} \end{cases} \tag{2}$$

In words, lenient bail is recommended to judges if the risk level is low or moderate. There is no recommendation otherwise. How does R impact judges' decisions?

- **If all recommendations do is inform judge predictions,** then the recommendation of lenient bail ($R = b = l$) communicates to the judge that the risk level is low or moderate ($r^A \in \{low, moderate\}$). However, under the status quo, judges already know risk levels and have integrated that information into misconduct predictions (since $Pr(m|l) = f(X, r^A)$). So, if the only channel through which recommendations matter is revealing algorithmic predictions, then we would expect no change to judge decisions in this setting.
- **If recommendations instead change payoffs**, the predictions are different. If recommendations change payoffs, then $c(m|l)$ becomes $c(m|l, R)$. Assume, in line with anecdotal evidence, that it is less costly to make a mistake when that mistake is consistent with a recommendation (because there is less liability). Then, $c(m|b = l, R = b = l) < c(m|b = l)$. Similarly, making a mistake that goes against a recommendation is more costly since this is seen as "going rogue." Then, $c(m|l, R = b = h) > c(m|l)$. In this case, judges choose to set bail based on two

[4] I make this assumption to fit with common practice in the real world and in my empirical setting. If one case is "low risk" while another is "moderate risk," it is unknown what probabilities of misconduct these levels imply; however, it is clear that the "moderate risk" case has a higher predicted probability of misconduct than the "low risk" case.

distinct threshold rules (one for when the recommendation applies and one for when the recommendation does not apply):

$$b = \begin{cases} R = b = l, \begin{cases} h, & \text{if } \frac{c(d|b=h)}{c(m|b=l,R=b=l)} < \frac{Pr(m|b=l)}{Pr(d|b=h)}, \\ l, & \text{otherwise;} \end{cases} \\ R = -, \begin{cases} h, & \text{if } \frac{c(d|b=h)}{c(m|b=l)} < \frac{Pr(m|b=l)}{Pr(d|b=h)}, \\ l, & \text{otherwise.} \end{cases} \end{cases} \tag{3}$$

Because $c(m|b = l, R = b = l) < c(m|b = l)$, the cost-ratio threshold when the recommendation applies ($\frac{c(d|b=h)}{c(m|b=l,R=b=l)}$) is larger than the cost-ratio under the status quo ($\frac{c(d|b=h)}{c(m|b=l)}$). In effect, the cost threshold shifts right under the lenient recommendation, making harsh decisions less frequent and lenient decisions more frequent.

Dueling predictions: Therefore, this toy model generates dueling predictions. If recommendations only serve to communicate algorithmic predictions, then they should have no effects in my empirical setting. If they change payoffs to decision-makers, they should increase lenient bail setting when the recommendation applies.

Anecdotal evidence on liability and algorithm recommendations: Recommendations can change misconduct costs in two ways. First, lenient recommendations may make lenient decisions less risky for decision-makers. The algorithm designer who sets the recommendation – in the Kentucky case, the state legislature – provides reputational cover to the judges. If someone commits misconduct, judges can point out that the lenient decision followed 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.

Second, suppose the recommendation is detention (harsh bail) and the judge releases the defendant (the judge is lenient). In that case, the judge sticks their neck out more than they would have without a recommendation. If a defendant commits misconduct, the judge could face higher costs through increased scrutiny and political backlash (loss of a future election).[5] This theory aligns with recent events in the US. The Milwaukee DA

[5] Angelova, Dobbie, and Yang (2023) find that judges make harsher decisions after an unrelated local defendant is arrested for a violent crime. This result could also be consistent with an error cost mechanism.

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).

5 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 in the state. I limit the sample to initial bail decisions made by district judges between March 18, 2011, and June 30, 2013, because that is when Pretrial Services used a consistent algorithm to predict pretrial misconduct (the KPRA), but algorithmic recommendations changed.

The final dataset consists of around 131,000 observations. Each row is constructed at the case level and contains information about the defendant, the relevant charges, the initial bail decision, the bail judge, and the algorithm-based risk components, scores, and levels. Appendix A.2 outlines the data preparation steps that determined the structure and features of the final dataset. As usual, the administrative data do not indicate what specific information judges and pretrial officers discussed in each bail decision.

The raw administrative data do not include the underlying risk scores, but I use data on the underlying risk score components along with the publicly available weights (from Table A.1) to construct them. Therefore, as the researcher, I observe risk scores while judges observe only the more discretized risk levels. This granularity is necessary for my differences-in-discontinuities identification strategy.

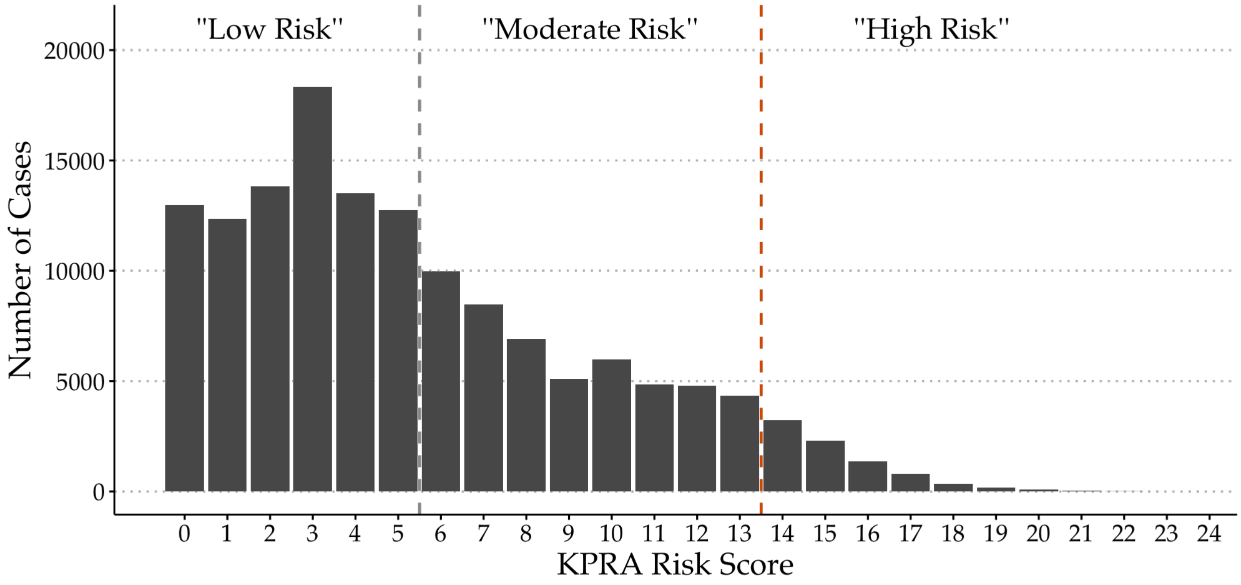
The Kentucky setting has a few features that distinguish it from other US bail settings. First, bail decisions are usually made in phone conversations between pretrial officers and judges rather than during in-person bail hearings, which are common in the US. Because judges make bail decisions over the phone, defendants are not present. Second, police

If salient misconduct increases public scrutiny of lenient judges, then lenient choices become more costly to judges, which reduces their prevalence.

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's actions. Appendix A.3 provides more background on bail setting in Kentucky.

Figure 2 demonstrates the distribution of risk scores across all cases in the administrative data. The distribution skews low risk: 90% of cases are in the low or moderate categories. Therefore, 90% of cases receive the lenient bail recommendation after the introduction of recommendations. However, only 32% of cases received lenient bail before the introduction of recommendations. 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.

Figure 2: The Risk Score Distribution



Notes: This histogram demonstrates the number of cases across the full risk score distribution. The dashed lines indicate the cut-offs between risk levels. The orange line is the threshold between "moderate" and "high" risk, and the gray line is the threshold between "low" and "moderate" risk. Scores of 0-5 are low risk, scores of 6-13 are moderate risk, and scores of 14 and above are high risk.

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. Recall that many different decision thresholds are consistent with the same underlying risk rankings. In this way, I can conceptualize algorithmic recommendations as a form of what Cowgill and Stevenson (2020) call "algorithmic social engineering" – recommendations are derived from predictions, but manipulated to reflect some algorithmic designer's perspective. The threshold of 14 suggested that the state wanted judges to set lenient bail more frequently than 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 have very different policy implications. While many researchers focus on how algorithms can change decisions in a "locally optimal" (or an "allocative") sense (Kleinberg, Lakkaraju, et al. 2018), the threshold choice hints at how algorithmic recommendations may have their independent causal effects.

6 Algorithmic Recommendations Change Judge Decisions

Throughout my study period, risk levels (low, moderate, high) were available to judges when they were setting bail. To study the effects of recommendations, I leverage a change that occurred in June 2011. Risk level calculation did not change, but judges were given explicit recommendations on setting bail. The new recommendation was to set lenient bail (no money bail) for cases with low or moderate risk scores.

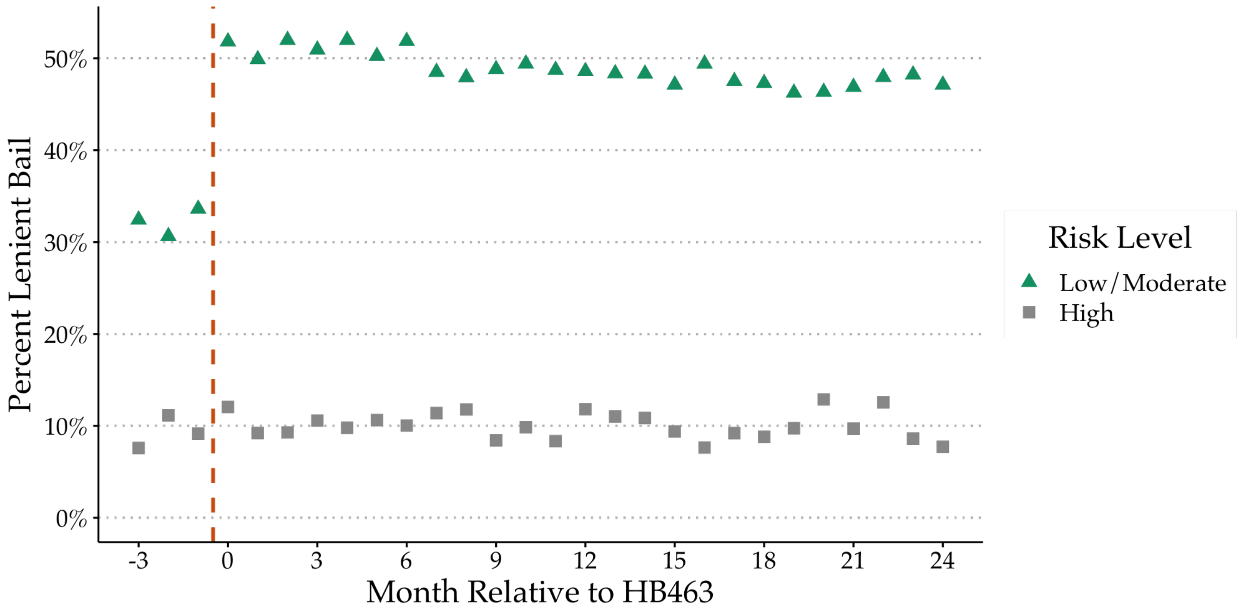
Does the new algorithmic recommendation change human decisions, even though the algorithm's predictions stay the same? I answer this question using two empirical methods: differences-in-differences and differences-in-discontinuities.

6.1 Differences-in-Differences Results

In my differences-in-differences framework, high risk cases are the control group because they experience no change in recommendations. In contrast, low and moderate risk cases are the treatment group because they experience a change in recommendations. Figure 3 illustrates the rate of lenient bail for low or moderate risk cases and high risk cases over time.[6] Once recommendations go into effect, there is a stark increase in lenient bail for low/moderate cases of about 15-20 percentage points. There is no similar increase for the high risk group. The underlying assumption of using a differences-in-differences approach is parallel pre-trends. The raw visual evidence in Figure 3 provides promising evidence for this assumption.

[6] Note that there are only a few pre-policy time periods because the method of calculating the risk score changed in March 2011. In order to keep the meaning of risk scores and risk levels consistent, I drop data from before March 2011.

Figure 3: Lenient Bail Rates by Risk Level over Time



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

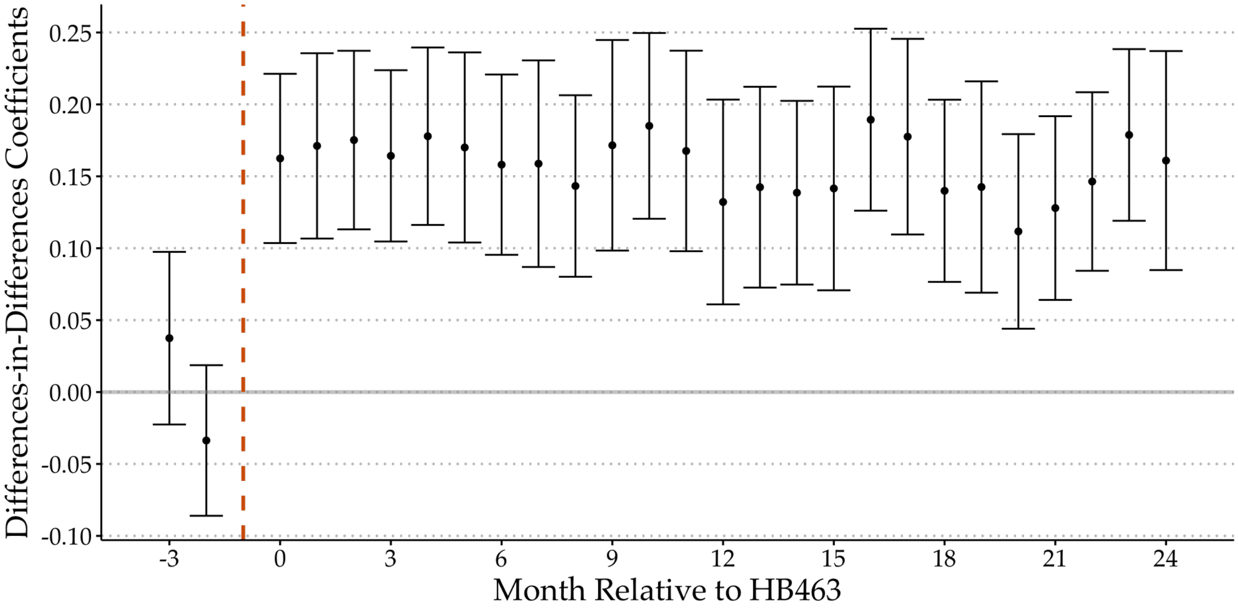
To formally estimate causal effects and test for pre-trends, I estimate a standard specification of the form

$$lenient_{itj} = \sum_{m \neq -1} [\beta_m \times I(score_i < 14)] + X_{itj} + \epsilon_{itj}, \tag{4}$$

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, meaning the risk level is low or moderate (rather than high). Distinct coefficients are estimated for each month m relative to HB463 adoption, and $m = -1$ is the omitted group. I include a 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. I cluster standard errors by judge.

Figure 4 shows the dynamic differences-in-differences coefficients by plotting the values of β_m. Before the recommendation introduction, the coefficients are close to zero and do not demonstrate evidence of pre-trends. The results are not sensitive to the choice of control variables. Figure A.1 shows that results with zero detailed controls are nearly identical to those with controls based on all observed case variables.

Figure 4: Dynamic Differences-in-Differences Estimates



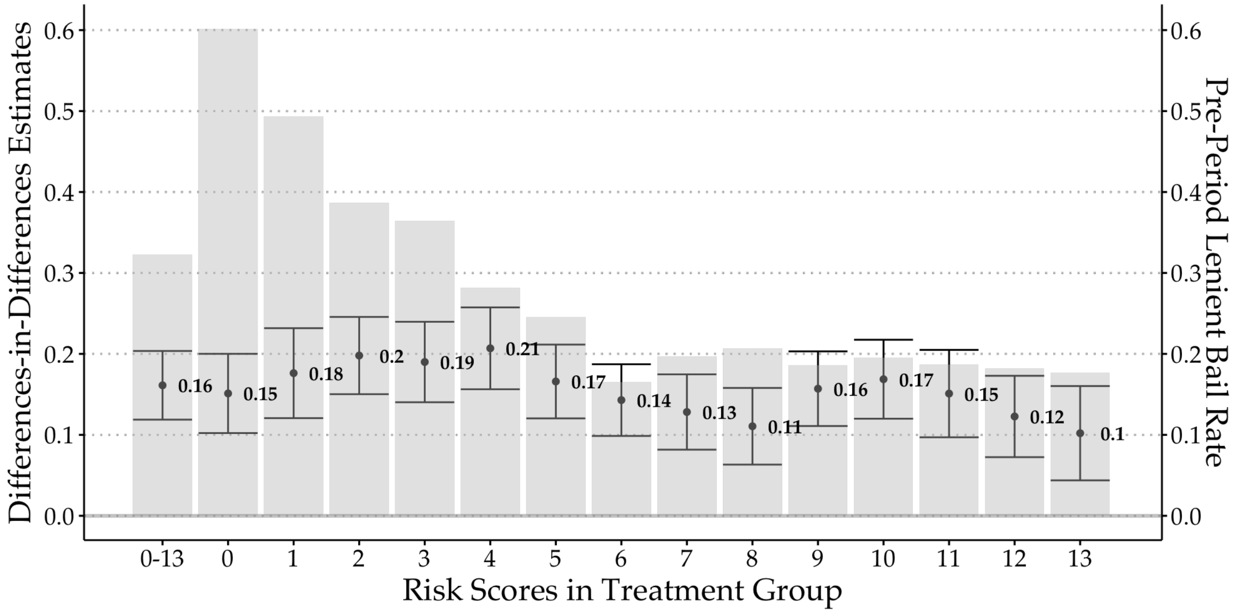
Notes: This figure shows the difference-in-differences coefficients for months relative to recommendation introduction. The orange dashed line denotes the omitted period of the month before recommendation introduction.

To obtain a summary coefficient, I estimate pooled differences-in-differences coefficients and present these results across specifications in Table A.3. Pooling time periods, I observe that the algorithmic recommendation increased 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%. These economically meaningful results are consistent with the theory that algorithmic recommendations change the costs of errors to decision-makers.

How do effects vary across the risk score distribution? So far, estimated effects apply to the entire low and moderate risk score distribution. If the recommendations change the cost of errors, we should see results across the whole risk score distribution. To test this, I estimate pooled differences-in-differences coefficients for each risk score in the low/moderate distribution – that is, scores between 0 and 13. In the raw data, each risk score group between 0 and 13 experienced a discontinuous increase in lenient bail at the time of HB463 (see Figure A.2). Figure 5 shows this result holds when estimating the differences-in-differences coefficients as well. The figure shows the pooled differences-in-differences coefficients and the baseline lenient bail rates (in shaded gray bars). There are statistically significant effects across the entire distribution, and the estimates range from 10 to 20 percentage points.

Even though the point estimates are similar in magnitude across the distribution, the relative effects are larger near the moderate-high risk cut-off because they have lower lenient bail baseline rates. For illustration, the coefficient for cases with scores of 0 is 14.8 percentage points, a 25% relative increase off the 60% baseline rate. In comparison, the coefficient for cases with scores of 13 is 10.1 percentage points, a 60% relative increase off the 18% baseline rate. The estimated coefficients across the distribution are similar regardless of specification and control choices, as demonstrated by Figure A.3.

Figure 5: Pooled Differences-in-Differences Estimates across Risk Score Bandwidths



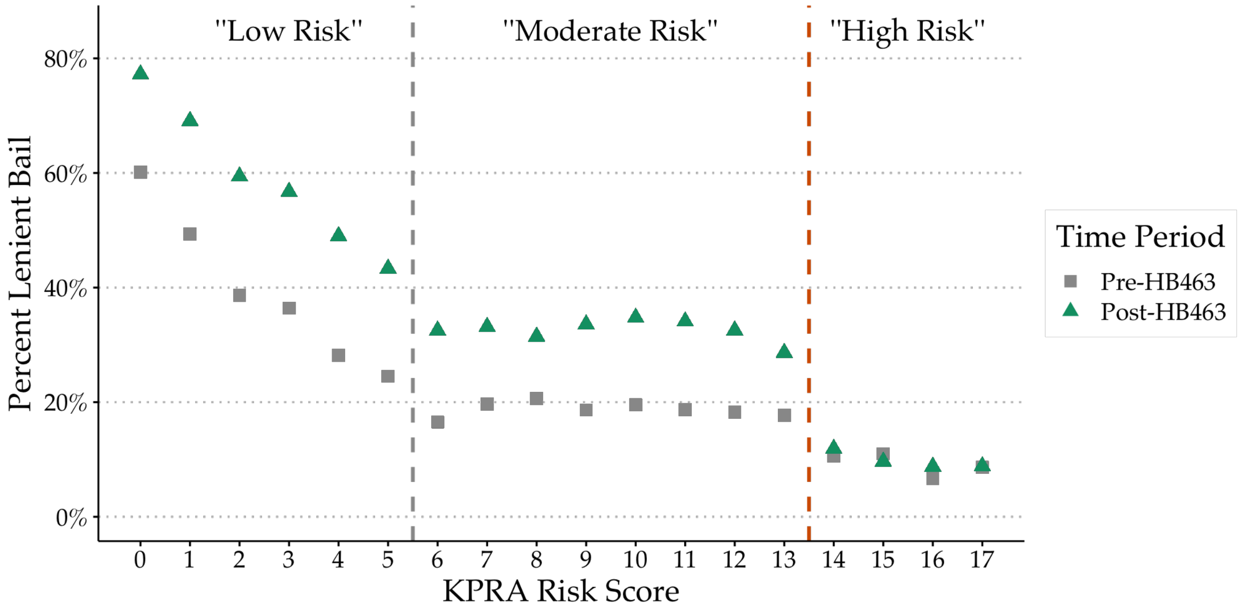
Notes: This figure shows the pooled difference-in-differences coefficients across different treatment groups based on risk scores. The control group consists of cases with high risk scores, and the treated group consists of cases with low or moderate risk scores (varying from 0 to 13). Specifications are estimated separately for all risk score treatment groups. The specification includes controls for day of week, month-year, exact risk score, top charge level/class, defendant demographics (race, gender), judge, county, and all risk score components listed in Table A.1 except for verified address and support. The black error bars show the 95% confidence interval for each differences-in-differences coefficient. The light-shaded gray bars show the baseline rate of lenient bail for that risk score group in the pre-period, which allows for relative interpretation of effect sizes.

6.2 Differences-in-Discontinuities Results

I can also estimate causal effects using a different identification strategy, focusing on marginal cases near the recommendation threshold. After June 2011, the lenient bail recommendation applied only to cases with risk scores below 14. Therefore, cases with similar risk scores received different recommendation treatments based on which side of the critical threshold they were on.

If the lenient bail recommendation were the only factor that changed discontinuously over the threshold, a simple regression discontinuity using the post-period data would identify the lenient recommendation effect. However, other relevant factors changed discontinuously at that threshold as well. Conveniently for identification, these confounding factors were also present in the pre-period. Therefore, I can leverage pre-period data at the same discontinuity to difference out confounding factors with a differences-in-discontinuities approach. This approach then allows me to isolate the effect of interest – the effect of the lenient bail recommendation – for cases near the threshold.

Figure 6: 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 to the left of the orange line received a lenient bail recommendation, but those with scores to the right did not. The gray rectangles show the rates before HB463, while the green triangles show those after HB463.

Figure 6 demonstrates the differences-in-discontinuities approach visually. The figure shows the percentage of cases that received lenient bail based on cases' risk scores and the time period. Points on the left represent cases with the lowest risk scores, while points on the right represent cases with the highest risk scores.[7] I show lenient bail rates across

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

the score distribution in the pre-period (before the introduction of recommendations) and post-period (after the introduction of recommendations).

There were no changes in recommendations for high risk cases (points to the right of the orange dashed line) across time periods, but there were changes for low or moderate risk cases (points to the left of the orange dashed line). For cases that did not experience a change in recommendation, lenient bail rates are nearly identical in the pre- and post-periods. However, for cases that did experience a change in recommendations, lenient bail rates are 10-20 percentage points higher in the post-period.[8] This raw visual evidence is consistent with lenient recommendations having a causal effect on lenient bail rates because rates increase discontinuously where the recommendation kicks in at the critical threshold (the orange dashed line) in the post-period, and the same increase is not present in the pre-period (before the introduction of recommendations).

To formally estimate the effect of the recommendation at the margin, I use a differences-in-discontinuities approach pioneered by Grembi, Nannicini, and Troiano (2016). I estimate regression discontinuity coefficients before and after HB463 and take the difference to isolate the effect of the lenient recommendation. Using data from the post-period, I estimate the effect of crossing the moderate-high threshold using nonparametric methods following Calonico, Cattaneo, and Titiunik (2014) and Calonico, Cattaneo, and Farrell (2020) for optimal bandwidth selection and bias-corrected inference. My preferred estimate yields a 13.7 percentage point effect. This method uses a triangular kernel and the optimal bandwidth based on Calonico, Cattaneo, and Farrell (2020), but particular estimation choices do not erode the effect (see Figure A.4). Since cases with scores of 14 receive lenient bail only 16% of the time, crossing the threshold in the post-period makes the case almost twice as likely to receive lenient bail (even though the underlying risk prediction is very similar).

If the only factor that changed across the moderate-high threshold was the lenient bail recommendation, the regression discontinuity estimate would be equivalent to the recommendation effect of interest. However, two other factors change discontinuously across

[8] As an aside, Figure 6 also demonstrates a clear downward trend in lenient bail for the low risk scores as they get higher (from 0 to 5). However, moderate risk scores receive similar lenience across the score range (from 6 to 13). One likely explanation is that even though judges do not receive the underlying risk scores, it is obvious to them which cases are the lowest risk. In cases with the lowest risk (scores near 0), the person arrested has little or no criminal history background, which is quickly evident on their bail phone call with pretrial officers. Meanwhile, when an arrested person has a handful of risk factors, they necessitate a more extended conversation, making judges less likely to be able to tell the difference between someone who has a low score in the moderate group (e.g., a 6) and someone who has a high score in the moderate group (e.g., a 13).

the threshold. First, the risk level given to judges for the case changes. Cases scored as 14 receive a *high risk* label and no recommendation, but cases scored as 13 receive a *moderate risk* level and a lenient recommendation. Second, Figure A.5 shows that while most characteristics do not display a sharp discontinuity around the critical threshold in the post-period, one exception is prior felony convictions. Defendants with cases that are marginally high risk are discontinuously more likely to have a prior felony conviction than defendants with cases that are marginally moderate risk. Therefore, the estimated 13.7 percentage point effect is some combination of the effect of changing risk levels, the effect of a prior felony conviction, *and* the effect of the recommendation.

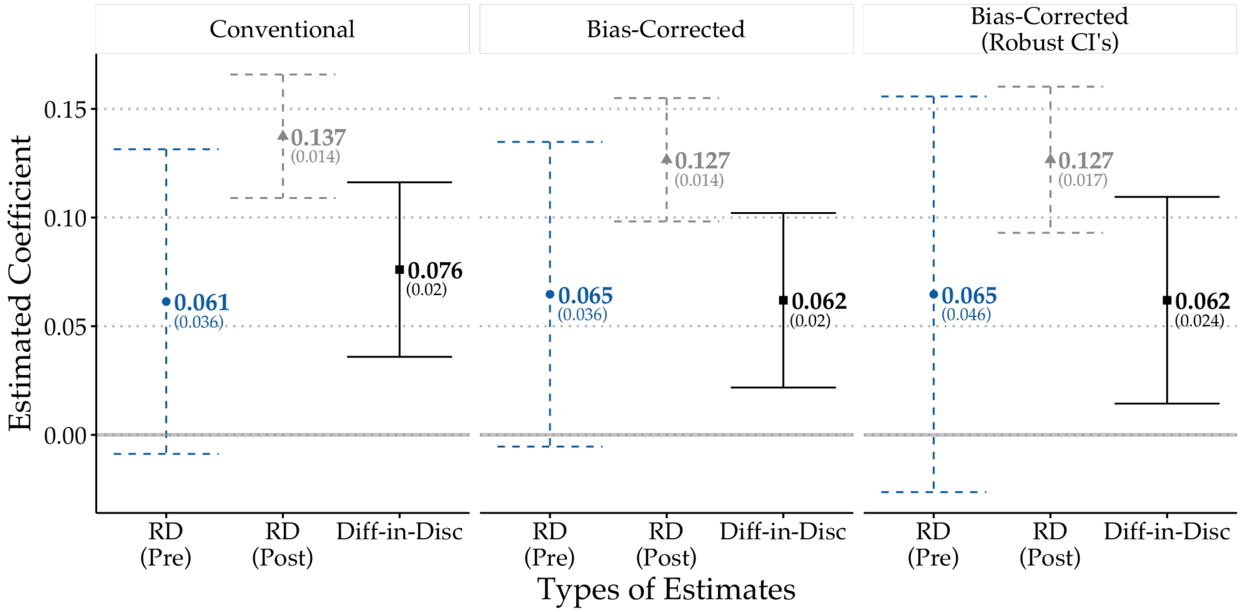
I can disentangle the recommendation effect from the other two components by leveraging the fact that I can observe bail decisions around the same discontinuity in the pre-period. In a regression discontinuity design, a central assumption is that nothing but the treatment (the presence of lenient recommendations, in my case) changes discontinuously at the threshold. My differences-in-discontinuities approach weakens this assumption, allowing for discontinuities at the threshold (confounders) as long as those same discontinuities are present in both time periods (Grembi, Nannicini, and Troiano 2016). Risk levels were present in the pre-period, and the movement in covariates across the risk score distribution was similar. In particular, Figure A.6 shows that the discontinuous uptick in the likelihood of prior felony conviction is nearly identical in the pre-period and the post-period, supporting the validity of differences-in-discontinuities assumptions in this setting.

The regression discontinuity estimate in the pre-period estimates the risk levels effect (the effect of switching from high to moderate) combined with the prior felony conviction effect. Again, I use nonparametric methods following Calonico, Cattaneo, and Titiunik (2014) and Calonico, Cattaneo, and Farrell (2020) for optimal bandwidth selection and bias-corrected inference. My preferred estimate in the pre-period aligns with the methods for my preferred estimate in the post-period: I use a triangular kernel and the optimal bandwidth based on Calonico, Cattaneo, and Farrell (2020), and again the effect is similar across different estimation choices (see Figure A.7). My preferred estimate in the pre-period is a 6.1 percentage point effect. This estimate is less than half the magnitude of the regression discontinuity in the post-period (13.7 percentage points). The pre-period estimate is meaningfully noisier than the post-period estimate because of the asymmetric nature of the data. There are many more months available for estimation in the post-period.[9]

[9] My pre-period data span only three months. There was a change to the scoring system three months

Finally, I can take the difference between the pre-period regression discontinuity estimate and the post-period regression discontinuity estimate to isolate the effect of the lenient recommendation. The preferred pre- and post-period regression discontinuity results yield a differences-in-discontinuities estimate of 7.6 percentage points. This estimate is statistically significant and economically meaningful: the lenient recommendation caused an almost 50% increase in lenient bail at the margin (an increase of 7.6 percentage points off a baseline of 16% for cases with scores of 14).

Figure 7: Regression Discontinuity and Differences-in-Discontinuities Estimates at the Moderate-High Threshold



Notes: The blue points and dotted lines show the coefficients and 95 percent confidence interval for pre-period regression discontinuity estimates at the moderate-high threshold. The gray points and dotted lines show the coefficients and 95 percent confidence intervals for post-period regression discontinuity estimates at the moderate-high threshold. The black dots and lines show the coefficients and 95 percent confidence intervals for differences-in-discontinuities estimates at the moderate-high threshold.

Figure 7 summarizes the pre-period regression discontinuity, post-period regression discontinuity, and differences-in-differences results across different estimation methods (conventional, bias-corrected, and bias-corrected with robust confidence intervals), following Calonico, Cattaneo, and Farrell (2020). Results are similar across these estimation options. Overall, the differences-in-discontinuities results are consistent with the estimated differences-in-differences results across the risk score distribution shown in Figure 5. The results further illustrate that algorithmic recommendations have independent effects,

before the policy change. To keep the scores' and levels' definitions fixed for valid identification, I am necessarily limited to those three months of pre-period data.

which is consistent with the theory that recommendations change the costs of errors to human decision-makers.[10]

7 Addressing Possible Confounds

Both the differences-in-differences and differences-in-discontinuities strategies leverage the fact that recommendations were introduced for some cases (low and moderate risk cases) but not others (high risk cases). To correctly attribute the estimated effects to recommendations, it must be the case that at the time of HB463, nothing else differentially impacted low and moderate risk cases relative to high risk cases.

In this vein, there is a potential identification concern due to institutional details related to HB463. While the calculation of risk levels was the same before and after HB463, and the risk levels were available before and after HB463, the policy change made it *mandatory* for judges to consider them. Therefore, some judges and pretrial officers may not have discussed risk levels on the bail calls before HB463.[11] In that case, HB463 changed the presence of algorithmic predictions *and* recommendations, which complicates how we interpret the previous differences-in-differences and differences-in-discontinuities results.

7.1 Differences-in-Discontinuities Context

In the differences-in-discontinuities approach, I estimate the differences-in-discontinuity coefficient at the moderate-high threshold to recover the effect of algorithmic recommendations, which I'll call R. Intuitively, I leverage the fact that the post-period regression discontinuity at this threshold is the sum of the recommendation effect (R), the levels effect at the threshold (L_{mh}, the effect of being labeled moderate instead of high risk for marginal cases), and the effect of increased prior felony conviction (F).[12] Meanwhile, the pre-period regression discontinuity at the threshold is the sum of the levels effect at the threshold (L_{mh})

[10] My results are also consistent with previous research that showed that discontinuous changes in algorithm risk labels have causal impacts on criminal proceedings (Cowgill 2018b). Both sets of results show that *how* algorithms are communicated matters for human decisions.

[11] Because the administrative data do not indicate which information judges discussed in bail decisions, I cannot directly test this possibility using tabulations in the data.

[12] Recommendation and level changes are sharp discontinuities over the moderate-high threshold. But, the prior felony conviction change is a fuzzy discontinuity because the share of cases with prior felony convictions increases from around 40% to 60% when crossing the moderate-high threshold. For notational simplicity, I refer to F as the effect of increased prior felony conviction, but it could also be denoted $0.2F'$, where F' is the sharp effect of moving from 0% to 100% of cases with prior felony convictions.

and the effect of increased prior felony conviction (F). Therefore, the difference between the two (the differences-in-discontinuities coefficient) leaves the desired recommendation effect (R).

If $\omega \in [0,1]$ is the share of judges who did not consider risk levels before HB463, then the interpretation of some of these estimates changes. The post-period regression discontinuity still recovers the desired recommendation effect plus the levels and increased prior felony effects, $R + L_{mh} + F$. However, the pre-period regression discontinuity coefficient recovers the increased prior felony effect plus a *diluted* version of the levels effect because only some judges considered levels in the pre-period. Assuming the ω share of judges is similar in their risk level responses to the remaining $1 - \omega$ share of judges, then the pre-period regression discontinuity estimates $\omega L_{mh} + F$ instead of $L_{mh} + F$. Accordingly, the difference-in-discontinuity approach estimates the recommendation effect plus an additional levels effect, $R + (1 - \omega)L_{mh}$ (instead of just the desired recommendation effect, R).

Since $\omega \geq 0$, the differences-in-discontinuities estimate is necessarily an upper bound for the recommendation effect. If ω is close to 1 (almost all judges considered risk levels before), then the extra term goes to 0, and the identification strategy recovers the recommendation effect well. If ω is close to 0 (almost no judges considered risk levels before), then the previous strategy does not recover recommendation effects unless L_{mh} is near 0.

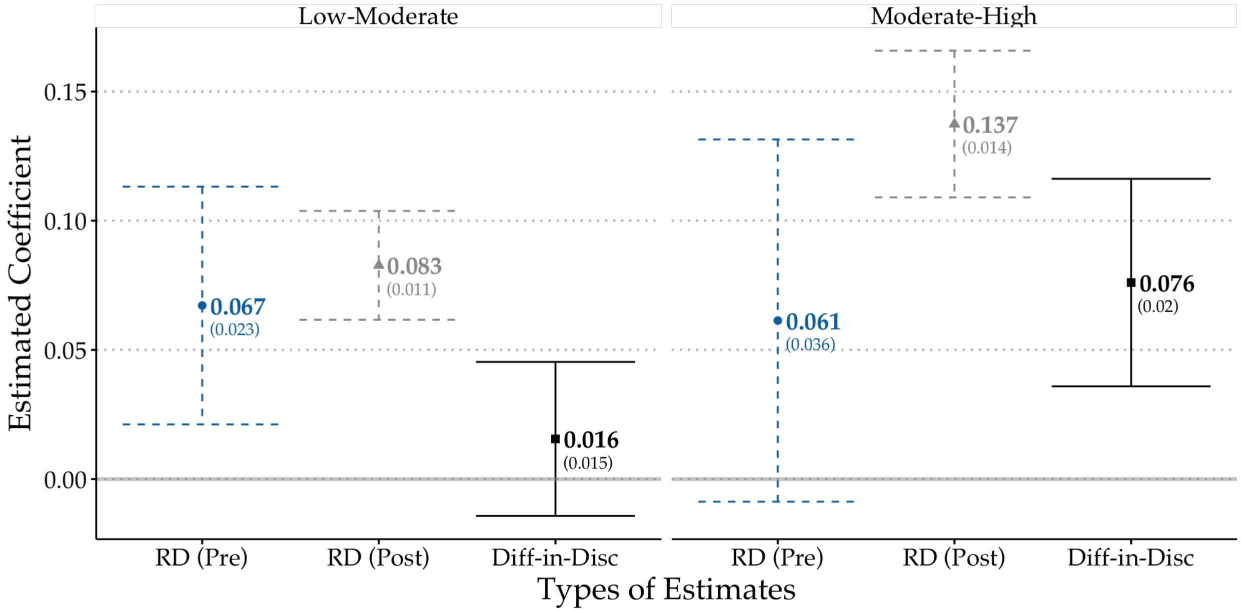
I can leverage another discontinuity in the risk score distribution to investigate the magnitude of ω. Cases also experience a discontinuous change in their risk level around the low-moderate discontinuity. Importantly, there is no change in the presence of recommendations over that discontinuity. Recommendations are either present for both groups, as in the post-period, or not, as in the pre-period. In the post-period, the regression discontinuity at this threshold recovers L_{lm}, the effect of being labeled low risk rather than moderate risk for marginal cases. Assuming that risk score consultation is similar around the low/moderate and moderate/high thresholds before HB463, then the pre-period regression discontinuity recovers ωL_{lm} and the differences-in-discontinuities estimate equals $(1 - \omega)L_{lm}$.[13]

Intuitively, if the differences-in-discontinuities estimate for the low-moderate threshold is near 0, then ω is near 1 and confounding is limited. To test this, I directly estimate the pre-period RD, post-period RD, and differences-in-discontinuities at the low/moderate

[13] This assumption requires an assumption about the homogeneity of ω at both points in the risk score distribution. However, it does not require any assumptions about how the level effect around the low/moderate discontinuity (L_{lm}) relates to the level effect around the moderate/high discontinuity (L_{mh}).

threshold. Figure 8 shows the results. The differences-in-discontinuities coefficient is 1.6 percentage points in magnitude and is not statistically significant. Therefore, this potential source of confounding is limited.

Figure 8: Regression Discontinuity and Differences-in-Discontinuities Estimates at Critical Thresholds



Notes: The blue dots and dotted lines show the point estimates and 95 percent confidence intervals for regression discontinuities using pre-period data. The gray dots and dotted lines show the 95 percent confidence intervals for regression discontinuities using post-period data. The black dots and lines show the point estimates and 95 percent confidence intervals for the differences-in-discontinuities results. I show estimates for both critical thresholds in the data: the low-moderate and the moderate-high thresholds.

Even though the results at the low-moderate threshold are not large, I can still use them to provide more conservative bounds on my original estimates of the recommendation effect. The pre-period regression discontinuity at the low-moderate threshold recovers ωL_{lm}, while the post-period regression discontinuity recovers the undiluted L_{lm}. Therefore, it is also the case that $RD_{lm}^{pre} = \omega RD_{lm}^{post}$. Plugging in the estimates from Figure 8 yields $0.067 = \omega 0.083$, which implies $\omega = 0.81$. Therefore, using the empirical estimates to solve this system of equations implies that risk levels were consulted in about 81% of cases.

I can use this estimated ω with the estimates at the moderate-high threshold to adjust my estimated effects. Table 1 compares the original estimates (assuming no confounding or $\omega = 1$) to adjusted estimates with $\omega = 0.81$. The effect of risk levels and increased prior felony conviction is slightly higher (7.5 percentage points instead of 6.1 percentage points), and the effect of recommendations is slightly lower (6.2 percentage points instead of 7.6 percentage points) with this adjustment. However, the adjusted results are similar in magnitude to the original results. Therefore, the recommendation effects survive this challenge to identification in the differences-in-discontinuities context.

Table 1: Comparing Estimates with and without Estimated Confounding

Parameter	Original Estimate ($\omega = 1$)	Adjusted Estimate ($\omega = 0.81$)
$R + L_{mh} + F$	13.7	13.7
$L_{mh} + F$	6.1	7.5
R	7.6	6.2

Notes: This table compares the original estimates from the regression discontinuities and differences-in-discontinuities approaches at the moderate-high threshold with estimates adjusted with the estimated ω parameter. Data-driven estimation implies that $\omega = 0.81$. The first row is the sum of the recommendation effect (R), the levels effect at the threshold (L_{mh}, the effect of being labeled moderate instead of high risk for marginal cases), and the effect of increased prior felony conviction (F). The second row is the sum of the levels effect at the threshold (L_{mh}) and the effect of increased prior felony conviction (F). The final row is the desired algorithmic recommendation effect (R).

7.2 Differences-in-Differences Context

I can also address potential confounding in the differences-in-differences context. If $\omega \in [0, 1]$ is the share of judges who did not consider risk levels before HB463, then the differences-in-differences approach also picks up the effect of newly using risk levels for $(1 - \omega)$ of the cases. So, the pooled differences-in-difference coefficient (β^{DD}) is then a weighted average between our desired recommendation effect (R') and an effect of levels (the effect the judge hearing the case risk level as opposed to not, L).[14] Specifically, $\beta^{DD} = R' + (1 - \omega)L$.

Since $\omega \geq 0$, β^{DD} is necessarily an upper bound for the recommendation effect. If ω is close to 1 (almost all judges considered risk levels before), then $\beta^{DD} \approx R'$, and the straightforward differences-in-differences strategy recovers the recommendation effect well. If ω is close to 0 (almost no judges considered risk levels before), then $\beta^{DD} \approx R' + L$, and the previous strategy does not recover recommendation effects *unless* we think $L \approx 0$.[15]

[14] Note that the recommendation effect here is denoted as R' because this effect may differ from the recommendation from the prior section, R. The two effects may differ because they use different data samples for identification.

[15] While I have an estimate of ω from the prior differences-in-discontinuities section, I can recover R' only with an estimate of L, which I cannot estimate. I can estimate the effect of switching levels at the margin (L_{lm} and L_{mh}) with regression discontinuities. However, it is impossible to use the available observational variation to estimate the effect of hearing any risk level instead of not (L).

Therefore, I test whether recommendation effects still matter in a subset of cases where the expected effect of consulting risk levels is small ($L \approx 0$). Specifically, think of cases where the risk level does not provide new prediction information to the judges because they are obviously low risk. A prime example is cases that are due to misdemeanor arrests, an attribute that is very salient to judges, and cases that have risk scores of 0, meaning that the person affiliated with the case has 0 risk factors (0 failures to appear, 0 pending cases, 0 convictions, etc.). In other words, these are cases associated with low-level offenses where the defendant has no criminal history. The bail phone call is short in these cases, and it is clear to judges that the relevant defendant is low risk. Since $L \approx 0$ intuitively in this case, then β^{DD} is a valid approximation for the recommendation effect for this group.

Misdemeanors with 0 risk factors are 7% of cases in the data. Figure A.8 illustrates the rate of lenient bail for these cases in contrast to the rate of lenient bail for the high risk cases. Intuitively, judges know these cases are low risk because there are no risk factors to discuss on the bail call and the offense itself is a misdemeanor. Even if some judges had not consulted risk levels before the policy change, the new "low risk" label should not introduce new prediction information to the judge. Regardless, we see an increase of 10-15 percentage points around the policy date.

Using this set of obviously low risk cases, I estimate dynamic and pooled differences-in-differences coefficients following the methodology in Section 6.1. Figure A.9 shows that the coefficients increase after the policy change in a way that diverges from any existing pre-trends. The results are not sensitive to the choice of control variables. Figure A.10 shows that results with no detailed controls are nearly identical to those with controls based on all observed case variables.

Table A.4 shows the pooled differences-in-differences results across different sets of controls. The estimated coefficients are similar in percentage point terms to those estimated for the entire sample in Table A.3. However, they are relatively smaller because the baseline rates are higher for the lowest risk sample. The 14-15 percentage point increase in lenient bail is a 22% increase relative to the baseline of 66%. These results demonstrate recommendation effects survive in a sub-sample of cases where potential confounding should be minimal. Therefore, the recommendation effects in the differences-in-differences context are robust to concerns about confounding variation.

8 Heterogeneous Effects by Defendant Race

One important reason why algorithm-based tools (like algorithmic recommendations) have been adopted in criminal legal settings is that they have the potential to standardize decisions across settings. When different treatment conditional on observables contributes to racial disparities, mandated standardization necessarily reduces these gaps. For instance, Feigenberg and Miller (2021) show that for similar-looking cases and defendants, racially heterogeneous jurisdictions in the US South are more punitive than racially homogeneous jurisdictions. If algorithmic recommendations meant all jurisdictions became equally severe, then racial gaps in criminal justice outcomes would be reduced.

However, decision-makers do not always adhere to algorithmic recommendations. In fact, decision-makers themselves may vary in their response to recommendations. If their responses are correlated with defendants' races, then algorithmic recommendations could worsen rather than alleviate racial gaps. Therefore, the effects of algorithmic recommendations on racial disparities are ambiguous. Do algorithmic recommendations differently impact white and Black defendants who face the same recommendation?

To answer this question, I investigate if there is racial heterogeneity in my causal estimates of algorithmic recommendation effects. Specifically, I focus on Black and white defendants in Kentucky since these two racial groups make up 99.8% of the cases in my data.[16] About 84% of cases feature white defendants, while 16% feature Black defendants. Black people are overrepresented in the data relative to the full population in Kentucky, which was approximately 8% Black people in 2010 (Price 2011).

Black defendants have a higher distribution of risk scores and higher consequent risk levels than white defendants. Figures A.11 and A.12 compare the risk level and risk score distributions, respectively, for white and Black defendants. The average risk score for Black defendants is 6.4, while the average score for white defendants is 5.3. Moreover, white defendants are 10 percentage points more likely to be classified as low risk. The different scores and levels are due to differences across groups in score inputs. Table A.5 shows the different risk score component averages by racial group. The most significant contributing factors to the different scores are higher failure to appear rates, prior felony rates, and prior violent conviction rates (all of which negatively impact risk score) for Black defendants relative to white defendants.

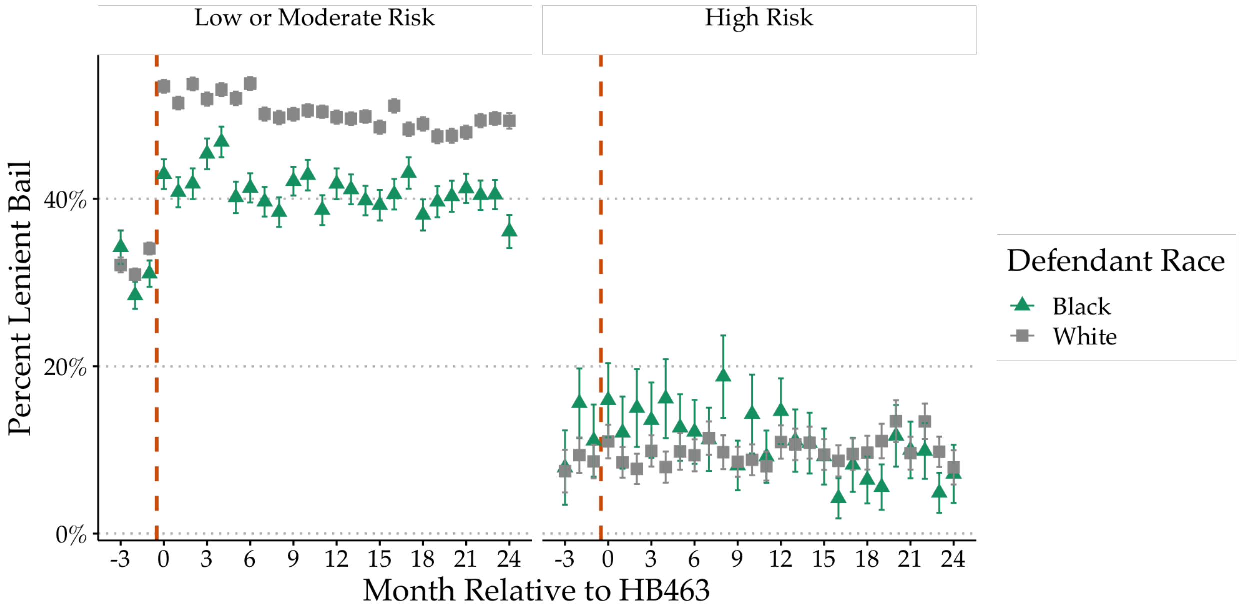
[16] Regarding ethnicity, only 1.5% of defendants are recorded as Hispanic. Moreover, in 26% of cases, ethnicity is listed as unknown. Owing to the small sample of people identified as Hispanic and the significant amount of missing data, I will not discuss results by ethnicity.

If the lenient recommendation were the sole determinant of bail after recommendation introduction, Black people would have been 3.3 percentage points less likely to receive lenient bail than white people. This gap would have been solely generated by the different risk level distributions across groups. However, in practice, Black people were 9.3 percentage points less likely to receive lenient bail than white people. The racial gap observed in the data is significantly larger than the gap generated with algorithmic recommendations and no human discretion, which suggests that deviation rates from lenient recommendations may vary by defendant race. The following section directly tests that possibility.

8.1 Do Recommendation Effects Vary by Defendant Race?

In the simple differences-in-differences framework, cases with low and moderate risk scores are the treated group (since they experience a new lenient recommendation), while cases with high risk scores are the control group (since they experience no change). To investigate racial heterogeneity, I also split the cases according to the defendant's race.

Figure 9: 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. The orange dotted line shows when HB463 went into effect. I illustrate cases with Black defendants as green triangles, and I illustrate cases with white defendants as gray squares. Cases with risk scores under 14 (low or moderate risk level) are in the left panel, and cases with risk scores of 14 or over (high risk level) are in the right panel.

Figure 9 shows the rates of lenient bail for (a) low or moderate risk cases over time by defendant race and (b) high risk cases over time by defendant race. For the low or moderate risk cases, Black and white defendants received similar rates of lenient bail before HB463. However, after HB463, white defendants with low or moderate risk scores were 10 percentage points more likely to receive lenient bail than Black defendants with low or moderate risk scores, even though all defendants with these risk scores received the lenient recommendation. Meanwhile, there was no consistent racial gap in outcomes for the high risk score group either before or after HB463. The raw visual evidence suggests that the lenient recommendation led to more lenience for white than Black defendants.

I can further validate this result by estimating differences-in-differences coefficients, following Specification 4, separately for each racial group. Figure A.13 illustrates the dynamic differences-in-differences coefficients separately for each group. The estimates for white defendants are consistently higher than those for Black defendants and are similar regardless of the exact sets of controls.

To estimate useful summary coefficients, I estimate pooled differences-in-differences coefficients for both racial groups. Columns 1 and 2 in Table 2 show the estimated differences-in-differences coefficients for the sample of only white defendants and only Black defendants, respectively. The algorithmic recommendation increased lenient bail by 17.5 percentage points for white defendants and 9.4 percentage points for Black defendants. Therefore, the recommendations had approximately half the effect on judge leniency for Black defendants as they did for white defendants.

I can demonstrate the same result by estimating a triple differences specification for the entire sample. The triple differences specification takes the form

$$\begin{aligned} lenient_{itj} = \beta_1[I(score_i < 14) \times Post_t] + \beta_2[I(score_i < 14) \times Black_i] + \\ \beta_3[Post_t \times Black_i] + \beta_4[I(score_i < 14) \times Post_t \times Black_i] + X_{itj} + \epsilon_{itj}, \end{aligned} \tag{5}$$

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 coefficient demonstrates the heterogeneity of the effect of recommendations by defendant race. I include the same vector of controls X_{itj} as in Specification 4, and I cluster standard errors by judge. Column 3 of Table 2 shows the triple differences coefficient when I estimate the specification for the full sample of defendants. The results are nearly identical to those shown in columns 1 and 2: the difference between the effects is around 8.0 percentage points, which is statistically significant at the 5% level.

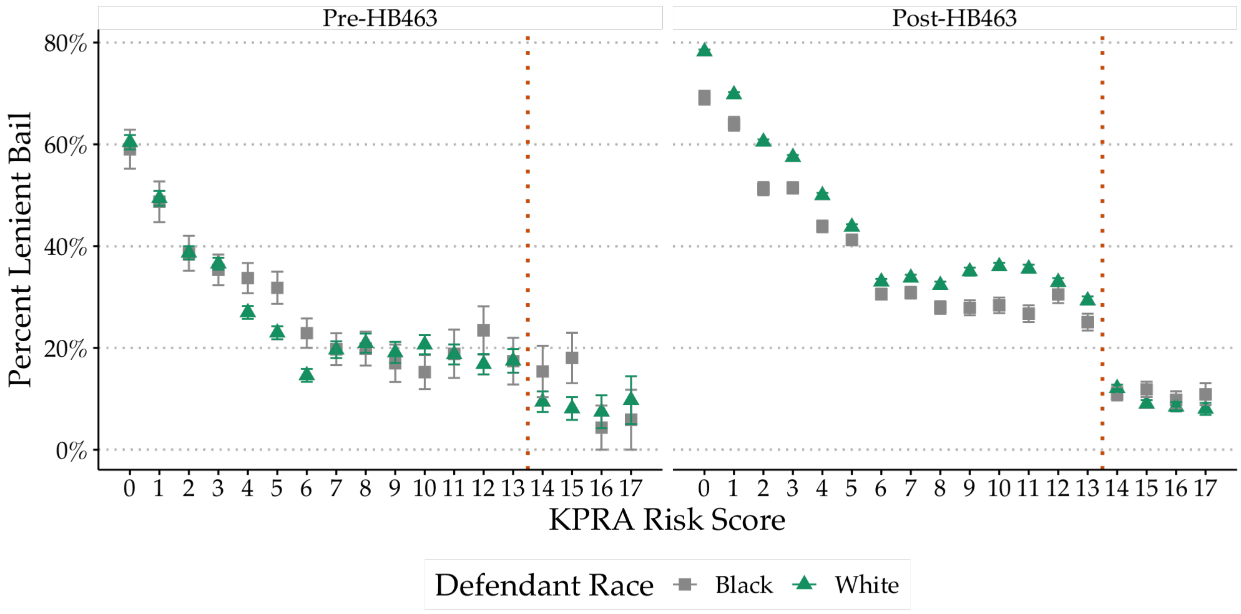
Table 2: Differences-in-Differences Results by Defendant Race

	Dependent variable: I(lenient bail)		
	DD *(White)*	DD *(Black)*	DDD
	(1)	(2)	(3)
I(score<14) x Post	0.175*** (0.021)	0.094** (0.037)	0.174*** (0.021)
I(score<14) x Black			0.026 (0.031)
Post x Black			−0.0004 (0.033)
I(score<14) x Post x Black			−0.080** (0.035)
Mean Dep. Var. *(Pre-HB463)*	0.312	0.298	0.310
Observations	119,427	22,662	142,089
R^2	0.274	0.248	0.270
Adjusted R^2	0.271	0.231	0.267

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 5, for the full sample. For Column (3), the coefficient of interest is the triple interaction of "I(score<14) x Post x Black." All specifications use the following controls: separate fixed effects for month-year, day of week, exact risk score, top charge level and class, judge, county, defendant gender, and defendant race. All specifications also control for the same characteristics factoring into risk scores from Specification 4. Standard errors are always clustered at the judge-level. *p<0.1; **p<0.05; ***p<0.01.

Instead of looking at recommendation effects across risk levels, I can investigate the differences across the entire risk score distribution. I can thereby compare white and Black defendants who share identical underlying risk scores (as well as identical algorithmic recommendations). Figure 10 shows that after recommendations are in effect, Black and white defendants with identical risk scores receive similar levels of lenient bail when the scores are too high for the lenient recommendation (to the right of the orange dashed line). However, when scores qualify for the lenient recommendation (to the left of the orange dashed line), white people are consistently more likely to receive lenient bail than Black defendants with identical risk scores. The figure also shows that this was not true in the pre-period. In the pre-period, lenient bail rates for white and Black defendants frequently overlap, and in a few cases (cases with scores of 3-5 or 14), Black defendants are more likely to receive lenient bail. Therefore, the recommendations generated Black-white gaps in lenient bail that were not present beforehand.

Figure 10: Lenient Bail Rates by Defendant Race, Risk Score, and Time Period

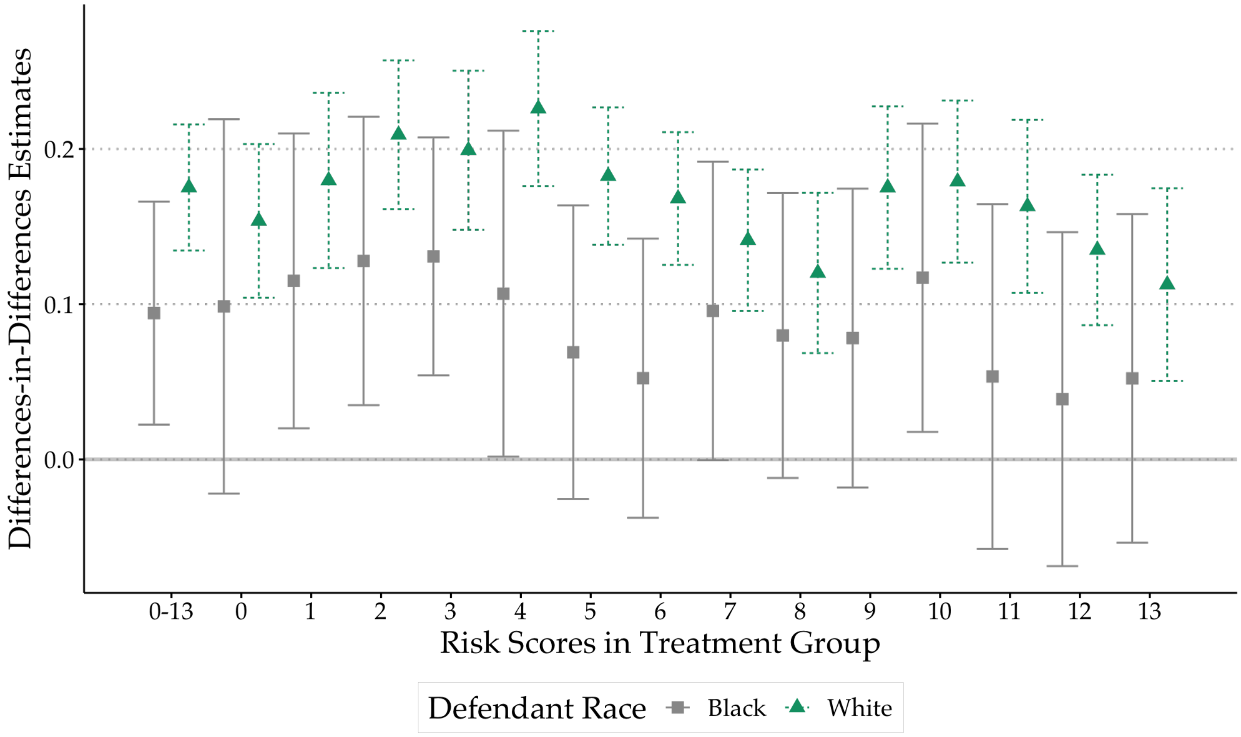


Notes: This figure shows the rates of lenient bail for Black and white defendants across the risk score distribution both before and after HB463. The orange dotted line shows the critical threshold for the lenient bail recommendation in the post-HB463 period. Cases with Black defendants are shown as green triangles, while cases with white defendants are shown as gray squares. Rates are shown for cases before HB463 in the left panel and for cases after HB463 in the right panel.

To formalize the motivational evidence, I estimate differences-in-differences coefficients for each risk score in the low-moderate distribution (scores of 0-13) and compare the results for Black and white defendants. This exercise is equivalent to generating Figure 5 separately by defendant racial group. Figure 11 shows the results.

The estimates for white defendants are consistently larger than those for Black defendants. Figure A.14 demonstrates that the estimated coefficients across the distribution are similar regardless of specification and control choices. Another way to validate these results is by estimating the triple differences specification across the risk score distribution. Figure A.15 shows that the triple differences coefficients are consistently negative across the distribution. In effect, the recommendations increased lenient treatment for white defendants more than for Black defendants with identical risk scores.

Figure 11: Differences-in-Differences Estimates across Risk Scores by Defendant Race



Notes: This figure shows the pooled difference-in-differences coefficients across treatment groups based on risk scores, estimated separately by defendant racial group. The control group consists of cases with high risk scores, and the treated group consists of cases with low or moderate risk scores (varying from 0 to 13). Specifications are estimated separately for each race (white or Black) by risk score group. The specification 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. The error bars show the 95% confidence intervals for each differences-in-differences coefficient.

8.2 Why Is There Racial Heterogeneity in Recommendation Effects?

Bail decisions in Kentucky are made by elected judges. The heterogeneity in recommendation effects by defendant race could come from different treatment of defendants within the same judges. However, it could also come from similar treatment of defendants within judges but different choice patterns across judges. Judges work in counties, meaning that different judges make decisions for very different defendant populations. Figure A.16 shows the vast variation across the state in Black defendant populations. In over half of Kentucky's counties, less than 5% of defendants are Black (see Figure A.17). However, in a handful of counties, more than 30% of defendants are Black. Because of this uneven spatial variation in racial demographics, different responses across judges could generate racial disparities at the state level.

To test this question, I re-estimate Specification 5 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 A.6. The coefficient of interest (the coefficient 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 (see column 3). The evidence in Table A.6 demonstrates that different responses across judges drive most of the differences in effects by defendant race. In other words, compared with judges who see more white defendants, judges who see more Black defendants respond less to lenient recommendations.[17]

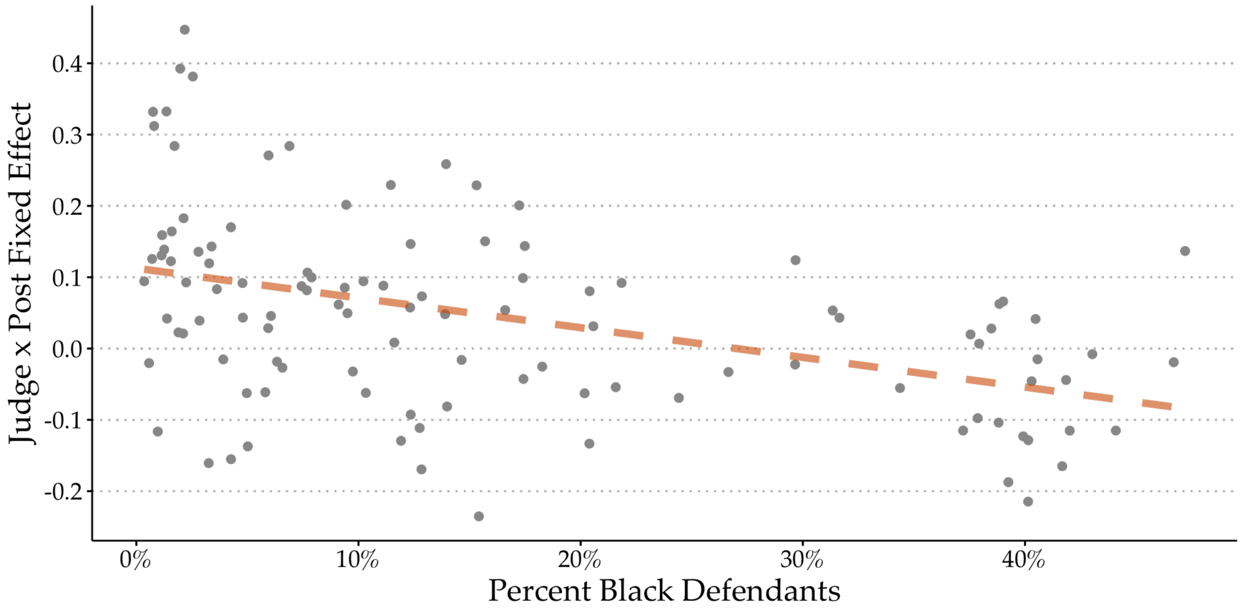
I can further demonstrate the link between defendant population and judge responsiveness. I subset to cases with risk scores below 14, so that time period is the only component that impacts recommendation presence. I then limit the sample to cases decided by judges with 50 or more bail decisions before and after HB463, so I focus on judges who see an adequate number of cases before and after. This restriction selects 114 judges out of the original 433, but omits only 2.5% of observations (because many of the unique judges in the data see very few cases). Therefore, 97.5% of cases in the original sample are in this sub-sample. With this restriction, I estimate the following specification:

$$lenient_{itj} = \beta_3[Post_t \times Black_i] + X_{itj} + \epsilon_{itj}, \tag{6}$$

[17] These results are consistent with prior work by Stevenson (2018), who found that judges who see more white defendants become more lenient in terms of pretrial release after HB463. Once she allowed for time-varying fixed effects for different circuit courts, there were no remaining changes to racial disparities.

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 the judge sees who are Black) in Figure 12.

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 6 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.

Figure 12 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 from the pooled baseline effect of 15 percentage points.

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 irrespective of whether I estimate the specification with no control variables, only case information control variables, or all possible control variables. Figure A.18 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.

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 instance, 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 associated with higher pretrial misconduct could respond less, because 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 county-level characteristics (crime, population, etc.).

To explore whether other factors explain the relationship between judge-post coefficients and the racial composition of the defendant population, I collect variables from interest from various 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 number 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 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 A.7 shows that the range of collected judge-level and county-level covariates do not drown out the magnitude of the relationship between defendant demographics and judge recommendation responsiveness. However, the relationship loses its statistical significance when the rural variable is included.

These results connect to the theory that recommendations change the costs of errors to judges. If judges think recommendations give them less political cover in more racially heterogeneous places, this belief will generate the empirical patterns observed in the data. This possible mechanism is conceptually similar to results from Feigenberg and Miller (2021), that 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%.[18] 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.

My results complicate public discussion about 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 meant to alleviate it.

9 Conclusion

As algorithms continue to be integrated into high-stakes decisions, studying their impact on human decisions becomes increasingly necessary. In this paper, I point out that algorithms' predictions are often translated into recommendations for decision-makers. However, there is little public discussion of these algorithmic recommendations. I highlight the recommendations' theoretical and empirical importance by demonstrating their independent causal effects on human decisions.

I demonstrate three key results in this paper using a well-suited natural experiment in the US criminal justice setting. First, my results show that algorithmic recommendations have first-order effects on human decisions. Judges make lenient bail decisions 50% more often when given a lenient recommendation.

Second, I provide evidence that recommendations matter for decision-making because they can change the costs of errors to decision-makers. Namely, judges may become more

[18] In Kentucky, the largest shares of Black defendants map onto the peak area of Feigenberg and Miller (2021)'s U-shaped 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.

lenient when their choices are consistent with recommendations, because the recommendation shields them from backlash. Therefore, recommendations can change more than just the allocation of decisions *(who gets which decision)* – they can change the overall composition *(how many decisions are lenient)*. If decision-makers and algorithmic designers disagree about the costs of errors, then recommendations could better align decision-maker incentives with social planner objectives (e.g., less money bail) (McLaughlin and Spiess 2022). Algorithmic recommendations are therefore a type of what Cowgill and Stevenson (2020) call "algorithmic social engineering": recommendations are derived from algorithmic predictions, but adjusted to meet certain policy objectives.

Finally, I show that algorithmic recommendations can have heterogeneous effects. Judges deviate from lenient recommendations more for Black defendants than for white defendants with identical algorithmic scores. The recommendations generate racial disparities because the adherence to algorithmic recommendations varies across decision-makers who serve different populations, as is consistent with prior work by Stevenson (2018). The interaction of the political economy of high-stakes decisions and the implementation of algorithmic systems can have unintended effects on group inequality.

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.

American Bar Association Criminal Justice Standards Committee. 2007. *ABA Standards for Criminal Justice: Pretrial Release, Third Edition*.

Angelova, Victoria, Will Dobbie, and Crystal Yang. 2023. "Algorithmic Recommendations and Human Discretion." *Working Paper 31747, National Bureau of Economic Research*.

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.

Arnold, David, Will Dobbie, and Crystal Yang. 2018. "Racial Bias in Bail Decisions." *Quarterly Journal of Economics* 133 (4): 1885–1932.

Austin, James, Roger Ocker, and Avi Bhati. 2010. "Kentucky Pretrial Risk Assessment Instrument Validation." *Research Paper, JFA Institute*.

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.

Bushway, Shawn, Emily Owens, and Anne Morrison Piehl. 2012. "Sentencing Guidelines and Judicial Discretion: Quasi-Experimental Evidence from Human Calculation Errors." *Journal of Empirical Legal Studies* 9 (2): 291–319.

Calonico, Sebastian, Matias Cattaneo, and Max Farrell. 2020. "Optimal Bandwidth Choice for Robust Bias-Corrected Inference in Regression Discontinuity Designs." *Econometrics Journal* 23 (2): 192–210.

Calonico, Sebastian, Matias Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82 (6): 2295–2326.

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.

Cornell Law School Legal Information Institute. 2024. "Bondsman." *Wex Legal Dictionary and Encyclopedia*. https://www.law.cornell.edu/wex/bondsman.

Covert, Bryce. 2022. "Why New York Jail Populations Are Returning to Pre-Pandemic Levels." *The Appeal*. https://theappeal.org/new-york-jail-population-increase/.

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 Megan Stevenson. 2020. "Algorithmic Social Engineering." *AEA Papers and Proceedings* 110: 96–100.

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

Davenport, Diag. 2023. "Discriminatory Discretion: Theory and Evidence from Use of Pretrial Algorithms." *Working Paper, Princeton University*.

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*.

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

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

Grembi, Veronica, Tommaso Nannicini, and Ugo Troiano. 2016. "Do Fiscal Rules Matter?" *American Economic Journal: Applied Economics* 8 (3): 1–30.

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

Hoffman, Mitchell, Lisa 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. 2018. “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 Rumman Chowdhury. 2021. “What Is an ’Algorithm’? It Depends on Who You Ask.” *MIT Technology Review*. https://www.technologyreview.com/2021/02/26/1020007/what-is-an-algorithm/.

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

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.

Price, Michael. 2011. “Kentucky Population Growth: What Did the 2010 Census Tell Us?” *Kentucky State Data Center Research Report* 1 (1): 1–13.

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

Skeem, Jennifer, 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*.

Sloan, CarlyWill, George Naufal, and Heather Caspers. Forthcoming. “The Effect of Risk Assessment Scores on Judicial Behavior and Defendant Outcomes.” *Journal of Human Resources*, Forthcoming.

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

Stevenson, Megan, and Jennifer Doleac. Forthcoming. "Algorithmic Risk Assessment in the Hands of Humans." *American Economic Journal: Economic Policy*, Forthcoming.

Stevenson, Megan, and Sandra Mayson. 2018. "Pretrial Detention and Bail." In *Reforming Criminal Justice, Volume 3, Edited by Erik Luna*, 21–47. Arizona State University.

Stevenson, Megan, and Christopher Slobogin. 2018. "Algorithmic Risk Assessments and the Double-Edged Sword of Youth." *Behavioral Sciences & the Law* 36 (5): 638–56.

Appendix

A.1 Kentucky Risk Assessment Institutional Details

Kentucky has used a few different risk assessment scoring tools over the years. The first 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 the Laura and John Arnold Foundation developed.

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 took the form of "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 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 older 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 13-23 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.2 Data Preparation

Determining the appropriate observation level: The data preparation work required defining an appropriate observation level. There can be multiple charges in a case, multiple cases in an interview, multiple bonds for a cases, and so on. These factors complicate the question of how to clean and prepare the data for analysis.

I took the following steps to define an interpretable and relevant level of observation. First, I aggregated data on charges at the case level. Second, I limited the data to pretrial interviews with defendants where one case was at issue. (This is necessary to think about bail decisions that apply to a single well-defined case rather than a potential bundle of cases.) Third, I focused on the first bail setting for each case. This bond instance is commonly called initial bail.

Dealing with changes to arrestable offenses: One challenge to studying HB463 is that it was a large bill of about 150 pages and 110 sections.[19] 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.[20] 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 identify the relevant cases. In other words, I subset to cases with offenses that were arrestable over the full study time period.

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

[20] 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."

A.3 Kentucky Pretrial Services Institutional Details

Unlike pretrial services in other states, Kentucky Pretrial Services is part of the judicial branch; it is a state entity that works for the courts.[21] Kentucky has one pretrial services agency, which serves all 120 counties in the state, meaning that data management and collection are unified and well-organized. Pretrial employees are housed in individual counties but do not work for the counties.

Kentucky is well known for its pretrial services for a few reasons. First, in 1976 it became the first state to ban commercial bail bonds. It was one of four states with this ban as of 2022 (Cornell Law School Legal Information Institute 2024). Second, 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.

Kentucky has been using phone calls for pretrial services since 1976. Kentucky uses phone calls because people are very spread out in parts of the state, which would make in-person bail hearings costly in terms of commute time.

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, which I listened to. These calls included the following information: name, age, risk score information, list of charges, and incident description. The incident description quoted information from the relevant police report.

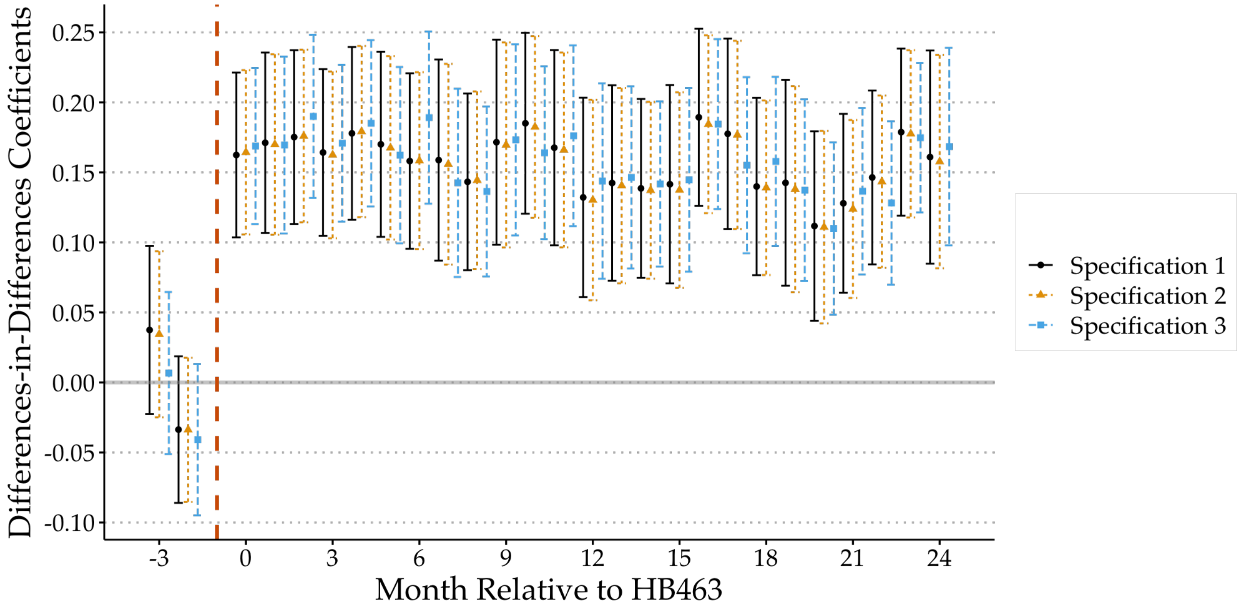
Note that while demographic information on race or gender is not explicitly in the calls, these details may be implicitly included. Gender is revealed through the use of pronouns (e.g., “he” and “she”) when the pretrial officer discusses the defendant. Meanwhile, names (especially in combination with 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.

[21] Information in the following paragraphs is sourced from an interview with Tara Blair, former executive officer of Kentucky Pretrial Services.

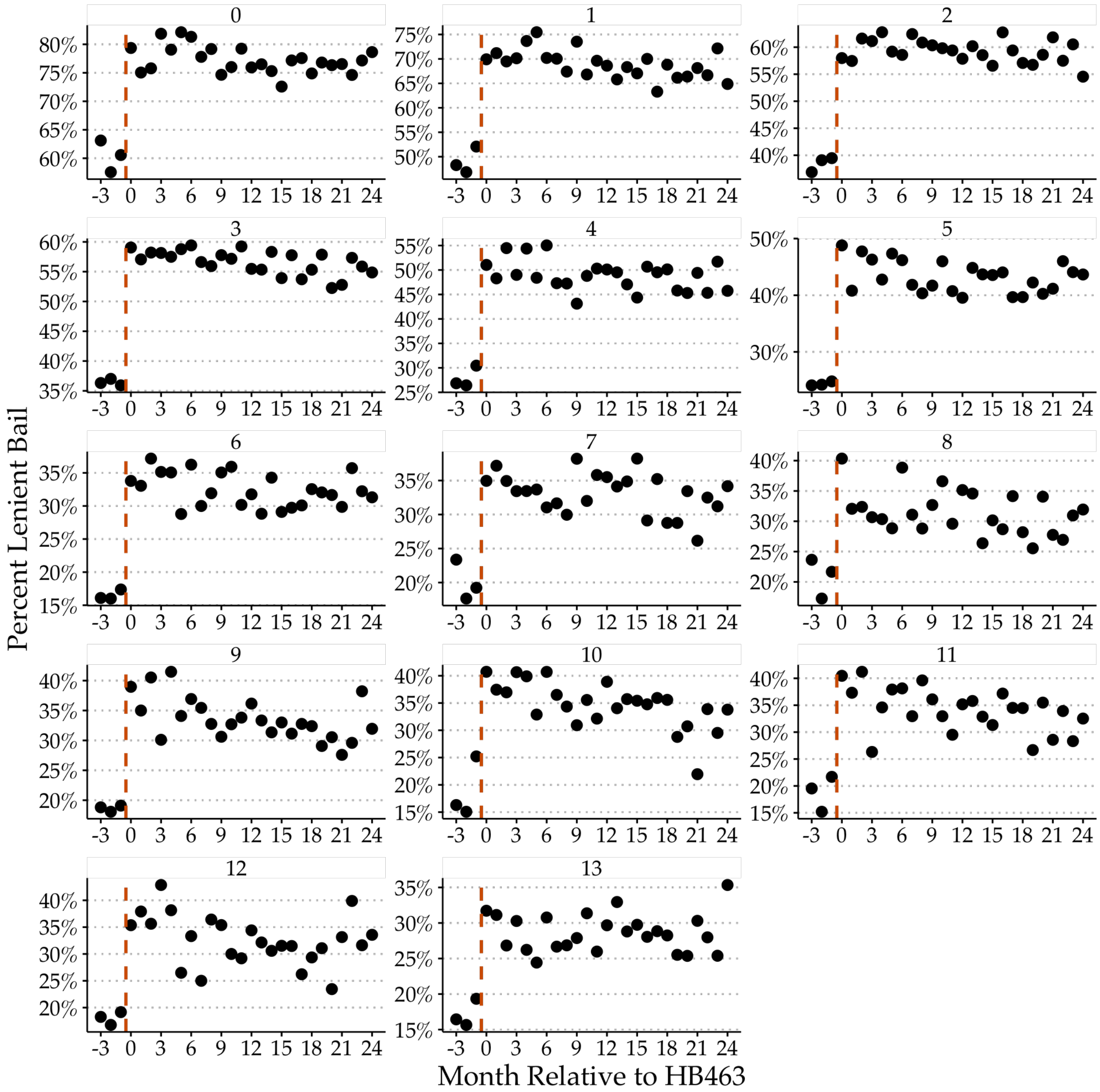
A.4 Appendix Exhibits for Section 6.1

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



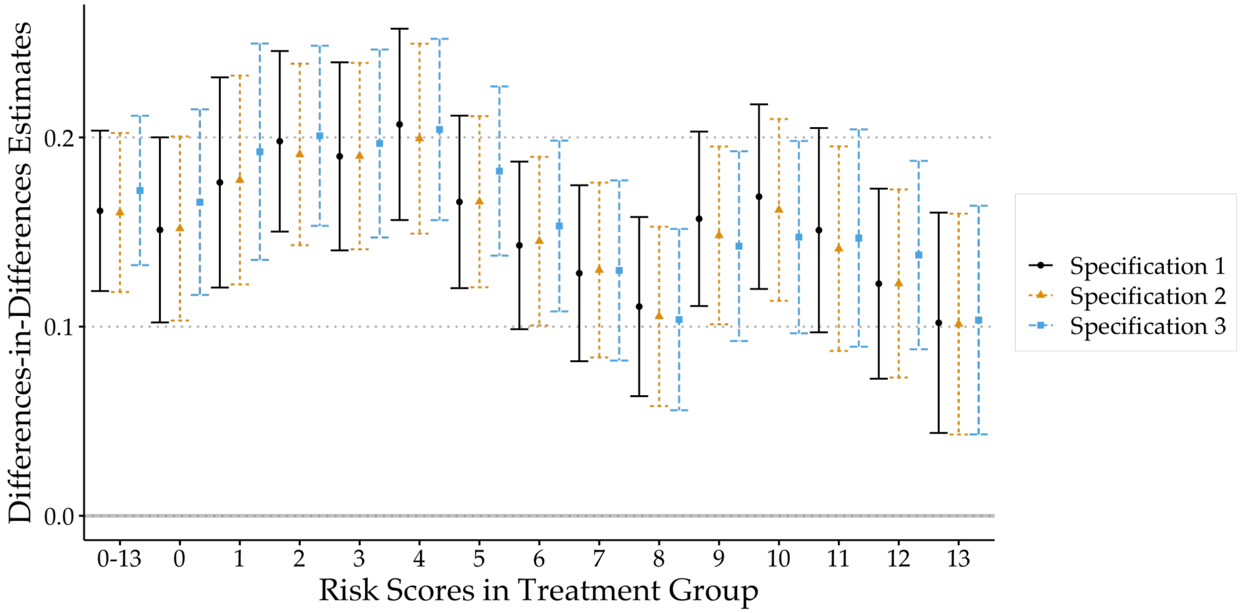
Notes: This figure shows the difference-in-differences coefficients for months relative to recommendation introduction across specifications. The control group consists of cases with high risk scores, and the treated group consists of cases with low or moderate risk scores. The orange dashed line denotes the omitted period of the month before the recommendation introduction. Specification 1 (black circles and error bars) is the main specification, also shown in Figure 4, 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. All error bars denote 95% confidence intervals.

Figure A.2: Lenient Bail Rates by Risk Score over Time



Notes: This figure shows the rate of lenient bail over months by risk scores for low and moderate risk cases. Months are indexed relative to the introduction of algorithmic recommendations. The orange dotted line shows when HB463 went into effect. Each plot is for a different discrete value of risk score between 0 and 13.

Figure A.3: Pooled Differences-in-Differences Estimates across Risk Score Values and Specifications



Notes: This figure shows the pooled difference-in-differences coefficients across different treatment groups based on risk scores and across different specifications. The control group consists of cases with high risk scores, and the treated group consists of cases with low or moderate risk scores (varying from 0 to 13). Specification 1 (black circles and error bars) is the main specification and 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. All error bars denote 95% confidence intervals.

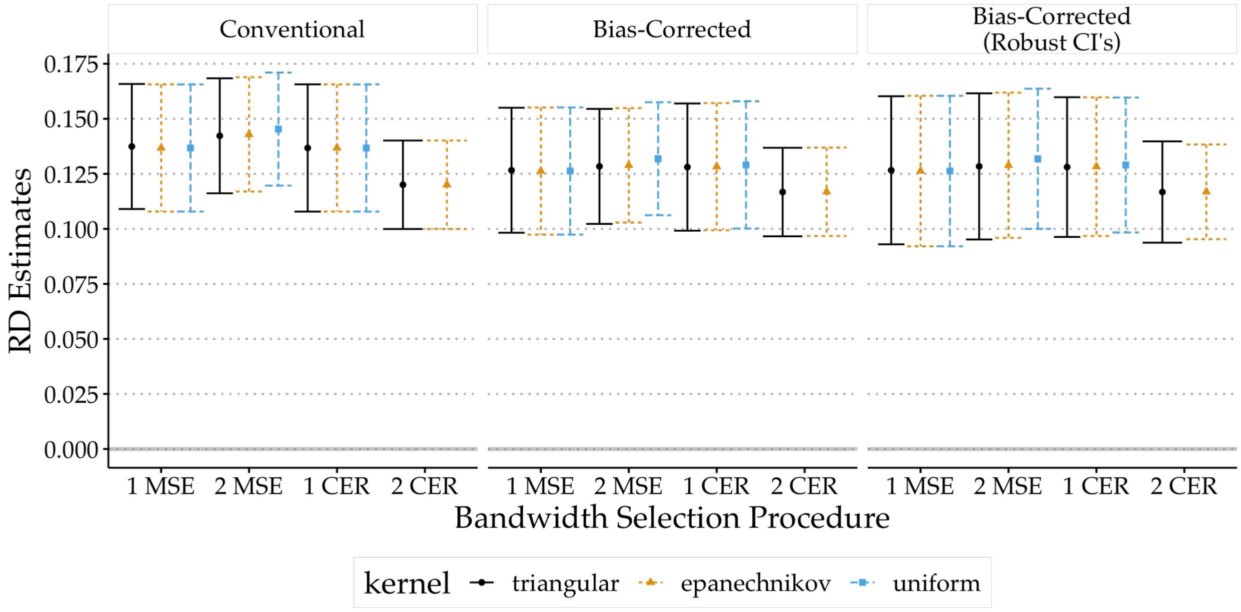
Table A.3: Differences-in-Differences Results across Specifications

	Dependent variable: I(lenient bail)		
I(score<14) x Post	0.161^{***}	0.160^{***}	0.172^{***}
	(0.022)	(0.021)	(0.020)
Pre-Mean Score<14	0.310	0.310	0.310
Time/Score FEs	Y	Y	Y
Charge/judge/county/demographic controls	Y	Y	N
Risk component controls	Y	N	N
Observations	142,466	142,466	142,466
R^2	0.270	0.264	0.133
Adjusted R^2	0.266	0.261	0.132
Note:	*p<0.1; **p<0.05; ***p<0.01		

Notes: This table displays estimated differences-in-difference coefficients in specifications with lenient bail as the dependent variable. The control group consists of cases with high risk levels, and the treated group consists of cases with low or moderate risk levels. The table shows results across different specifications. The complete set of controls includes fixed effects for month-year, day of week, exact risk score, top charge level and class, judge, county, defendant gender, and defendant race, and 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.

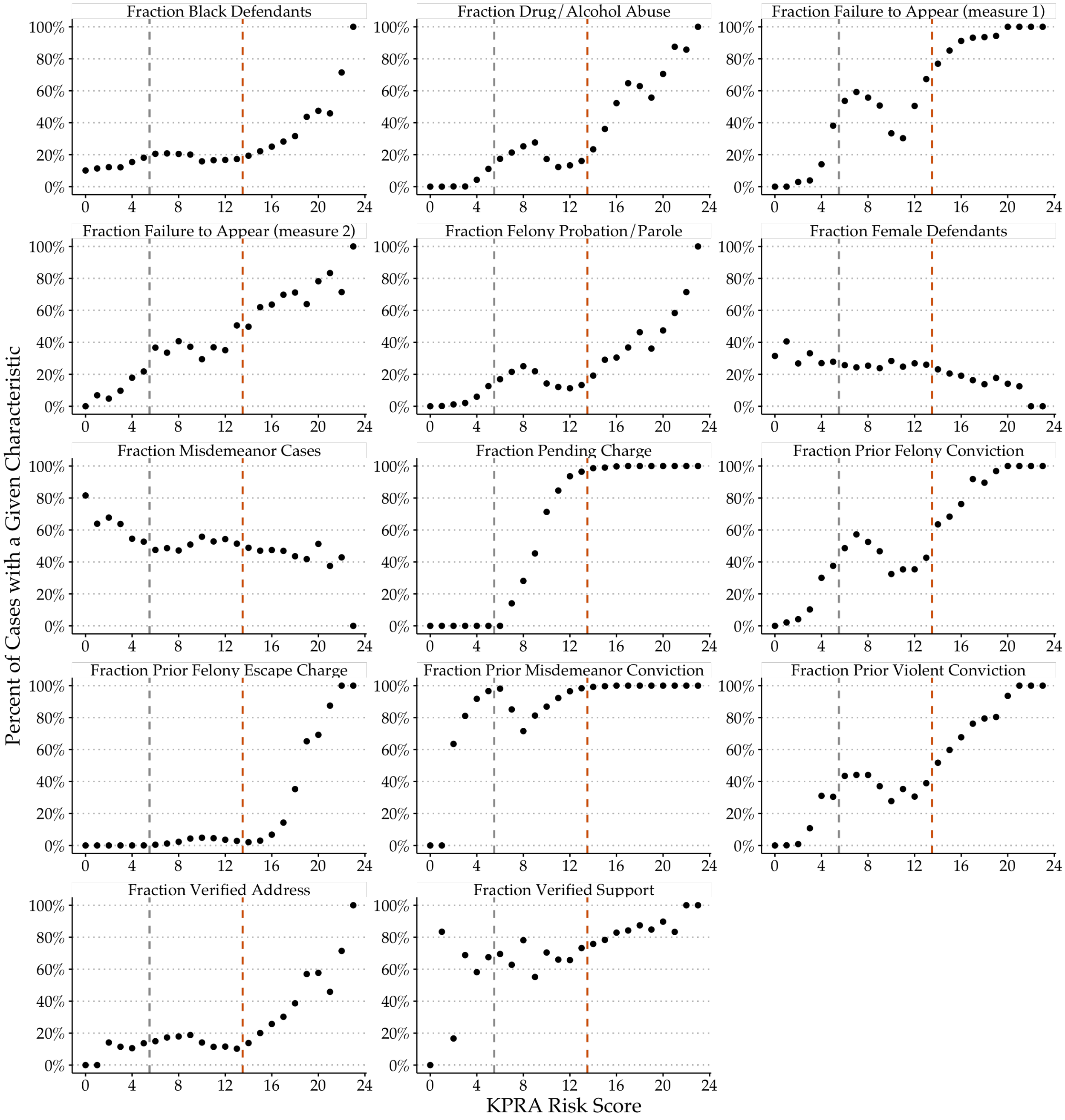
A.5 Appendix Exhibits for Section 6.2

Figure A.4: RD Robustness across Specification Choices (Post-Period, Moderate-High Threshold)



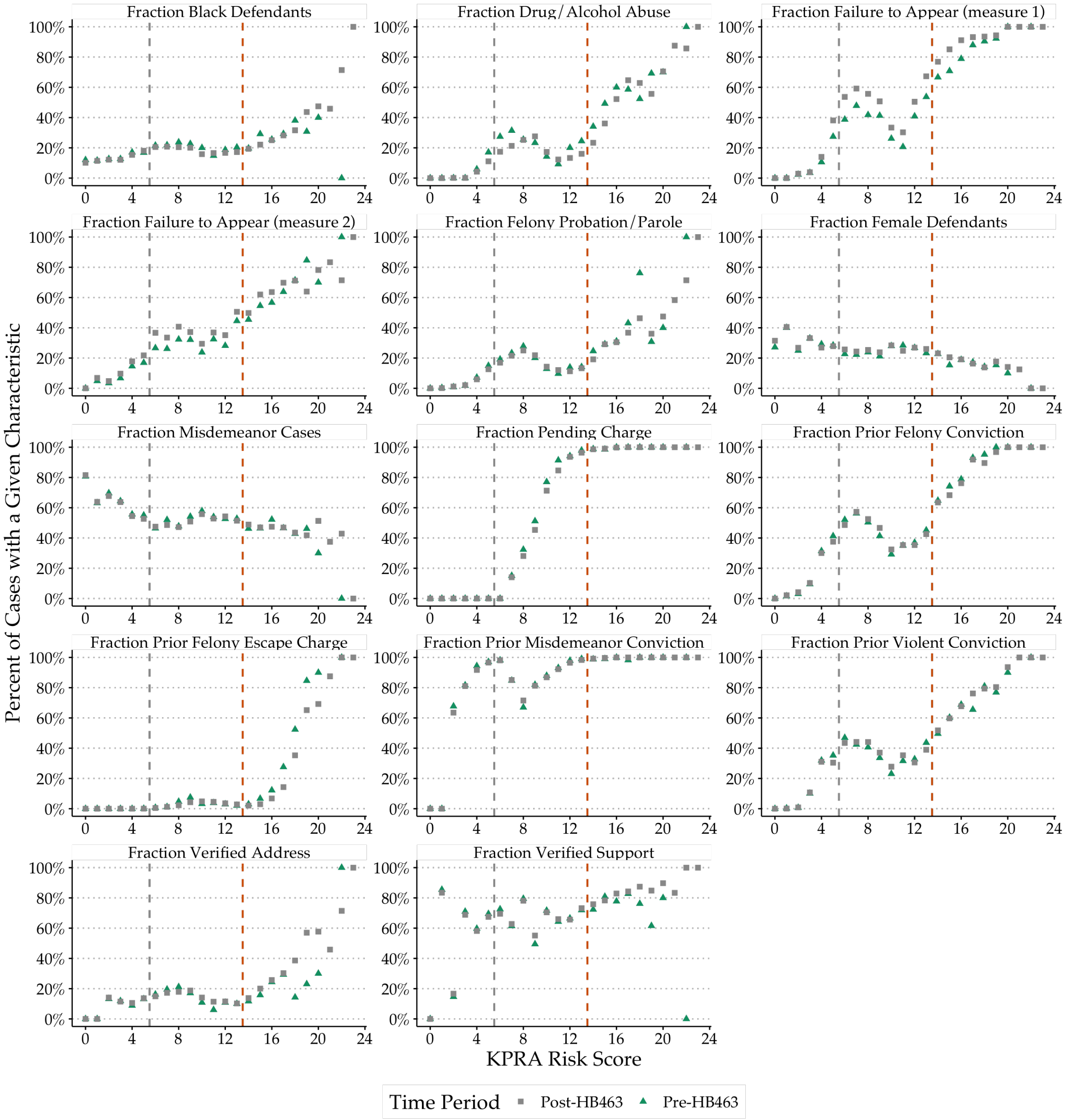
Notes: This figure plots the regression discontinuity estimates using post-period data around the moderate-high cut-off across a range of bandwidth selection procedures, kernels, and estimation adjustments. The shaded gray line at 0 shows the estimate of 0 percentage points. I show estimates with 95% confidence intervals. The plots are grouped based on whether I use conventional estimation, bias-corrected estimation, or bias-corrected estimation with robust bias-corrected confidence intervals. I show different kernel choices with distinct colors, shapes, and line types. The x-axis differentiates between the bandwidth selection procedures (1 MSE = one common Mean Square Error-optimal bandwidth selector; 2 MSE = two different Mean Square Error-optimal bandwidth selectors [below and above the cut-off]; 1 CER = one common Coverage Error Rate-optimal bandwidth selector; 2 CER = two different Coverage Error Rate-optimal bandwidth selectors [below and above the cut-off].

Figure A.5: Defendant and Case Covariates over Risk Score Distribution (Post-Period)



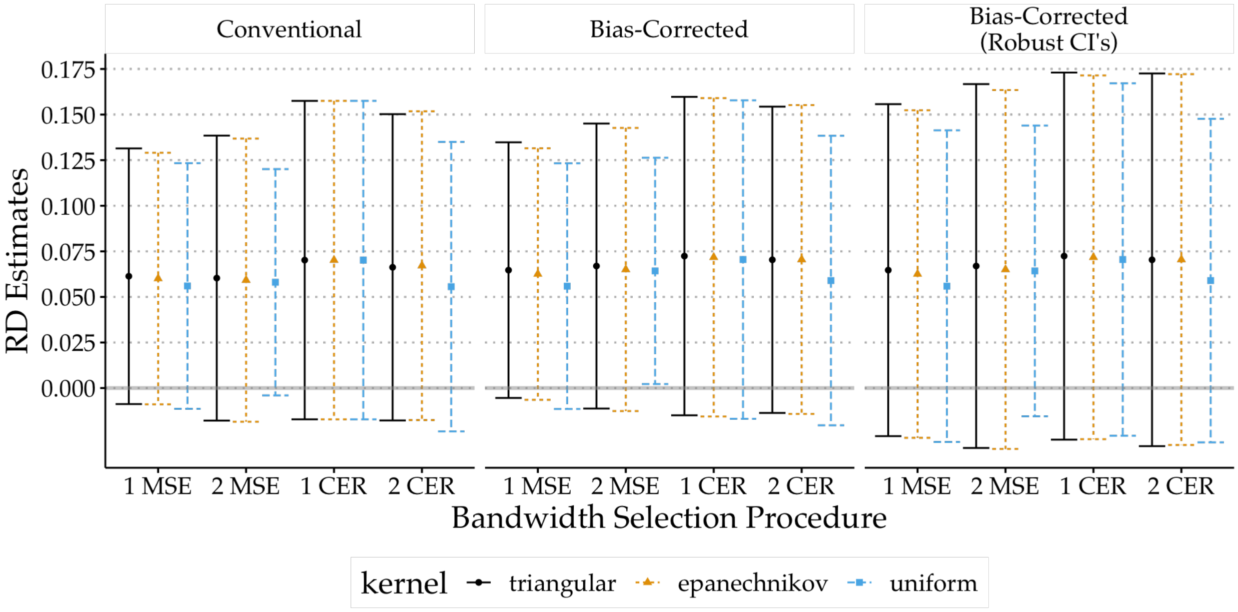
Notes: This figure shows the average defendant and case covariates for each discrete case risk score using data from the post-period. The dashed lines indicate the cut-offs between risk levels. The orange line is the threshold between "moderate" and "high" risk, and the gray line is the threshold between "low" and "moderate" risk. Scores of 0-5 are low risk, scores of 6-13 are moderate risk, and scores of 14 or over are high risk.

Figure A.6: Defendant and Case Covariates over Risk Score Distribution and Time Periods



Notes: This figure shows each discrete case risk score's average defendant and case covariates. The gray rectangles show the averages before HB463, while the green triangles show the averages after HB463. The dashed lines indicate the cut-offs between risk levels. The orange line is the threshold between "moderate" and "high" risk, and the gray line is the threshold between "low" and "moderate" risk. Scores of 0-5 are low risk, scores of 6-13 are moderate risk, and scores of 14 or over are high risk.

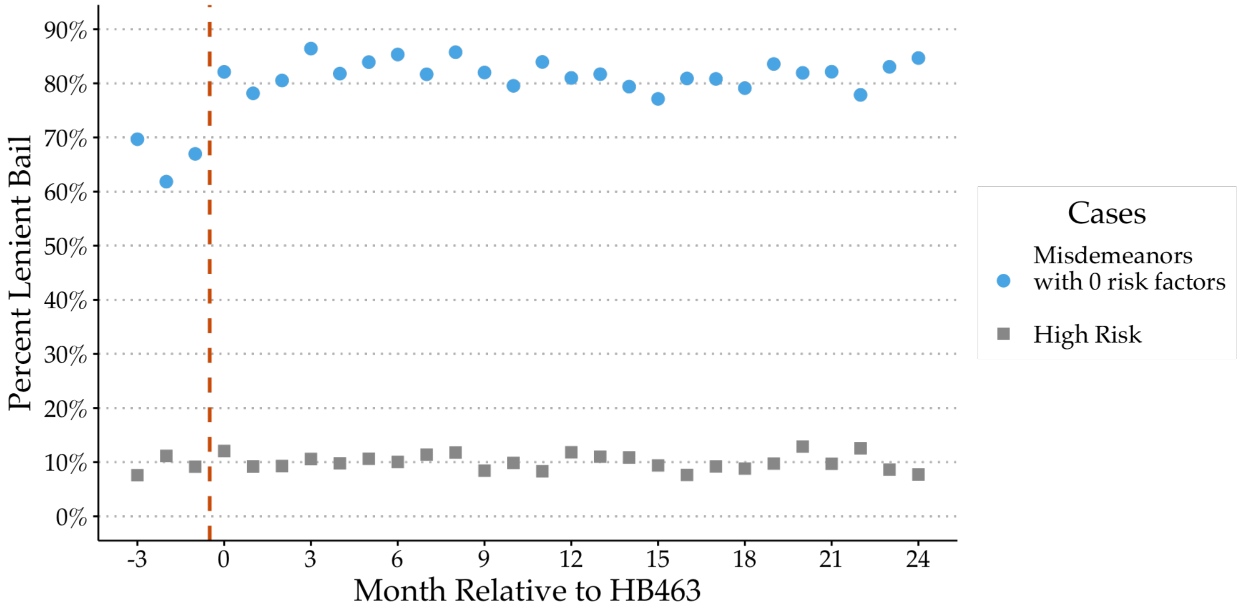
Figure A.7: RD Robustness across Specification Choices (Pre-Period, Moderate-High Threshold)



Notes: This figure plots the regression discontinuity estimates using pre-period data around the moderate-high cut-off across a range of bandwidth selection procedures, kernels, and estimation adjustments. The shaded gray line at 0 shows the estimate of 0 percentage points. I show estimates with 95% confidence intervals. The plots are grouped based on whether I use conventional estimation, bias-corrected estimation, or bias-corrected estimation with robust bias-corrected confidence intervals. I show kernel choices with distinct colors, shapes, and line types. The x-axis differentiates between the bandwidth selection procedures (1 MSE = one common Mean Square Error-optimal bandwidth selector; 2 MSE = two different Mean Square Error-optimal bandwidth selectors [below and above the cut-off]; 1 CER = one common Coverage Error Rate-optimal bandwidth selector; 2 CER = two different Coverage Error Rate-optimal bandwidth selectors [below and above the cut-off].

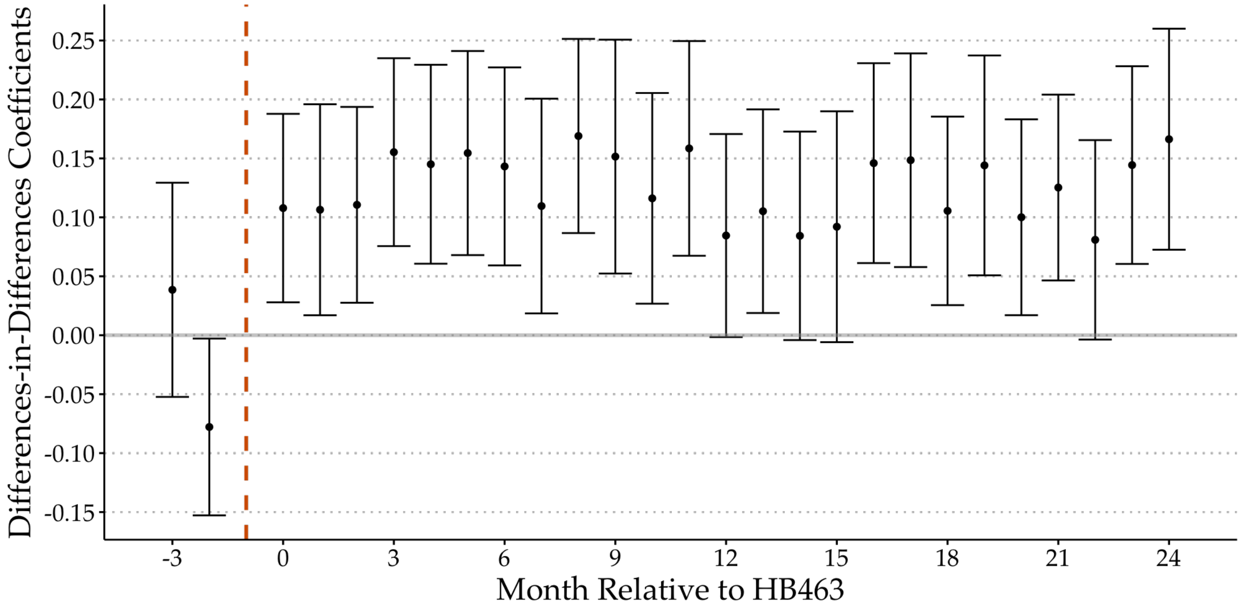
A.6 Appendix Exhibits for Section 7.2

Figure A.8: Lenient Bail Rates by Case Type 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. Misdemeanor cases with risk scores of 0 are shown as blue circles, while cases with high risk scores are shown as gray squares. The orange dotted line shows when HB463 went into effect.

Figure A.9: Dynamic Differences-in-Differences Estimates (Treated Group: Lowest Risk Cases)



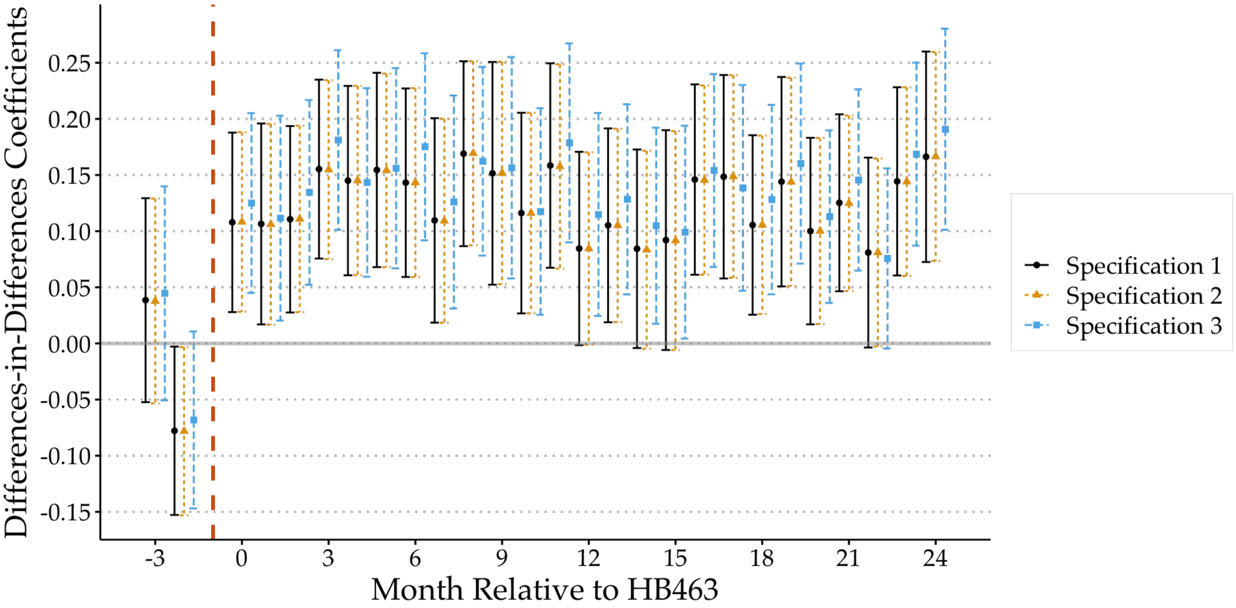
Notes: This figure shows the difference-in-differences coefficients for months relative to the recommendation introduction. The control group consists of cases with high risk scores, and the treated group consists of cases with risk scores of 0 and misdemeanor offenses. The orange dashed line denotes the omitted period of the month before the recommendation introduction. All error bars denote 95% confidence intervals.

Table A.4: Differences-in-Differences Results across Specifications (Treated Group: Lowest Risk Cases)

	Dependent variable: I(lenient bail)		
I(score<14) x Post	0.146***	0.147***	0.155***
	(0.026)	(0.026)	(0.027)
Pre-Mean Score<14	0.659	0.659	0.659
Time/Score FEs	Y	Y	Y
Charge/judge/county/demographic controls	Y	Y	N
Risk component controls	Y	N	N
Observations	18,904	18,904	18,904
R^2	0.552	0.552	0.490
Adjusted R^2	0.540	0.540	0.489

Notes: This table displays estimated differences-in-differences coefficients in specifications with lenient bail as the dependent variable. The control group consists of cases with high risk levels, and the treated group consists of misdemeanor cases with risk scores of 0. The table shows results across different specifications. The full set of controls includes fixed effects for month-year, day of week, exact risk score, top charge level and class, judge, county, defendant gender, and defendant race, and 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.

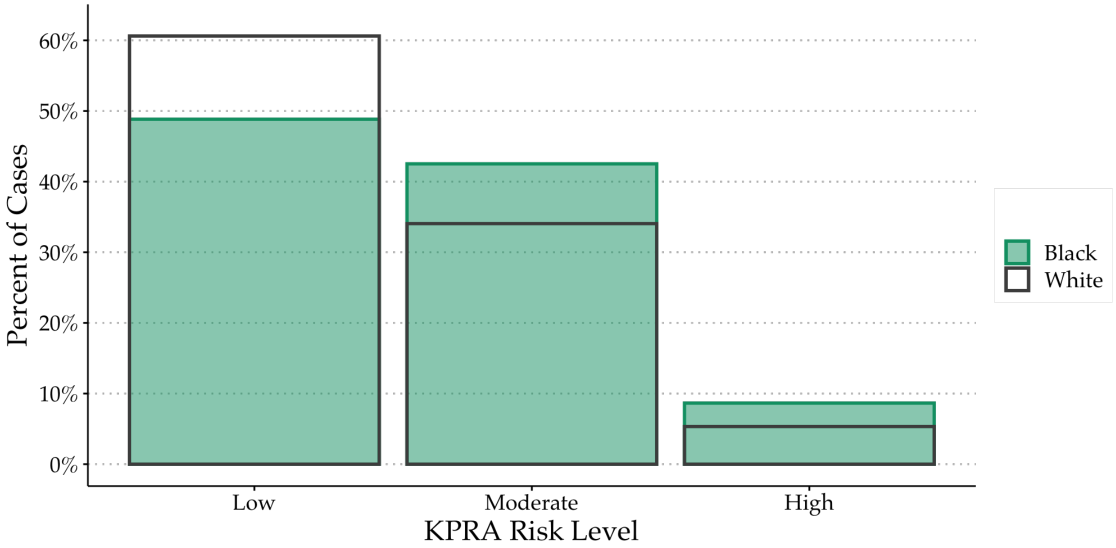
Figure A.10: Dynamic Differences-in-Differences Estimates across Specifications (Treated Group: Lowest Risk Cases)



Notes: This figure shows the difference-in-differences coefficients for months relative to recommendation introduction across specifications. The control group consists of cases with high risk scores, and the treated group consists of cases with risk scores of 0 and misdemeanor offenses. The orange dashed line denotes the omitted period of the month before the recommendation introduction. Specification 1 (black circles and error bars) is the main specification, also shown in Figure 4, 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. All error bars denote 95% confidence intervals.

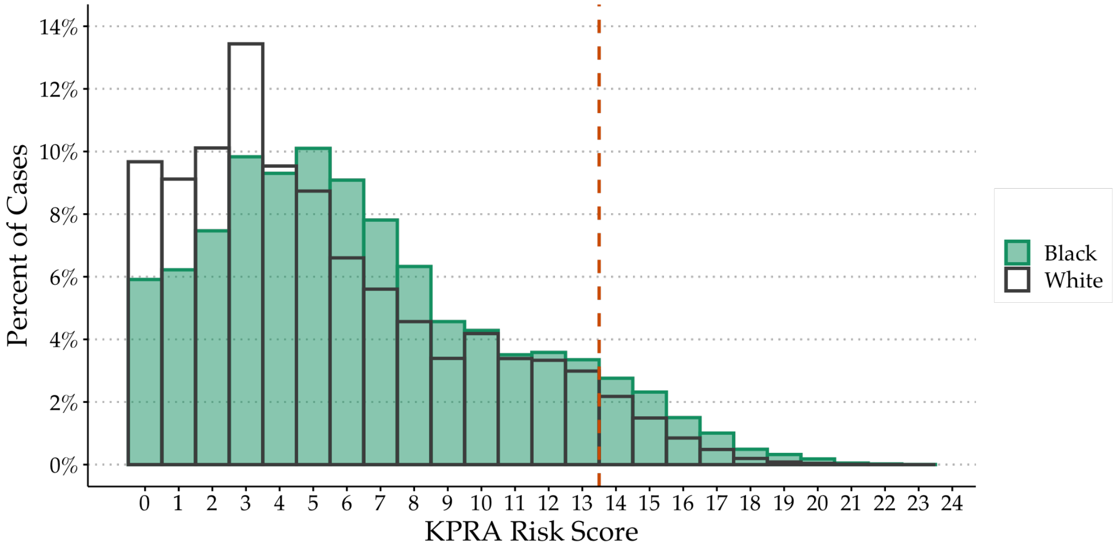
A.7 Appendix Exhibits for Section 8

Figure A.11: Distribution of Risk Levels by Defendant Race



Notes: This figure illustrates the distribution of cases across the risk level distribution for Black and white defendants. Solid green bars represent cases with Black defendants, while unfilled gray bars represent cases with white defendants.

Figure A.12: Distribution of Risk Scores by Defendant Race



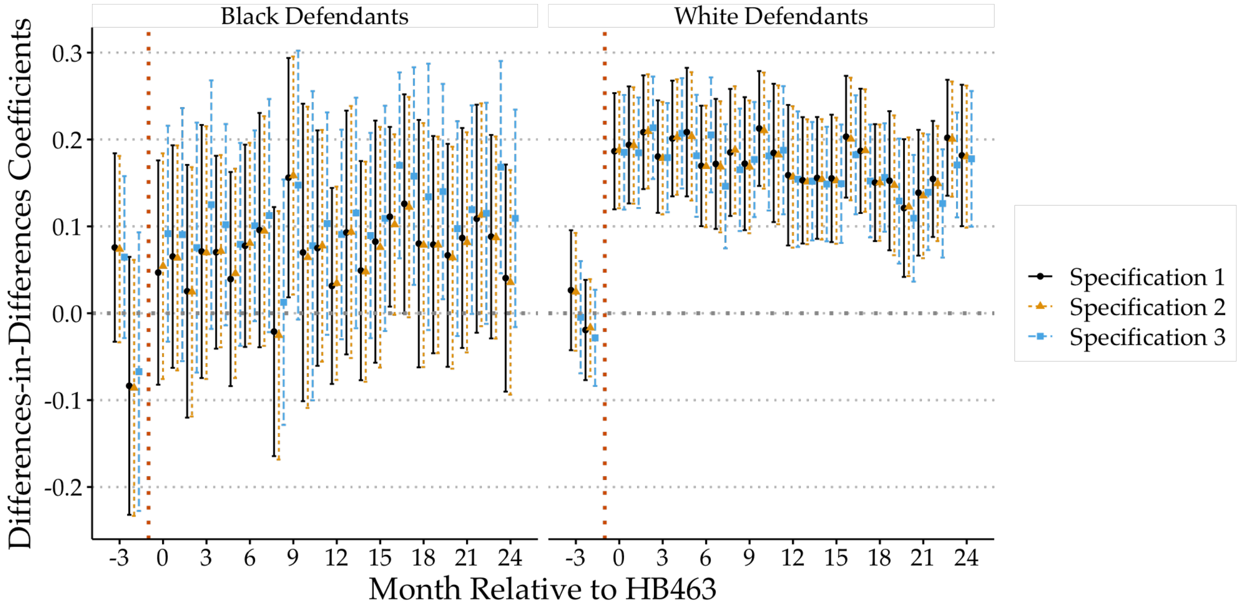
Notes: This figure illustrates the distribution of cases across the risk score distribution for Black and white defendants after HB463. Solid green bars represent cases with Black defendants, while unfilled gray bars represent cases with white defendants. The orange dashed line shows the threshold between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

Table A.5: Fraction of Defendants with Different Risk Score Factors by Defendant Race

Risk Component	**Black**	**White**	**p-value**
Verified Address	0.843	0.890	<0.001
Verified Support	0.414	0.411	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

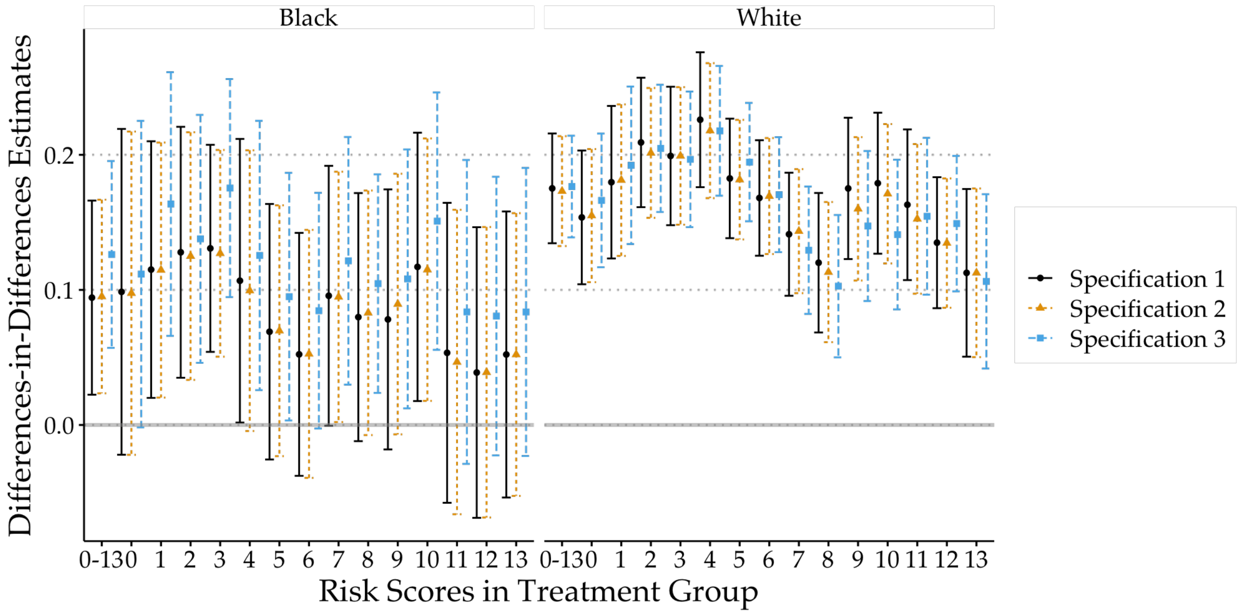
Notes: This table shows what fraction of defendants have characteristics that factor into the algorithm-based risk scores. The p-value tests if the fractions are statistically different in the Black and white defendant populations.

Figure A.13: Dynamic Differences-in-Differences Estimates across Specifications, by Race



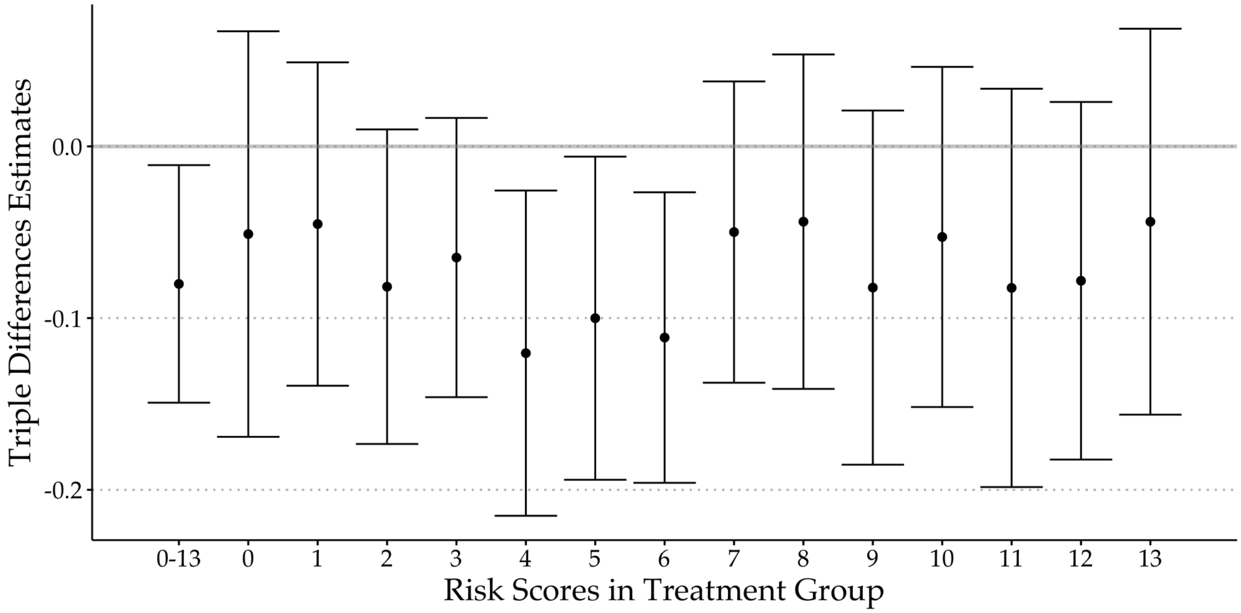
Notes: This figure shows the difference-in-differences coefficients for months relative to the recommendation introduction across specifications for both white and Black defendants. The orange dashed line denotes the omitted period of the month before the recommendation introduction. Specification 1 (black circles and error bars) is the main specification, also shown in Figure 4, 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. All error bars denote 95% confidence intervals.

Figure A.14: Differences-in-Differences Estimates by Defendant Race, Risk Score Value, and Specifications



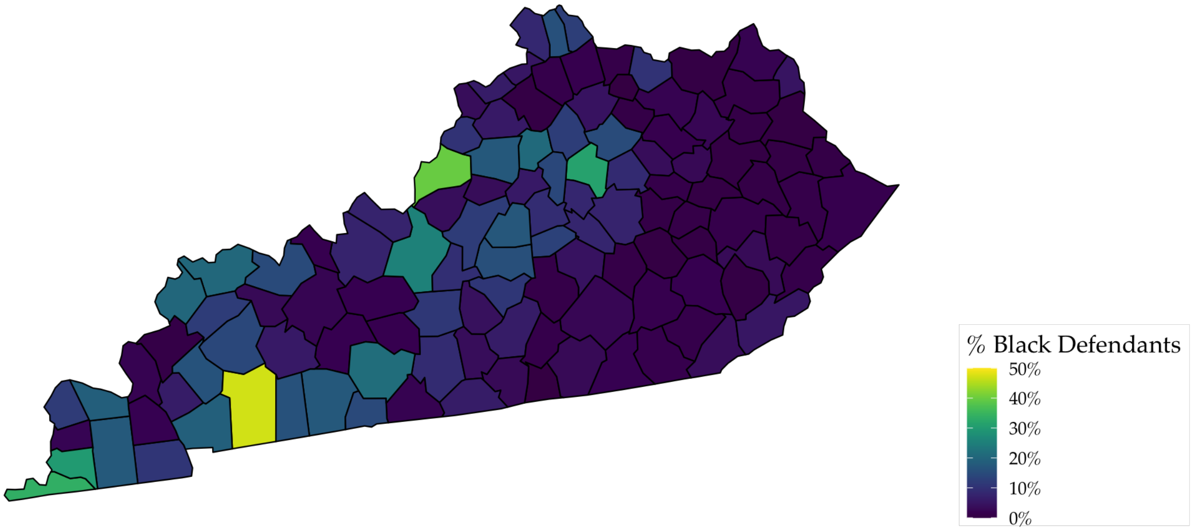
Notes: This figure shows the pooled difference-in-differences coefficients across different treatment groups based on risk scores. The figure also shows how these coefficients vary across different specifications. Results are shown separately for Black defendants and white defendants. The control group consists of cases with high risk scores, and the treated group consists of cases with low or moderate risk scores (varying from 0 to 13). Specification 1 (black circles and error bars) is the main specification, also shown in Figure 11 in the main text. This specification 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. All error bars denote 95% confidence intervals.

Figure A.15: Triple Differences Estimates across Risk Score Bandwidths



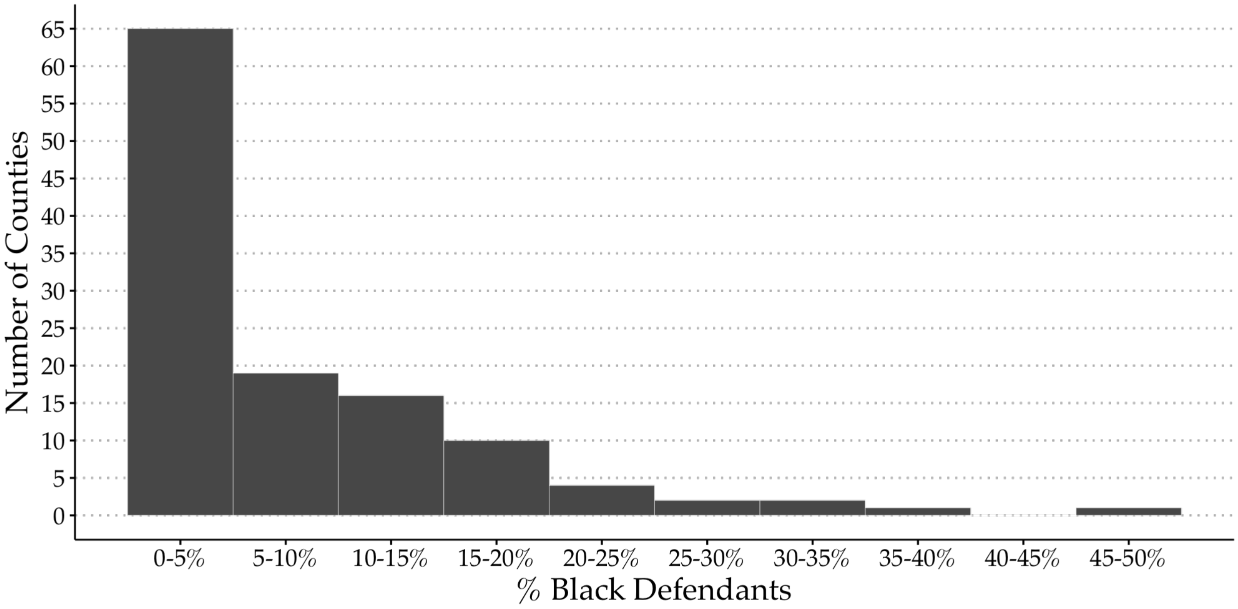
Notes: This figure shows the pooled triple differences coefficients by racial group across different treatment groups based on risk scores. The control group consists of cases with high risk scores, and the treated group consists of cases with low or moderate risk scores (varying from 0 to 13). Specifications are estimated separately for each risk score group. The coefficient of interest is the triple interaction of "I(score<14) x Post x Black." The specification 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. The error bars show the 95% confidence interval for each estimate.

Figure A.16: Choropleth of Percent Black Defendants across Kentucky Counties



Notes: This figure displays the percentage of defendants who are Black across all of Kentucky's 120 counties. The dark blue colors signify lower percentages, while the bright green and yellow colors signify higher percentages.

Figure A.17: Histogram of Percent Black Defendants across Kentucky Counties



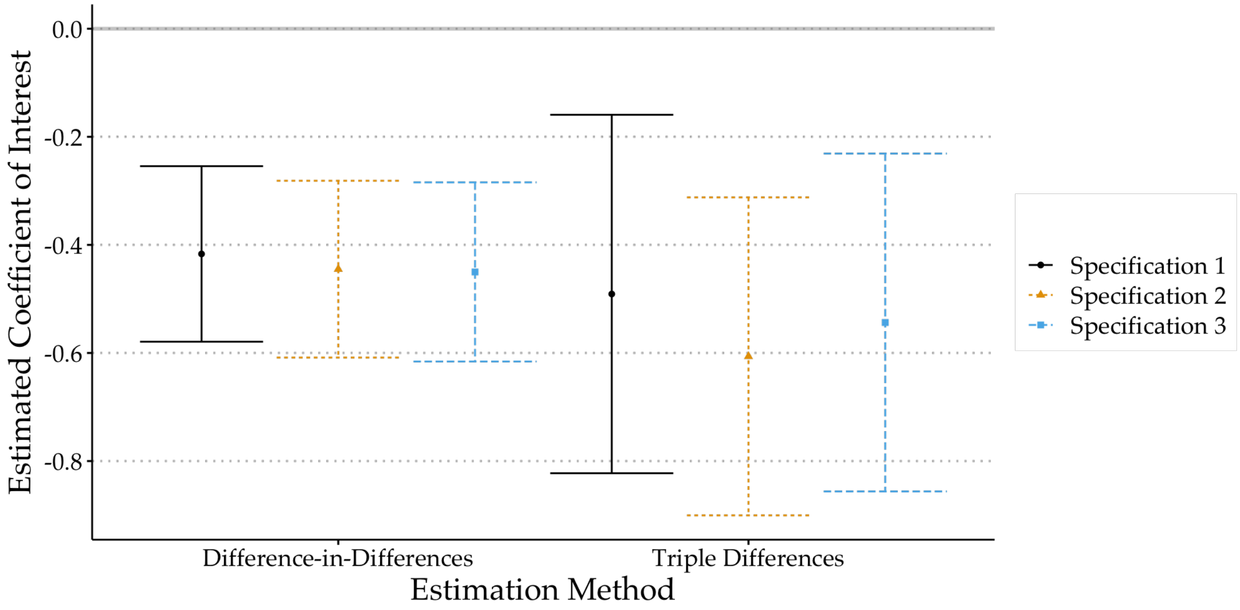
Notes: This figure is a histogram that displays the number of counties with different percentages of Black defendants. The binwidths are 5 percentage points each.

Table A.6: Triple Differences Results with Additional Interacted Fixed Effects

	Dependent variable: I(lenient bail)		
	DDD	DDD	DDD
	(1)	(2)	(3)
I(score<14) x Post	0.174^{***}		
	(0.021)		
I(score<14) x Black	0.026	-0.013	-0.013
	(0.031)	(0.036)	(0.028)
Post x Black	-0.0004	-0.003	0.001
	(0.033)	(0.031)	(0.025)
I(score<14) x Post x Black	-0.080^{**}	-0.017	-0.024
	(0.035)	(0.035)	(0.029)
Mean Dep. Var. *(Pre-HB463)*	0.310	0.310	0.310
Additional Controls	-	*judge-level-time varying FE's*	*county-level-time varying FE's*
Observations	142,089	142,089	142,089
R^2	0.270	0.282	0.276
Adjusted R^2	0.267	0.265	0.268

Notes: Column (1) shows the results from the triple differences specification, Specification 5, for the full sample. For Column (1), the coefficient of interest is the triple interaction of "I(score<14) x Post x Black." Columns (1), (2), and (3) all estimate triple differences specifications, but they vary in whether they include additional fixed effects. Column (2) allows for judge-level-time varying fixed effects (each judge has distinct fixed effects depending on the risk score (I(score<14)) and time period (Post)). Column (3) allows for county-level-time varying fixed effects (each county has distinct fixed effects depending on the risk score (I(score<14)) and time period (Post)). All specifications use the following controls: separate fixed effects for month-year, day of week, exact risk score, top charge level and class, judge, county, defendant gender, and defendant race. All specifications also control for the same characteristics factoring into risk scores from Specification 4. Standard errors are always clustered at the judge-level. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Figure A.18: 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 made at least 50 bail decisions before and after HB463. This restriction yields 114 judges. In the case of the three estimates on the left, only the subset of data with risk scores under 14 is used for estimation. Fixed effects are extracted from differences-in-differences approaches ("DD") that interact indicators for post-period and Black defendant with each other. In the case of the three estimates on the right, fixed effects are extracted from triple differences specifications ("DDD") that interact indicators for post-period, Black defendant, and risk score under 14 with each other. Estimates across different sets of control variables are illustrated with distinct colors, shapes, and line types. Specification 1 (black circles and error bars) is the main specification and 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.

Table A.7: Explaining Judge Responsiveness to Lenient Recommendations

	Dependent Variable = Judge x Post FE					
	(1)	(2)	(3)	(4)	(5)	(6)
Share Black Defendants	−0.383***	−0.391***	−0.378**	−0.320**	−0.275	−0.345*
	(0.084)	(0.088)	(0.151)	(0.157)	(0.178)	(0.187)
Black Judge		0.057	0.068	0.057	0.052	0.059
		(0.070)	(0.072)	(0.073)	(0.075)	(0.075)
Woman Judge		−0.023	−0.013	−0.015	−0.018	−0.021
		(0.025)	(0.027)	(0.027)	(0.029)	(0.028)
Years as Judge		−0.001	−0.003	−0.002	−0.002	−0.001
		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Contested in 2010			−0.036	−0.033	−0.039	−0.029
			(0.034)	(0.034)	(0.036)	(0.036)
Any Contest in District in 2010			−0.0003	−0.007	−0.005	−0.021
			(0.036)	(0.037)	(0.037)	(0.038)
log(Election Voters)			−0.003	−0.003	0.015	0.003
			(0.023)	(0.026)	(0.044)	(0.043)
Rearrest Rate Pre-HB463				0.576	0.569	0.639
				(0.381)	(0.385)	(0.392)
FTA Rate Pre-HB463				−0.198	−0.208	−0.212
				(0.381)	(0.385)	(0.381)
log(County Population)					−0.023	−0.045
					(0.042)	(0.045)
Rural County					−0.016	−0.030
					(0.045)	(0.046)
Total Crime Rate						−0.00005
						(0.00005)
Total Index Crime Rate						0.009
						(0.008)
Property Crime Rate						−0.009
						(0.008)
Violent Crime Rate						−0.009
						(0.008)
Constant	0.108***	0.128***	0.170	0.138	0.225	0.540
	(0.018)	(0.026)	(0.208)	(0.225)	(0.304)	(0.340)
N	94	94	94	94	94	94
R^2	0.185	0.206	0.221	0.242	0.245	0.302
Adjusted R^2	0.176	0.170	0.158	0.161	0.144	0.168

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 where a 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. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.