The Hidden Effects of Algorithmic Recommendations

Alex Albright*

April 2025

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. 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 30-40% for marginal cases. The results are consistent with algorithmic recommendations making mistakes less costly to judges by reducing their liability. In this way, algorithms may affect human decisions by changing incentives, in addition to informing predictions.

*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, Northwestern, 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 for editorial support, Sara Brandel for data agreement assistance, and Amisha Kambath for 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. 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 unique setting in which algorithmic predictions were already present, but algorithmic recommendations were introduced. 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. For instance, it is common for judges to set money bail, which requires defendants to post money to be released from jail. During my study period, judges making money bail decisions received information on alleged incidents and involved defendants. One piece of information available to them was an algorithmic prediction of pretrial misconduct (rearrest or failure to appear in court) for each case. Before June 2011, there were no recommendations based on these risk predictions. But, in June 2011, a new policy set recommendations for judges based on the predictions: judges were recommended to not set money bail when a case's predicted risk was low or moderate (rather than high). I call these recommendations "lenient bail" recommendations because not setting money bail is a more lenient decision than setting money bail. The institutional details yield useful variation for causal inference: lenient bail recommendations kicked in discontinuously across risk (under the "high" risk cut-off) and time (after June 2011).

Why might these recommendations change decision-maker behavior? I develop a model in which judges make bail decisions based on pretrial misconduct predictions 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 incentives because they impact liability. In particular, visible mistakes become less costly to judges when their decisions adhere to recommendations. The two theories generate dueling testable predictions in the Kentucky empirical setting. If recommendations only change predictions, the new recommendations should have no effects on bail setting. If recommendations change costs, then lenient bail should have increased for lower-risk cases.

To test these predictions, I estimate the causal effects of algorithmic recommendations. I leverage the fact that only some cases received lenient recommendations to implement 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 causal effects for the entire distribution of cases in the low and moderate risk groups.

My other identification approach – differences-in-discontinuities – 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. Since confounding factors around the threshold in the post-period were also present in the pre-period, I use a differences-in-discontinuities approach to recover the lenient recommendation effect. 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 the 2011 policy change increased lenient decisions by a striking 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 are included in the specifications. The relative effects are similarly large (around 40%) for the marginal moderate risk cases when using differences-in-discontinuities. These straight-forward results suggest that algorithmic recommendations change human decisions.

However, more empirical work is needed to isolate the causal effect of algorithmic recommendations. Both the differences-in-differences and differences-in-discontinuities strategies leverage the fact that recommendations were introduced for some cases but not others. To correctly attribute these estimated effects to the causal effect of recommendations, it must be the case that at the time of the policy change, nothing else differentially impacted low and moderate risk cases relative to high risk cases. While the calculation of risk was the same before and after the policy and risk levels were available in both periods, there is a potential confounding issue: the policy mandated judges consider algorithmic predictions. Therefore, if some judges did not consider algorithmic predictions before the policy, then the policy changed the presence of algorithmic predictions *and* recommendations. This means the straight-forward differences-in-differences and differences-in-discontinuities results are necessarily an upper-bound on the recommendation effect of interest.

I use two empirical approaches to isolate the desired recommendation effect from my original estimates. First, I use a data-driven approach to estimate the share of cases in which judges were using risk levels before the policy. I do this by estimating regression discontinuities before and after the policy at the low-moderate threshold because risk levels change at this cut-off but recommendations do not (low and moderate risk cases receive the same recommendation). I find that risk levels were used in about 80% of cases in the pre-period. I then use this rate to put bounds on the recommendation effect. I find that lenient recommendations increased lenient bail by 30-40% for marginal cases. Therefore, the majority of the original 50% effect is attributable to the independent causal effects of algorithmic recommendations.

My second way of addressing this potential confound is a simple intuitive approach. I test whether recommendations matter for a unique subset of cases where the expected effect of risk levels is small. Specifically, I look at cases that are associated with misdemeanor charges and have zero associated risk factors (zero failures to appear, zero pending cases, zero convictions, etc.). Intuitively, the risk level does not provide new information to judges for these cases because they are obviously low risk. Therefore, the differences-in-differences result for this group should only capture the effect of algorithmic recommendations. Consistent with meaningful effects of algorithmic recommendations, I find that lenient bail increases by about 15 percentage points for this extreme low-risk group.

Overall, my results shows algorithmic recommendations have economically important effects on human decisions, and these effects are independent of changes to algorithmic predictions. Bringing the results back to my testable predictions, my results are consistent with the theory that algorithmic recommendations change decision-maker incentives. In particular, algorithmic recommendations may change how decision-makers perceive their individual liability. If a recommended choice results in bad outcomes, some of the blame goes to the recommendation designer rather than all of the blame going to the individual decision-maker. In this way, the communication of predictive algorithms may change human decision-making by changing incentives in addition to directly providing new prediction information.

My paper makes new contributions to the literature on algorithms and human decisions. When studying the effects of algorithms, researchers may contrast a world without algorithms, where humans have complete discretion, to one with algorithms where there is no human discretion (Berk 2017; Mullainathan and Obermeyer 2022; Kleinberg, Lakkaraju, et al. 2018; Cowgill 2018a). However, since humans usually make the final decisions even when algorithms are present, this is often not the policy-relevant comparison. With that in mind, researchers may compare 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 greatly across these settings. For instance, some settings provide decision-makers with algorithmic predictions only, while others also provide explicit algorithmic recommendations. My paper demonstrates that these specifics matter for the causal effects of algorithms.[1] In particular, *algorithmic*

[1]By demonstrating the causal importance of algorithmic recommendations, my paper also contributes to an interdisciplinary literature 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

recommendations matter: they have independent effects on human decisions and merit independent attention in policy discussions and research.

This paper is one of the first to focus directly on the causal effects of algorithmic recommendations. In doing so, I complement two concurrent papers that also study these recommendations. First, McLaughlin and Spiess (2022) investigate the distinction between algorithmic predictions and recommendations by developing a theoretical model in which algorithmic recommendations may directly change preferences (rather than only changing beliefs).[2] While they develop theoretical results showing how recommendations can have independent effects on human decisions, I demonstrate these independent effects in practice. Second, Hausman (2024) studies how changes to algorithmic recommendations changed human decisions in another context: US Immigration and Customs Enforcement (ICE). He finds that ICE release rates decreased by 50% after release recommendations were removed – a result that is consistent with the large effects in my setting. However, ICE's algorithmic predictions changed at the same time as the recommendations, so it is not possible to disentangle the two effects in Hausman (2024)'s setting. My paper, therefore, progresses the evidence by rigorously isolating the causal effects of algorithmic recommendations.

To study the effects of algorithmic recommendations, I leverage a 2011 policy change in Kentucky, which was first studied by Stevenson (2018). Her paper was one of the first to discuss the important distinction between how algorithms change outcomes in theory and practice. While Stevenson (2018) studied the effects of the 2011 policy, I leverage this policy to study the effects of *algorithmic recommendations*. Therefore, instead of demonstrating the effects of a treatment bundle (a few things changed with the 2011 policy), I isolate the effects of my treatment of interest (algorithmic recommendations only). Doing this requires a series of novel steps: using court reports to find a time period when the calculation of algorithmic predictions was unchanged but recommendations changed, using court report documents to calculate cases' underlying risk scores, and using data-driven approaches to address confounding policy changes (changes to arrests and changes to risk level visibility).

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; Angelova, Dobbie, and Yang 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).

[2]A number of other papers discuss the idea that algorithms in general may change decision-maker incentives (Davenport 2023; Stevenson and Doleac Forthcoming; Stevenson and Doleac 2023), but McLaughlin and Spiess (2022) is unique in that they discuss algorithmic recommendations changing these incentives. In related work, Almog et al. (2024) studies how AI oversight changes the costs of errors to decision-makers. When decisions can be publicly overruled by AI, mistake costs increase.

The algorithmic recommendations studied in this paper 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. My paper shows that, even as prediction tools get more complex, advisory recommendations continue to have independent causal effects in the justice system.

The remainder of the paper proceeds as follows. Section 2 provides background on algorithms and decision-making to highlight the underappreciated role of algorithmic recommendations. Section 3 describes my empirical setting – bail decisions in Kentucky – and the variation used for identification. Section 4 develops a theoretical model, which generates testable predictions to differentiate between dueling theories of the effects of algorithm recommendations. Section 5 describes the administrative court data. Section 6 presents my results, which isolate the causal effects of algorithmic recommendations. Section 7 concludes.

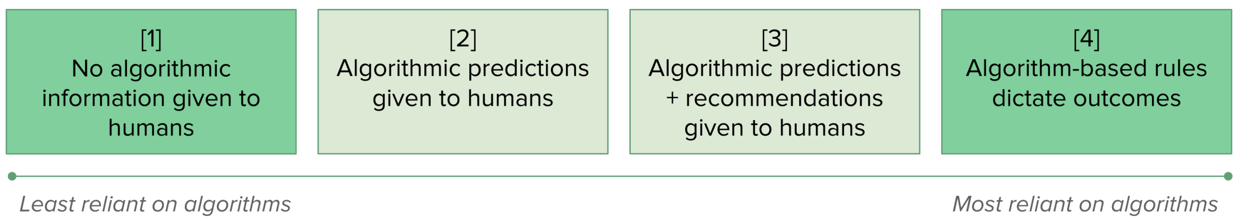
2 Background on Algorithms and Decisions

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.

In fact, 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.

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.

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



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.

Using the spectrum shown in Figure 1, I can spatially situate different strands of the literature on algorithms and decision-making. Research showing that algorithms alone can outperform human decision-makers (Berk 2017; Mullainathan and Obermeyer 2022; Kleinberg, Lakkaraju, et al. 2018; Cowgill 2018a) contrasts (1) with (4). Meanwhile, research showing how outcomes change when humans are given algorithms but have 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) contrasts (1) with either (2) or (3). My paper contributes to the literature by studying the underappreciated distinction between (2) and (3). In doing so, I 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. Consider, for example, the coordinated entry system in the US, which dictates how housing gets allocated to people experiencing homelessness. Individual risk assessment plays a role in ordering people in terms of their vulnerability or housing need, and then available housing is allocated in that order (County of Santa Clara 2025; Mingle 2020; U.S. Department of Housing and Urban Development 2017). In this context, the housing supply 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

In general, algorithmic predictions and recommendations are often introduced simultaneously, which makes it difficult to isolate the effects of recommendations. However, in Kentucky, algorithmic predictions were used both before and after the introduction of algorithmic recommendations. The nature of the introduction of algorithmic recommendations in Kentucky, therefore, provides a unique opportunity to estimate the independent effects of algorithmic recommendations.

The algorithmic predictions: From March 18, 2011 to June 30, 2013, Kentucky used one fixed algorithm to make predictions about cases. It was called the Kentucky Pretrial Risk Assessment (KPRA) and made predictions about pretrial misconduct (rearrest or failure to appear in court).

Kentucky 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 were 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).[3]

The algorithmic 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.[4] HB463 introduced recommendations for low and moderate risk cases but not for high risk cases. Importantly, the policy change did not change the calculation of the risk scores or levels; it introduced recommendations for how to use them.

Bail decisions before June 2011: In the pre-period, 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 and the alleged incident to a judge over the phone. One piece of information that could be discussed was the KPRA risk level. After hearing this information, the judge made a bail decision in a few minutes.[5]

Bail decisions after June 2011: In the post-period, bail decisions were made in a similar way as before, but with two changes. First, the new recommendations were part of the judge conversations: cases with low or moderate KPRA risk levels were recommended lenient bail. Second, risk levels were a mandated part of the conversation (previously, they had been optional).

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

[3]See Appendix A.1.2 for more background and details on risk calculation in Kentucky.

[4]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.

[5]See Appendices A.1.1 and A.1.3 for more background on the Kentucky bail setting.

4 Theoretical Model and Testable Predictions

In this section, I present a theoretical model to demonstrate that different theories of recommendations yield different empirical predictions in my setting. I contrast two theories: first, that recommendations only matter by conveying predictive information; second, that recommendations change decision-maker incentives. I focus on these two theories because the first captures the conventional focus on algorithmic predictions, while the second acknowledges that recommendations can have direct, independent effects. The two distinct empirical predictions are testable with my data, allowing my empirical findings to inform which theory better explains the observed effects.

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 not for defendants. I call money bail harsh bail ($b = h$) and no money bail lenient bail ($b = l$).

If the judge sets harsh bail, there is some probability the defendant is detained $Pr(d = 1|b = h)$, and there is some probability the defendant is released $(1 - Pr(d = 1|b = h))$. If the defendant is detained, the judge incurs a cost $c(d = 1|b = h)$, which is the perceived cost to the judge of detaining someone in jail. If the defendant is released, they may commit misconduct with probability $Pr(m = 1|b = h)$. If they don't commit misconduct, the judge faces no costs. If they do, the judge faces cost $c(m = 1|b = h)$, which is the cost of misconduct, given the choice of harsh bail. In total, under the harsh bail choice, the judge incurs cost

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

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

If the judge sets lenient bail, they again face a total cost based on the probabilities and costs of detention and misconduct. However, since detention is not possible under lenient bail, only misconduct probabilities and costs remain. Therefore, under lenient bail, the judge incurs cost

$$C(b = l) = Pr(m = 1|b = l)c(m = 1|b = l).$$

What determines the probabilities and costs in the expressions for $C(b = l)$ and $C(b = h)$? I 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, $c(m = 1|b = h) = 0$, which means the expression for judge costs under harsh bail simplifies to

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

On the other hand, judges face blowback for setting lenient bail for people who commit misconduct, because the choice looks like a mistake, meaning $c(m = 1|b = l) >> 0$. Meanwhile, there is no way for anyone to assess whether harsh bail was correct when defendants are detained, because (mechanically) they cannot commit misconduct. Therefore, $Pr(d = 1|b = h) \neq 0$.

How do judges predict $Pr(m = 1|b = l)$? They have a vector of case information X and algorithm-based risk level information. The risk level information is a mapping from $Pr^A(m = 1|b = l)$ (the algorithm's prediction of misconduct under the lenient choice) to r^A. Risk levels provide relative risk information rather than absolute risk information.[6] To align with the empirical environment, I assume $r^A \in \{low, moderate, high\}$. The judge prediction is some function of observables and the algorithm's risk level: $Pr(m = 1|b = l) = f(X, r^A)$.

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

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

Adding algorithmic recommendations: Now, I complicate the status quo set-up by introducing algorithmic recommendations. I call the algorithmic recommendation R, and I define R to align with the recommendation introduced in my empirical environment, as described in Section 3. Therefore, it is based on algorithmic risk level r^A as follows:

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

In words, if the risk level is low or moderate, the recommendation to judges is to set lenient bail ($R = "b = l"$). Otherwise, there is no recommendation. How does R impact judges' decisions?

[6] 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.

Theory 1: Recommendations only impact judge predictions. 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 (because $Pr(m = 1|b = l) = f(X, r^A)$). 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.

Theory 2: Recommendations change payoffs. If recommendations change payoffs, the predictions are different. If recommendations change payoffs, then $c(m = 1|b = l)$ becomes $c(m = 1|b = l, R)$. Under this theory, it is less costly to make a mistake when that mistake is consistent with a recommendation because the recommendation lowers the level of liability. In equations, this theory means $c(m = 1|b = l, R = ''b = l'') < c(m = 1|b = l)$. In this case, judges choose to set bail based on two distinct threshold rules (one for when the recommendation applies and one for when the recommendation does not apply):

$$b = \begin{cases} \text{if } R = ''b = l'', \begin{cases} h, & \text{if } \frac{c(d=1|b=h)}{c(m=1|b=l,R=''b=l'')} < \frac{Pr(m=1|b=l)}{Pr(d=1|b=h)}, \\ l, & \text{otherwise;} \end{cases} \\ \text{if } R = -, \begin{cases} h, & \text{if } \frac{c(d=1|b=h)}{c(m=1|b=l)} < \frac{Pr(m=1|b=l)}{Pr(d=1|b=h)}, \\ l, & \text{otherwise.} \end{cases} \end{cases} \tag{3}$$

Dueling predictions: What does this theory predict happens to bail decisions after the recommendation's introduction?

- For cases where the new recommendation does not apply (high risk cases), the bail decision is determined by the rule $\frac{c(d=1|b=h)}{c(m=1|b=l)} < \frac{Pr(m=1|b=l)}{Pr(d=1|b=h)}$ in both situations. Therefore, the theory predicts bail setting rates will not change for high risk cases.
- However, for cases where the recommendation applies (low and moderate risk cases), the theory predicts a change in bail setting. Previously, without recommendations, equation 1 showed that whether judges set harsh or lenient bail depended on the expression $\frac{c(d=1|b=h)}{c(m=1|b=l)} < \frac{Pr(m=1|b=l)}{Pr(d=1|b=h)}$. Now, with the recommendation, how judges set bail depends on the expression $\frac{c(d=1|b=h)}{c(m=1|b=l,R=''b=l'')} < \frac{Pr(m=1|b=l)}{Pr(d=1|b=h)}$. The one difference between the two expressions is that with the recommendation, the denominator in the costs ratio becomes $c(m = 1|b = l, R = ''b = l'')$ instead of $c(m = 1|b = l)$. Because $c(m = 1|b = l, R = ''b = l'') < c(m = 1|b = l)$, this means the cost ratio is higher with the recommendation. In effect, fewer cases will have probabilities ratios

$\frac{Pr(m=1|b=l)}{Pr(d=1|b=h)}$ high enough to to meet the threshold for harsh bail. Therefore, harsh decisions should become less frequent and lenient decisions should become more frequent.

The theoretical model generates dueling and testable predictions for the low and moderate risk cases. 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 because the lenient decision becomes lower cost.

Anecdotal evidence on liability and algorithmic recommendations: Lenient recommendations may make lenient decisions less risky for decision-makers because they reduce liability. 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 that were 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).

There is also anecdotal evidence showing (symmetrically) that deviation from recommendations may increase liability. Suppose the recommendation is harsh bail, but a judge sets lenient bail instead. In this case, the judge sticks their neck out more than they would have in the absence of any recommendation. If a defendant commits misconduct, the judge could face higher costs through increased scrutiny and political backlash (loss of a future election).[7] This theory aligns with anecdotal evidence in a few high-stakes contexts. For example, consider the Milwaukee DA who faced calls for removal after setting low bail for a person who later committed a lethal crime. 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). In a different high-stakes setting, 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).

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

5 Kentucky Administrative Court Data

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 use the raw data to construct my final dataset using the following steps.

i. Defining the appropriate observation level: The raw data consists of many datasets at different levels of observation. My desired observation level is at the case-level. Since there can be multiple charges in a case, multiple cases in an pretrial interview, and multiple bail decisions (over time) for a cases, I take the following steps to define an interpretable and relevant level of observation. First, I aggregate data on charges up to the case level. Second, I subset to pretrial interviews with defendants where one case is 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 focus on the first bail setting for each case, commonly called initial bail.

ii. Sample restrictions: I impose several sample restrictions. First, I limit the sample to initial bail decisions made by district judges between March 18, 2011, and June 30, 2013. I make this restriction because that is the time period during which (a) the KPRA was used and (b) its calculation did not change. As such, there is no change to the calculation of risk levels during my chosen study period.

Second, I impose restrictions to eliminate concerns that HB463 changed the sample composition itself. One challenge to studying a change resulting from HB463 is that it was a large bill of about 150 pages and 110 sections (Kentucky Legislature 2011). The bill introduced more policy changes beyond introducing algorithmic recommendations to bail decision making. Therefore, a key empirical concern is incorrectly attributing estimated effects to the recommendations when they are instead due to concurrent policy changes.

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

To address any resulting change in the composition of cases, I omit cases from my sample

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

that were supposed to result in citation after HB463 according to the bill's language. Therefore, my sample of cases excludes the group that could have been simultaneously impacted by policing policy changes. I use "Standard Operating Procedures" documentation (SOP 10.1) from the Louisville Metro Police Department to identify the relevant cases based on underlying charge codes. In short, I restrict my sample to cases with offenses that were arrestable before and after HB463.

iii. Constructing risk scores: The raw administrative data do not include the underlying KPRA risk scores; they only include the risk levels. However, the administrative data do have information on all the components are used to calculate risk scores. I use observation of these components combined with Austin, Ocker, and Bhati (2010)'s explanation of the corresponding weights to construct the underlying scores. Table A.1 demonstrates the weights and the components. 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.

Final dataset: The resulting dataset consists of about 131,000 observations. Each row is an observation 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. As usual in the pretrial context, the administrative data do not indicate what specific information judges and pretrial officers discussed in each bail decision.

Motivational Descriptive Statistics: Figure A.1 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. Therefore, the new recommendations set a much lower threshold for lenient bail than implicitly existed beforehand. If the state wanted to set a threshold to align with the pre-existing level of bail setting, the lenient recommendation would have kicked in for cases with scores below 4 rather than below 14.

The chosen recommendation threshold was a normative decision on the part of the state rather than a natural consequence of any underlying risk-scoring system. Recall that many different decision thresholds are consistent with the same underlying risk rankings. In this way, algorithmic recommendations can be thought of 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 chosen threshold of 14 suggests that the state wanted judges to set lenient bail more frequently than they were doing under the status quo. Conversely, if the state had chosen a threshold of 2, that would suggest the state wanted judges to set lenient bail less frequently. Both thresholds (14 and 2) are defined based on the same underlying algorithmic 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), these descriptive statistics hint at how algorithms may change the overall composition of decisions because of explicit algorithmic recommendations.

6 Algorithmic Recommendations Change Judge Decisions

To study the effects of algorithmic recommendations, I leverage a policy change implemented in June 2011 that impacted bail decisions in Kentucky. As a result of the policy, 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 levels.

I leverage the fact that only some cases received lenient recommendations to implement both (a) differences-in-differences and (b) differences-in-discontinuities approaches. These estimated effects are the causal effect of recommendations if nothing else differentially impacted low and moderate risk cases relative to high risk cases at the time of the policy. Risk level calculation was the same before and after the policy and risk levels were available in both periods, however, the policy made it mandatory for judges to consider these algorithmic predictions. Therefore, I take additional steps to test and adjust the estimated effects to align with the desired recommendation effects.

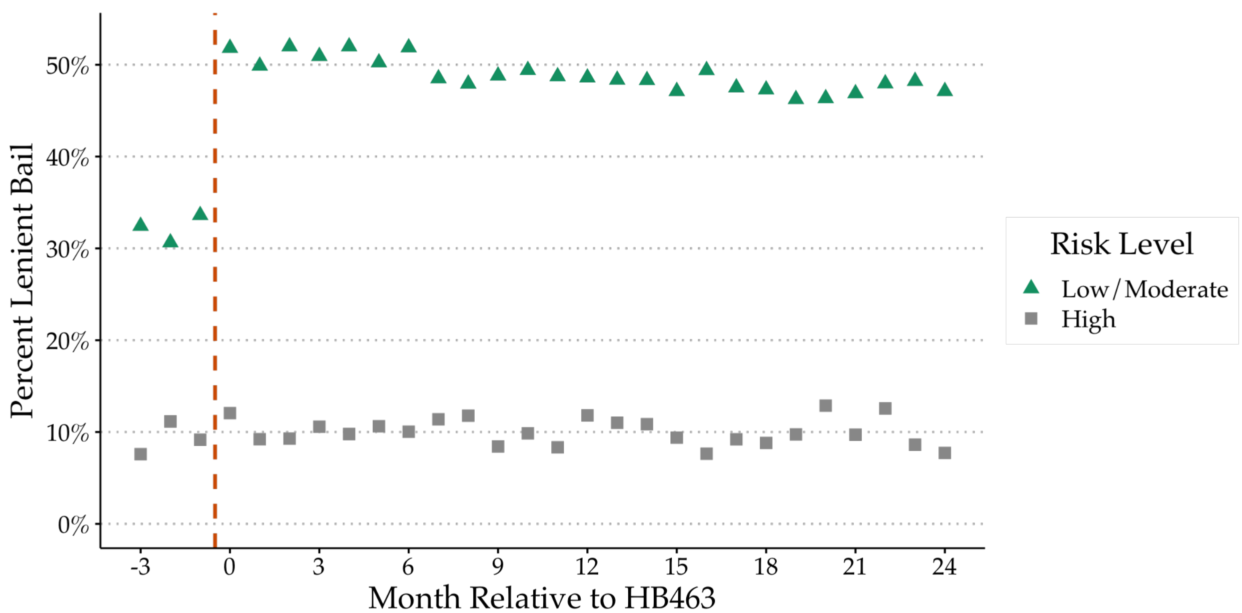
In Section 6.1, I demonstrate the straight-forward (naive) differences-in-differences and differences-in-discontinuities results. In Section 6.2, I address concerns about potential confounding related to the usage of risk levels with two different approaches. In the end, I find that the majority of the naive (unadjusted) effects are attributable to the independent causal effects of algorithmic recommendations. Lenient recommendations increase judges’ lenient bail decisions by 30-40% for marginal cases.

6.1 Naive Estimates

6.1.1 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 2 illustrates the rate of lenient bail for low or moderate risk cases and high risk cases over time.[9] 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 2 provides promising evidence for this assumption.

Figure 2: Lenient Bail Rates by Risk Level over Time



Notes: This figure shows the percent of cases receiving lenient bail over time, split 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

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

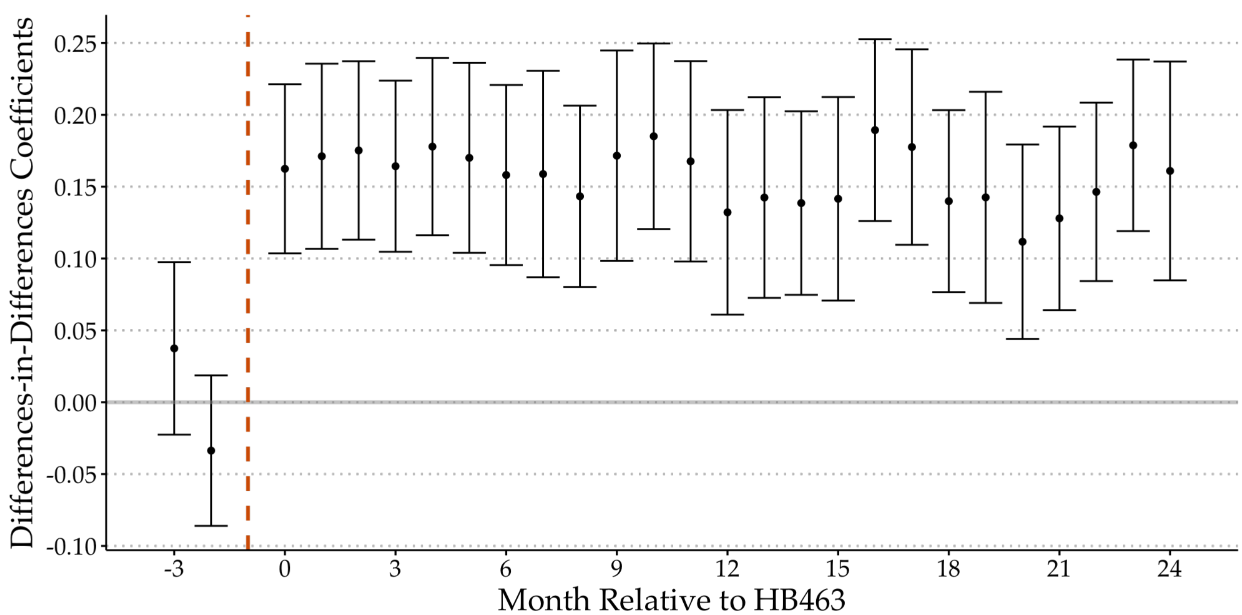
where $lenient_{itj}$ is an indicator for if the bail for case i at time t decided by judge j is lenient (no money bail) and $I(score_i < 14)$ is an indicator for if the risk score for case i is below

[9] Note that there are only a few pre-policy time periods because the method of calculating the KPRA risk score changed in March 2011, as described in Section A.1.2. To keep the meaning of risk scores and risk levels consistent, I exclude data from before March 2011 when constructing the analysis dataset (as previously detailed in Section 5).

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 3 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 B.1 shows that results with zero detailed controls are nearly identical to those with controls based on all observed case variables.

Figure 3: Dynamic Differences-in-Differences Estimates

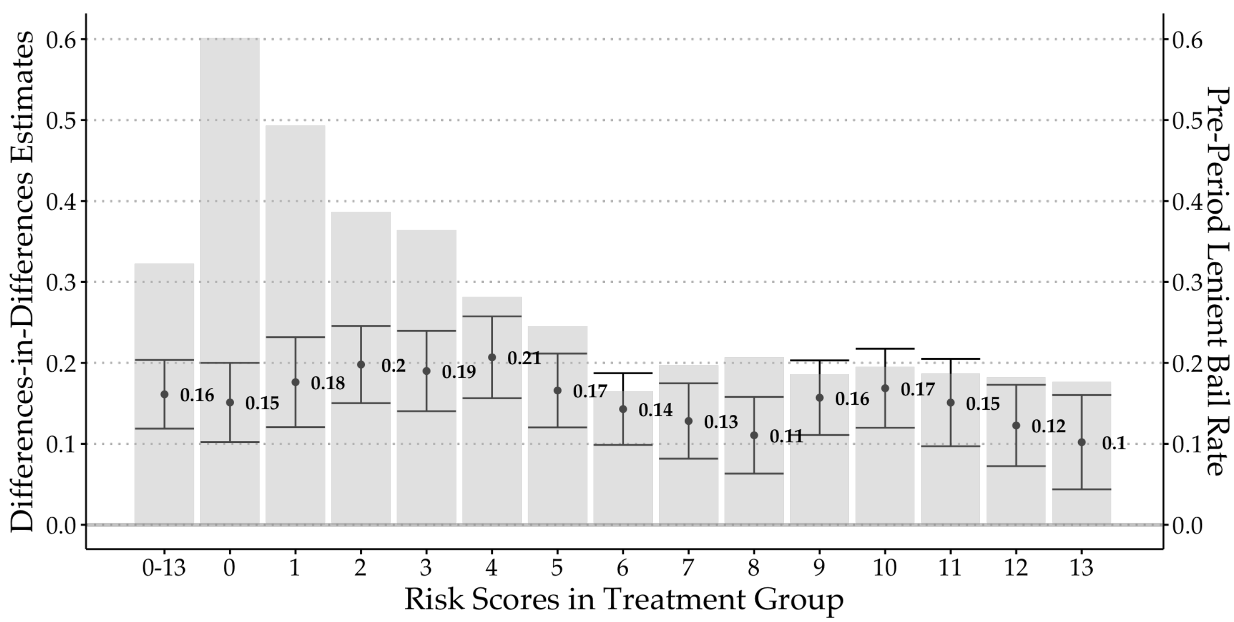


Notes: This figure shows the difference-in-differences coefficients for months relative to recommendation introduction. The outcome variable is the binary variable for lenient bail. The orange dashed line denotes the omitted period of the month before recommendation introduction.

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 find that algorithmic recommendations 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 B.2). Figure 4 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.

Figure 4: 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 outcome variable is the binary variable for lenient bail. 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.

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 B.3.

6.1.2 Differences-in-Discontinuities Results

I also estimate recommendation 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.

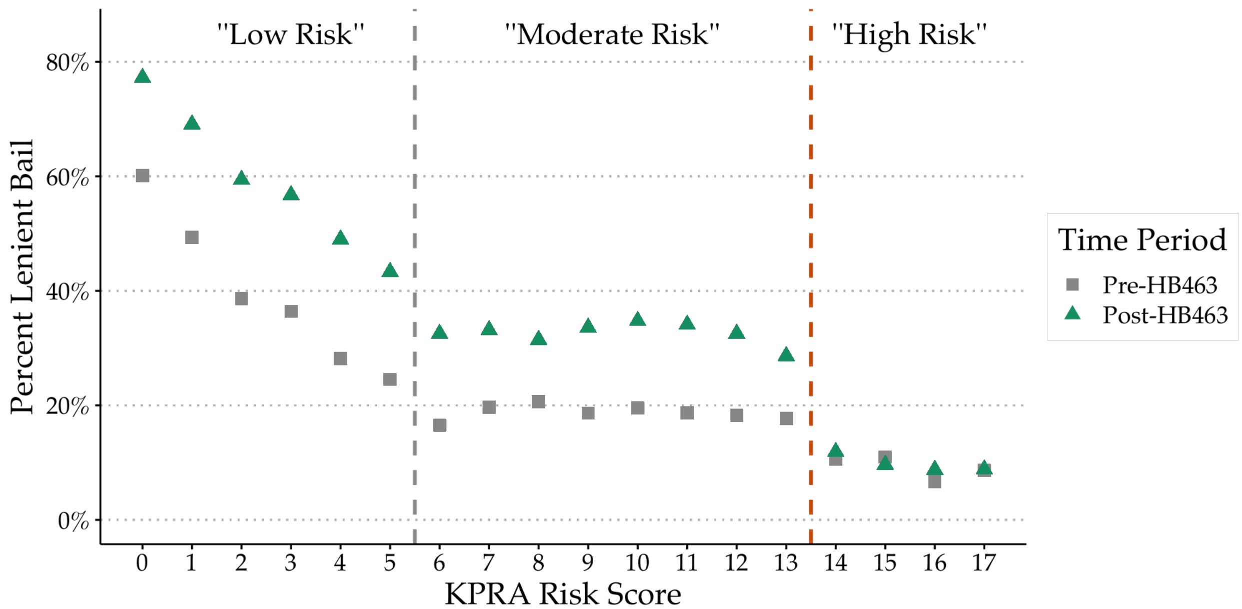
If the lenient bail recommendation were the only factor that changed discontinuously over the threshold, a simple regression discontinuity using the post-period data would identify the lenient recommendation effect. However, other relevant factors changed discontinuously at that threshold as well. Conveniently 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 5 demonstrates the differences-in-discontinuities approach visually. It 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.[10] I show lenient bail rates across 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

[10] 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 A.1 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.

Figure 5: 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.

rates are 10-20 percentage points higher in the post-period.[11] 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

[11] As an aside, Figure 5 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).

Calonico, Cattaneo, and Titiunik (2014) and Calonico, Cattaneo, and Farrell (2020) for optimal bandwidth selection as well as robust bias-corrected confidence intervals and inference procedures.[12] This approach yields a 12.7 percentage point effect, demonstrated in Figure 6. Since cases with scores of 14 receive lenient bail only 16% of the time, crossing the threshold in the post-period makes the case 80% more 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 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.2 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 12.7 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.

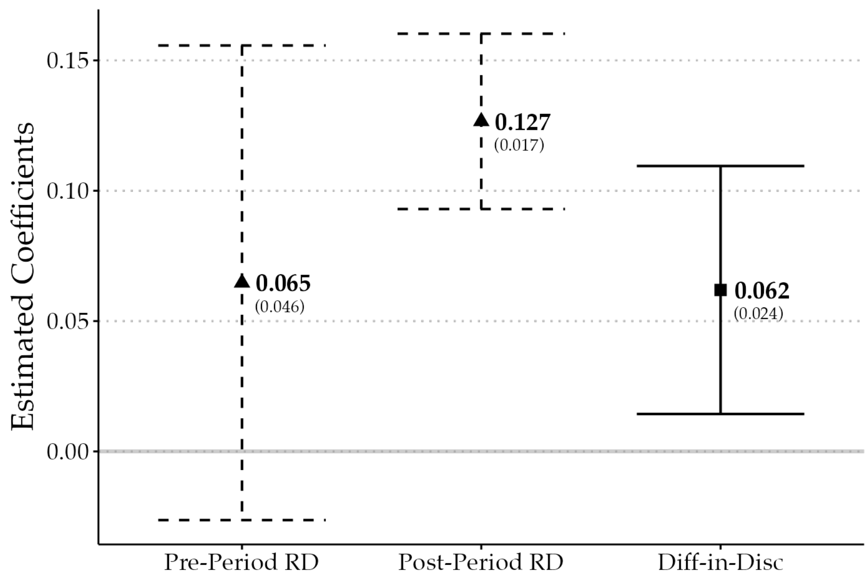
I can disentangle the recommendation effect from the other two components by leveraging the fact that I can observe bail decisions around the same discontinuity in the pre-period. In a regression discontinuity design, a central assumption is that nothing but the treatment (the presence of lenient recommendations, 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.3 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

[12] In practice, I use the `rdrobust` package developed by Calonico, Cattaneo, Farrell, and Titiunik (2023).

robust bias-corrected inference and confidence intervals. Figure 6 shows that the estimated effect in the pre-period is 6.5 percentage points. This estimate is about one half of the magnitude of the regression discontinuity in the post-period (12.7 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).

Figure 6: Regression Discontinuity and Differences-in-Discontinuities Estimates at the Moderate-High Threshold



Notes: This graph contrasts three estimation objects at the moderate-high threshold of the risk score distribution: the pre-period regression discontinuity estimate, the post-period regression discontinuity estimate, and the differences-in-discontinuities estimate. The outcome variable is the binary variable for lenient bail, and the running variable is the underlying risk scores. The two regression discontinuities are shown with triangles for the point estimates, and dotted lines for the 95% confidence intervals. These estimates use robust bias-corrected confidence intervals and inference procedures developed by Calonico, Cattaneo, and Titiunik (2014) and Calonico, Cattaneo, and Farrell (2020). The differences-in-discontinuities is shown with a rectangle for the point estimate, and regular black lines for the 95% confidence intervals. It is the difference between the two regression discontinuities.

Finally, I 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 resulting differences-in-discontinuities estimate is 6.2 percentage points, shown in Figure 6. This estimate is statistically significant and economically meaningful: the lenient recommendation led to an approximately 40% increase in lenient bail at the margin (an increase of 6.2 percentage points off a baseline of 16% for cases with scores of 14).

Overall, the differences-in-discontinuities result is consistent with the estimated differences-in-differences results across the risk score distribution shown in Figure 4. The results suggest that algorithmic recommendations have their own independent effects, consistent with the theory that recommendations change the costs of errors to human decision-makers.[13]

6.2 Testing and Adjusting the Naive Estimates

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 in Section 6.1 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 the implementation of 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.[14] In that case, HB463 changed the presence of algorithmic predictions *and* recommendations, which complicates how we interpret the previous naive results. In the following subsections, I address concerns about this potential confounding in the context of both the (a) differences-in-discontinuities and (b) differences-in-differences approaches. I find that the vast majority of the naive (unadjusted) effects are attributable to the independent causal effects of algorithmic recommendations.

6.2.1 Differences-in-Discontinuities Results

In my naive 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

[13]These 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.

[14]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.

high risk for marginal cases), and the effect of increased prior felony conviction (F):[15] $RD_{mh}^{post} = R + L_{mh} + F$.

Meanwhile, the pre-period regression discontinuity at the threshold is the sum of the levels effect at the threshold (L_{mh}) and the effect of increased prior felony conviction (F): $RD_{mh}^{pre} = L_{mh} + F$. Therefore, the difference between the two (the differences-in-discontinuities coefficient) isolates the desired recommendation effect (R):

$$\text{Diff-in-Disc}_{mh} = RD_{mh}^{post} - RD_{mh}^{pre} = R.$$

But, what if some judges do not consult the risk levels before HB463 but do after? If $\omega \in [0, 1]$ is the share of cases in which judges consult risk levels before HB463, then we can adjust the previous estimates as follows. The post-period regression discontinuity still recovers the desired recommendation effect plus the levels and increased prior felony effects: $RD_{mh}^{post*} = R + L_{mh} + F$. 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. If I assume the judges who didn't consult risk levels before HB463 respond to risk levels similarly as those who did, then I can write the pre-period regression discontinuity estimate as: $RD_{mh}^{pre*} = \omega L_{mh} + F$. Accordingly, the difference-in-discontinuity approach estimate is the sum of the desired recommendation effect and an effect that depends on the share ω and the level effect L_{mh}:

$$\text{Diff-in-Disc}_{mh}^{*} = RD_{mh}^{post*} - RD_{mh}^{pre*} = R + (1 - \omega) L_{mh}.$$

Since $\omega \geq 0$, the differences-in-discontinuities estimate from Section 6.1.2 is necessarily an upper bound for the recommendation effect. If ω is close to 1 (risk levels were consulted in almost all cases in the pre-period), then the extra term goes to 0, and the original identification strategy recovers the recommendation effect well. But, if ω is close to 0 (risk levels were consulted in almost no cases in the pre-period), then the previous strategy does not recover recommendation effects well unless L_{mh} is near 0.

I can leverage another discontinuity in the risk score distribution to estimate the share ω. Like at the medium-high threshold, cases also experience a discontinuous change in their risk level at the low-moderate threshold. However, importantly, there is no change in

[15] 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.

the presence of recommendations at the low-moderate threshold. Recommendations are either present for both groups (post-period) or absent for both groups (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: $RD_{lm}^{post*} = L_{lm}$.

If I assume that judges consult risk levels at the same rates near the low-moderate and moderate-high thresholds before HB463, then I can write the pre-period regression discontinuity as $RD_{lm}^{pre*} = \omega L_{lm}$.[16] Accordingly, the resulting differences-in-discontinuities estimate at the low-moderate threshold is:

$$\text{Diff-in-Disc}_{lm}^{*} = RD_{lm}^{post*} - RD_{lm}^{pre*} = (1-\omega)L_{lm}.$$

If the differences-in-discontinuities estimate for the low-moderate threshold is near 0, then ω is near 1 (almost all judges were consulting risk levels in the pre-period) or L_{lm} is near 0. I can directly estimate both the differences-in-discontinuities coefficient and the magnitude of L_{lm} (because $RD_{lm}^{post*} = L_{lm}$). Figure 7 demonstrates the results.

Figure 7 shows that the differences-in-discontinuities coefficient is 2.0 percentage points in magnitude and is not statistically significant, while the post-period regression discontinuity estimate is 9.9 percentage points and statistically significant. Since the differences-in-discontinuities estimate is near 0 while L_{lm} is not, this suggests that ω is close to 1 and confounding is limited.

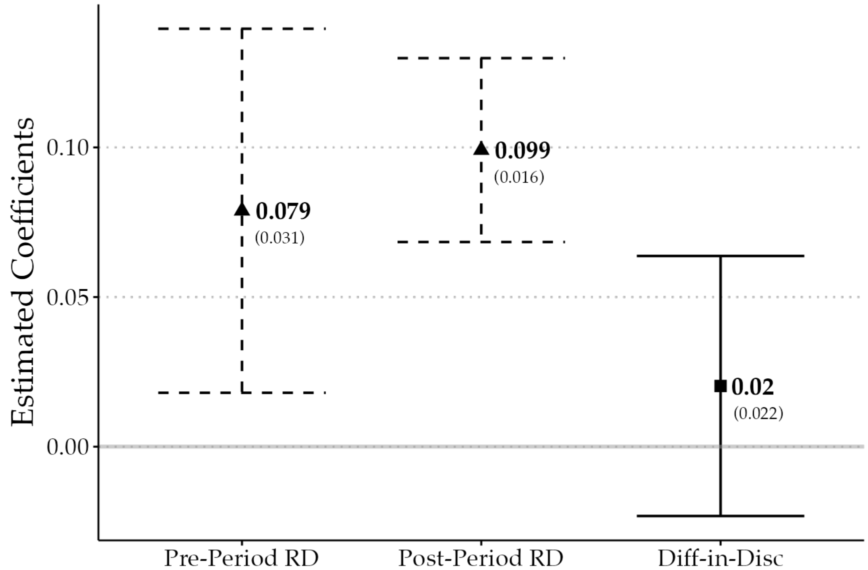
Moreover, I can use the estimates from Figure 7 to provide direct bounds on my recommendation effect estimates. Since $RD_{lm}^{pre*} = \omega L_{lm}$ and $RD_{lm}^{post*} = L_{lm}$, then it is also the case that $RD_{lm}^{pre*} = \omega RD_{lm}^{post*}$. Plugging in the coefficients from Figure 7 yields $0.079 = \omega 0.099$, which implies $\omega = 0.80$. Therefore, the empirical estimation from the low-moderate threshold implies that risk levels were consulted in about 80% of cases before HB463.

I can use this estimated ω parameter with my previous equation for the observed differences-in-differences estimate at the moderate-high threshold to bound the recommendation effect. Recall that $\text{Diff-in-Disc}_{mh}^{*} = R + (1-\omega)L_{mh}$. Plugging in the estimated differences-in-discontinuity coefficient, ω, and rearranging terms yields the following expression for R:

$$R = 0.062 - (0.20)L_{mh}$$

[16] The use of the ω parameter assumes that risk level usage is the same at the moderate-high and low-moderate thresholds, but this does not embed any assumptions about the magnitude of the level effect itself, L_{lm}.

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

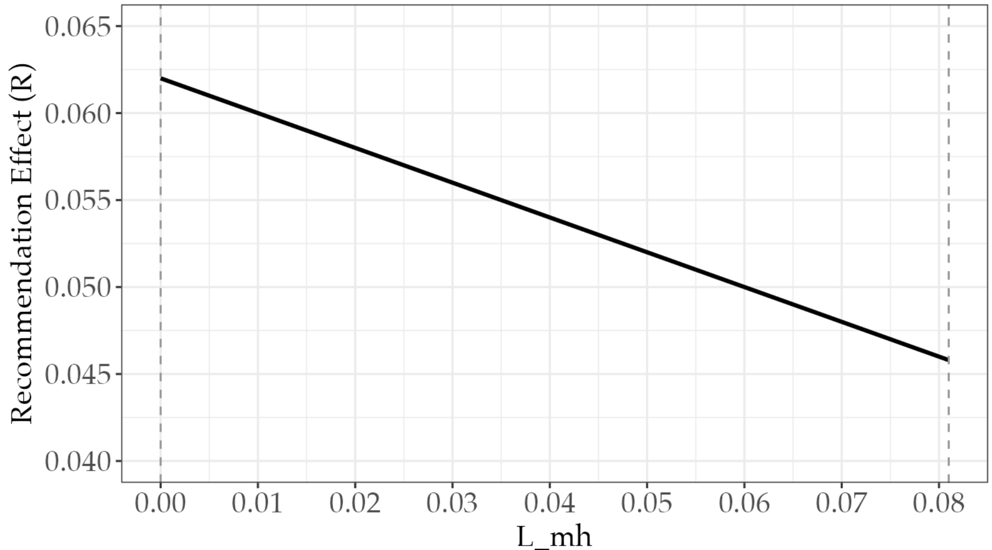


Notes: This graph contrasts three estimation objects at the low-moderate threshold of the risk score distribution: the pre-period regression discontinuity estimate, the post-period regression discontinuity estimate, and the differences-in-discontinuities estimate. The outcome variable is the binary variable for lenient bail, and the running variable is the underlying risk scores. The two regression discontinuities are shown with triangles for the point estimates, and dotted lines for the 95% confidence intervals. These estimates use robust bias-corrected confidence intervals and inference procedures developed by Calonico, Cattaneo, and Titiunik (2014) and Calonico, Cattaneo, and Farrell (2020). The differences-in-discontinuities is shown with a rectangle for the point estimate, and regular black lines for the 95% confidence intervals. It is the difference between the two regression discontinuities.

Therefore, R depends on the magnitude of L_{mh}, which I can also bound based on previous expressions and estimates. Recall that $RD^{pre*}_{mh} = \omega L_{mh} + F$. Plugging in the estimated regression discontinuity and ω yields the expression $0.065 = (0.80)L_{mh} + F$. Assuming that $F \geq 0$ and $L_{mh} \geq 0$ (since these factors should only make judges more strict), then it must be the case $L_{mh} \in [0, 0.081]$. As such, Figure 8 plots the possible values of the recommendation effect R (where $R = 0.062 - (0.20)L_{mh}$) based on the value of L_{mh} (where $L_{mh} \in [0, 0.081]$).

Figure 8 implies the algorithmic recommendation effect must lie between 4.6 and 6.2 percentage points. These magnitudes mean that the algorithmic recommendation increased lenient decisions by 30-40%. This bounding exercise therefore demonstrates that the recommendation's causal effects are still large and economically meaningful even after carefully accounting for confounding.

Figure 8: Bounds for the Isolated Recommendation Effect



Notes: This figure demonstrates the bounds for the isolated recommendation effect, R, as a function of another unknown parameter L_{mh} (the effect of the risk level labeling). The black line shows the relationship between R and L_{mh} as the line defined by the equation $R = 0.062 - (0.20)L_{mh}$, where $L_{mh} \in [0, 0.081]$.

6.2.2 Differences-in-Differences Results

I also address potential confounding in the differences-in-differences setting. If $\omega \in [0, 1]$ is the share of cases in which risk levels were consulted before HB463, then the differences-in-differences estimation approach picks up the effect of newly using risk levels for $(1 - \omega)$ of the cases. As a result, the pooled differences-in-difference coefficient (β^{DD}) is then a weighted average between my desired recommendation effect (R') and an effect of levels (the effect of the judge hearing the case's risk level L).[17] 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, then $\beta^{DD} \approx R'$, and the straightforward differences-in-differences strategy recovers the recommendation effect well. If ω is close to 0, then $\beta^{DD} \approx R' + L$, and the previous strategy does not recover recommendation effects *unless* $L \approx 0$.[18]

Since confounding should be minimal when $L \approx 0$, I test whether recommendation effects matter in a subset of cases where I expect the risk levels effects to be very small. Specifically,

[17] 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.

[18] 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).

I focus on cases for which risk levels do not provide new prediction information to the judges because they are very 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 zero risk factors (zero failures to appear, zero pending cases, zero 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 obvious 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 particular group of defendants.

Misdemeanor cases with zero risk factors are 7% of cases in the data. Figure A.4 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 6.1.1. Figure A.5 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 B.4 shows that results with no detailed controls are nearly identical to those with controls based on all observed case variables.

Table 1 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, the relative effects are smaller because the baseline lenient bail rates are higher for this low 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 is necessarily minimal.

Therefore, algorithmic recommendation effects survive concerns about confounding variation when using both mathematical bounding approaches and intuitive subsetting approaches. Both methods demonstrate that algorithmic recommendations have causal effects that are independent of algorithmic predictions.

Table 1: 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. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

7 Conclusion

This paper studies how predictive algorithms impact human decisions. Conventional wisdom assumes that algorithms affect human decisions by providing people with data-driven predictions. In this paper, I demonstrate that algorithms matter in another way. Algorithms often provide decision-makers with explicit recommendations, and these algorithmic recommendations have independent effects on human decisions.

I demonstrate the importance of algorithmic recommendations by isolating their causal effects on human decisions. I use a unique setting in the U.S. criminal justice system paired with administrative data to demonstrate that algorithmic recommendations have first-order effects on human decisions. In my setting, lenient recommendations increase judges' lenient bail decisions by 30-40% for marginal cases.

These economically meaningful effects are not attributable to changes in algorithmic predictions. Instead, the evidence is consistent with recommendations changing human decisions because they change the incentives of making different decisions. Recommendations may change decisions because they change the cost of errors – liability is lower when adhering to recommendations, but higher when deviating from recommendations.

In this way, recommendations can change more than just the allocation of decisions – they can change the overall composition of decisions. If decision-makers and algorithm designers disagree about the costs of errors, then recommendations may better align decision-maker incentives with social planner objectives (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.

These results help inform policy discussions related to algorithms and human decisions. For instance, when Obermeyer et al. (2019) found that a healthcare algorithm was racially biased, the company that developed the algorithm 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" (Johnson 2019). In other words, because doctors consider more than just algorithmic recommendations when making decisions, the company argued their algorithmic recommendations don't necessarily impact final outcomes. My results demonstrate that algorithmic recommendations do have strong causal effects on human decisions even when many other pieces of data are available. Therefore, algorithmic recommendations (like this referral recommendation) merit independent study and scrutiny because they have meaningful demonstrable effects on high-stakes decisions.

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.

Almog, David, Gauriot, Romain, Page, Lionel, and Martin, Daniel. 2024. "AI oversight and human mistakes: Evidence from Centre Court" *Research Paper, Northwestern University*.

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, Max Farrell, and Rocio Titiunik. 2023. "`rdrobust`: Robust Data-Driven Statistical Inference in Regression-Discontinuity Designs." https://cran.r-project.org/web/packages/rdrobust/rdrobust.pdf.

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.

County of Santa Clara, Office of Supportive Housing. 2025. "Coordinated Entry - Get Connected to Housing Support." https://osh.santaclaracounty.gov/housing-and-shelter-support/shelter-and-housing-support/housing-support/coordinated-entry-get.

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 Univer-*

sity.

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*.

Hausman, David. 2024. "Risk Assessment as Policy in Immigration Detention Decisions." *Working Paper, University of Berkeley School of Law*.

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

Johnson, Carla. 2019. "Racial bias in health care software aids whites over blacks." *Seattle Times*. https://www.seattletimes.com/seattle-news/health/racial-bias-in-health-care-software-aids-whites-over-blacks/.

Kentucky Legislature. *House Bill 463: An Act Relating to Public Safety and Making an Appropriation Therefor, and Declaring an Emergency.* 2011. Available at: https://apps.legislature.ky.gov/record/11rs/hb463.html. Accessed: September 4, 2024.

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*.

Mingle, Katie. 2020. "Chapter 4: The List." 99% Invisible Podcast. Transcript available at: https://99percentinvisible.org/episode/according-to-need-chapter-4-the-list/transcript.

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 Jennifer Doleac. 2023. "The Counterintuitive Consequences Of Sex Offender Risk Assessments At Sentencing." *University of Toronto Law Journal* 73 (1): 59–72.

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.

U.S. Department of Housing and Urban Development. 2017. "Notice Establishing Additional Requirements for a Continuum of Care Centralized or Coordinated Assessment System." Notice CPD-17-01. Washington, DC: U.S. Department of Housing and Urban Development. Available at: https://www.hud.gov/sites/documents/17-01cpdn.pdf.

Appendix

A.1 Background on Kentucky Bail Setting

A.1.1 Kentucky Compared to other US Settings

The Kentucky setting has a several features that distinguish it from other US bail settings. First, in many other states, pretrial data is managed locally, meaning that data needs to be collected at the county-level. However, Kentucky has one pretrial services agency serving all of its 120 counties, so I am able to use data from the entire state. Second, 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.[19] Because judges make bail decisions over the phone, defendants are not present. Third, police have full authority to charge in Kentucky, which means there is no prosecutorial review before the judge makes a bail decision. Thus, judges' bail decisions do not follow any prosecutor's actions. Finally, Kentucky does not have a commercial bail bonds industry – it is one of four states with this ban as of 2022 (Cornell Law School Legal Information Institute 2024). This means that if someone cannot afford money bail in Kentucky, they cannot contract with a bail bonds agent to make bail.[20]

A.1.2 Background on Kentucky Risk Assessment

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.

[19] 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.

[20] However, 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.

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."

A.1.3 Information Used in Bail Decisions

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.)

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).

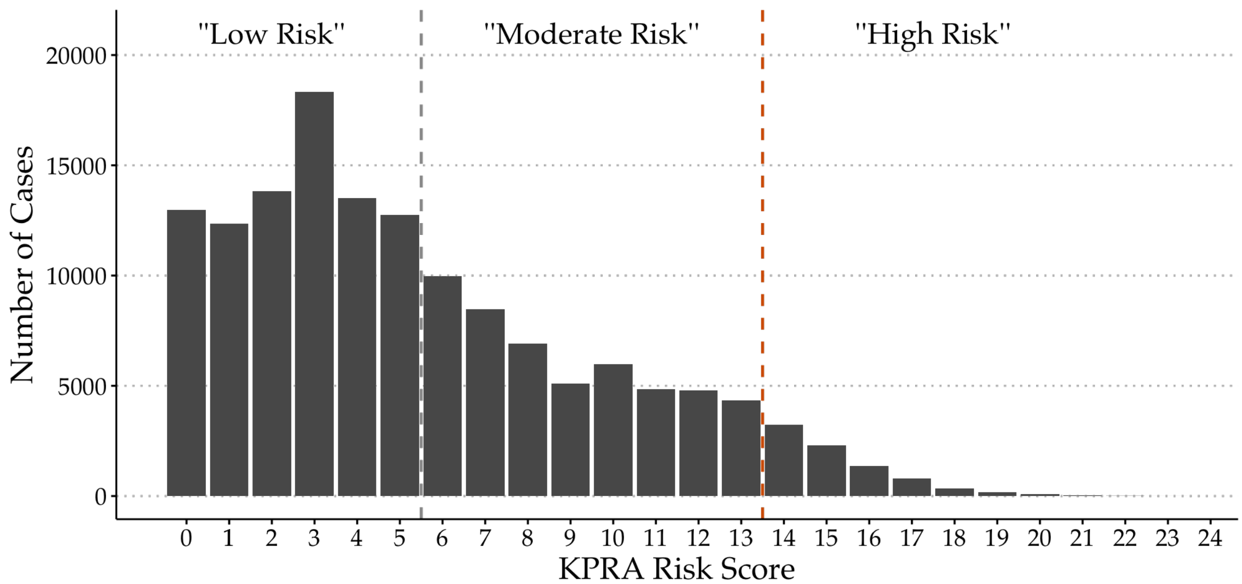
Table A.2: Kentucky Pretrial Risk Assessment Factors (Before March 18, 2011)

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).

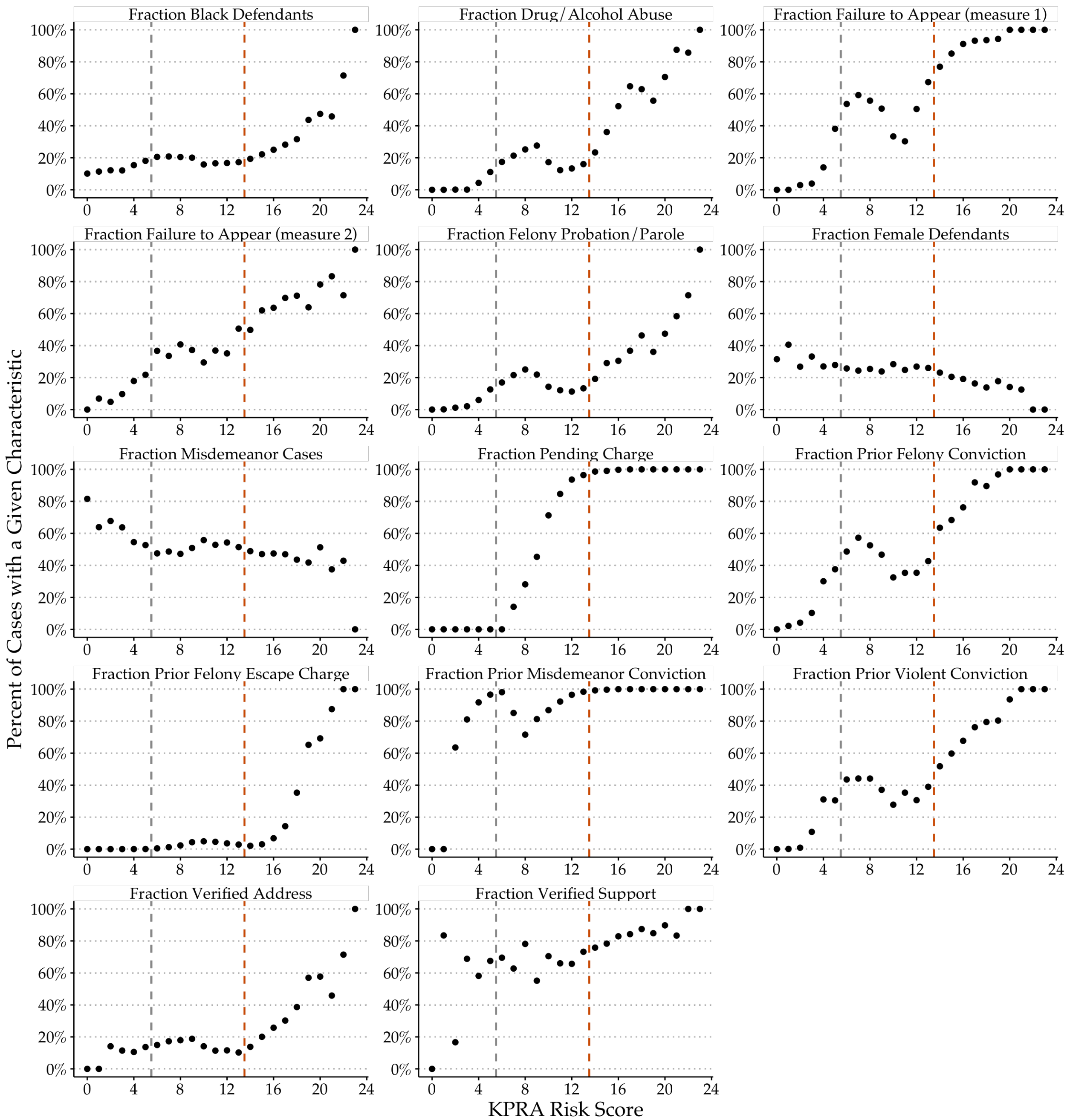
A.2 Supplementary Figures and Tables

Figure A.1: The Risk Score Distribution



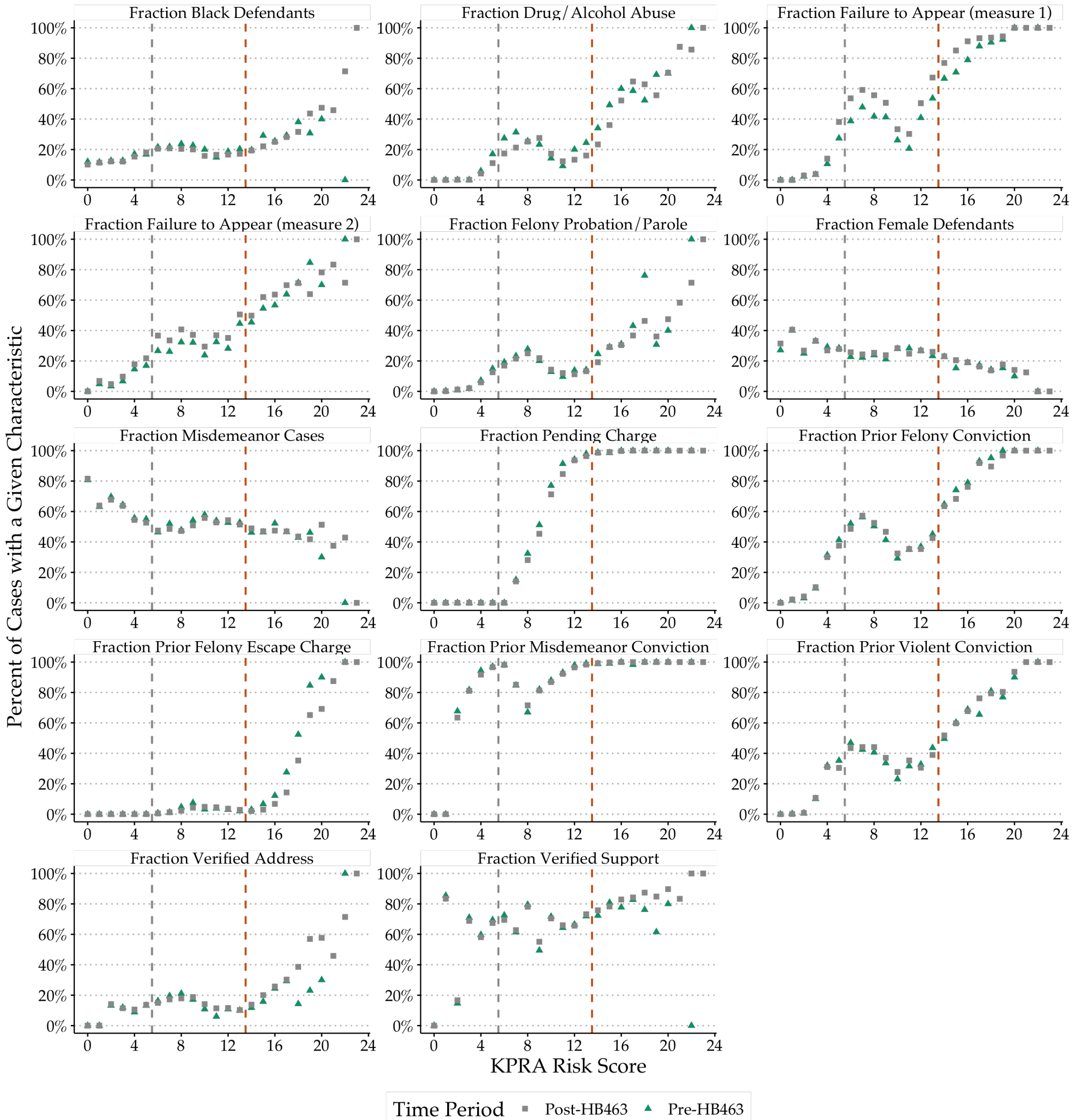
Notes: This histogram demonstrates the number of cases across the full risk score distribution. The dashed lines indicate the cut-offs between risk levels. The orange line is the threshold between "moderate" and "high" risk, and the gray line is the threshold between "low" and "moderate" risk. Scores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

Figure A.2: 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.3: Defendant and Case Covariates over Risk Score Distribution and Time Periods



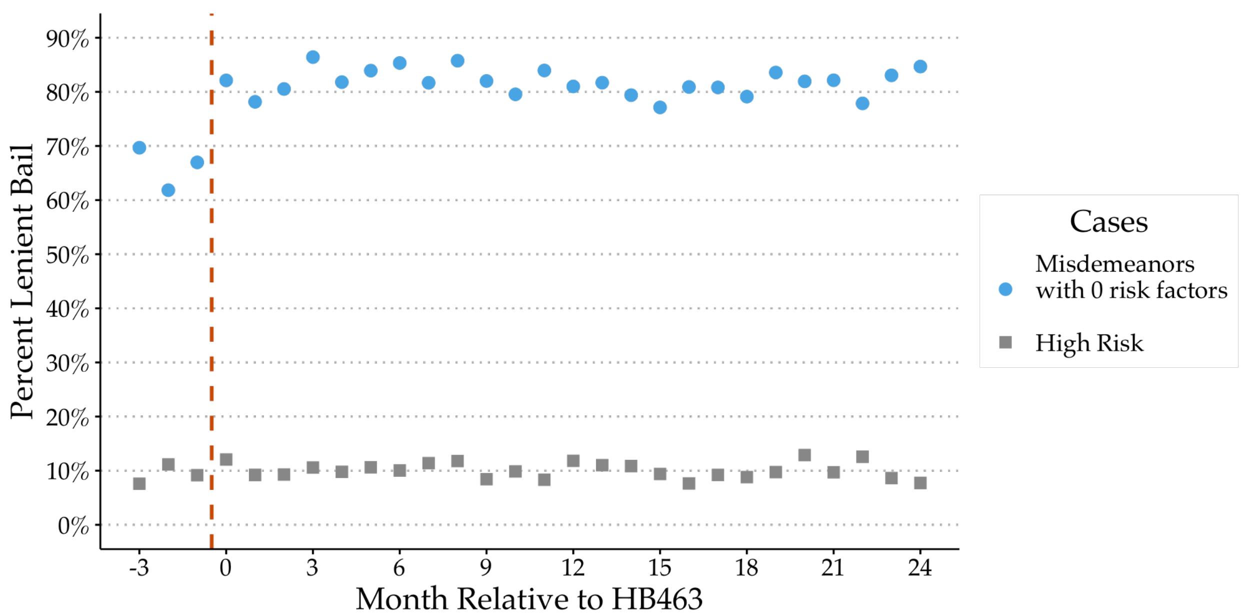
Notes: This figure shows each discrete case risk score's average defendant and case covariates. The gray rectangles show the averages before HB463, while the green triangles show the averages after HB463. The dashed lines indicate the cut-offs between risk levels. The orange line is the threshold between "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.

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		

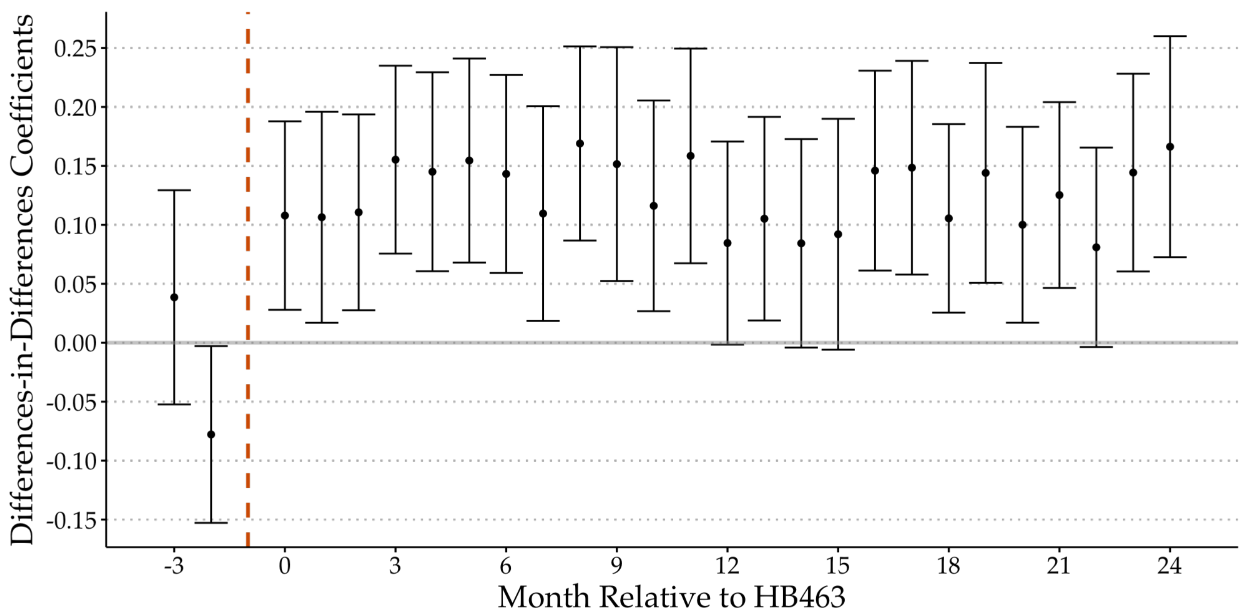
Notes: This table displays estimated differences-in-difference coefficients in specifications with lenient bail as the dependent variable. The control group consists of cases with high risk levels, and the treated group consists of cases with low or moderate risk levels. The table shows results across 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.

Figure A.4: 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.

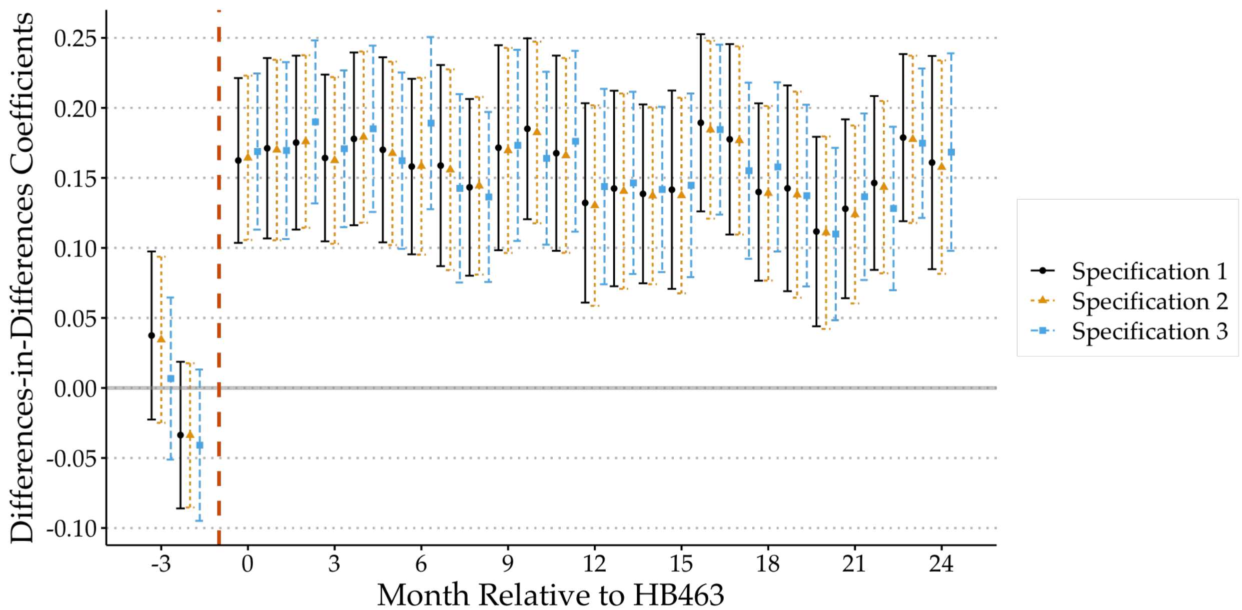
Figure A.5: Dynamic Differences-in-Differences Estimates (Treated Group: Lowest Risk Cases)



Notes: This figure shows the difference-in-differences coefficients for months relative to the recommendation introduction. The outcome variable is the binary variable for lenient bail. 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.

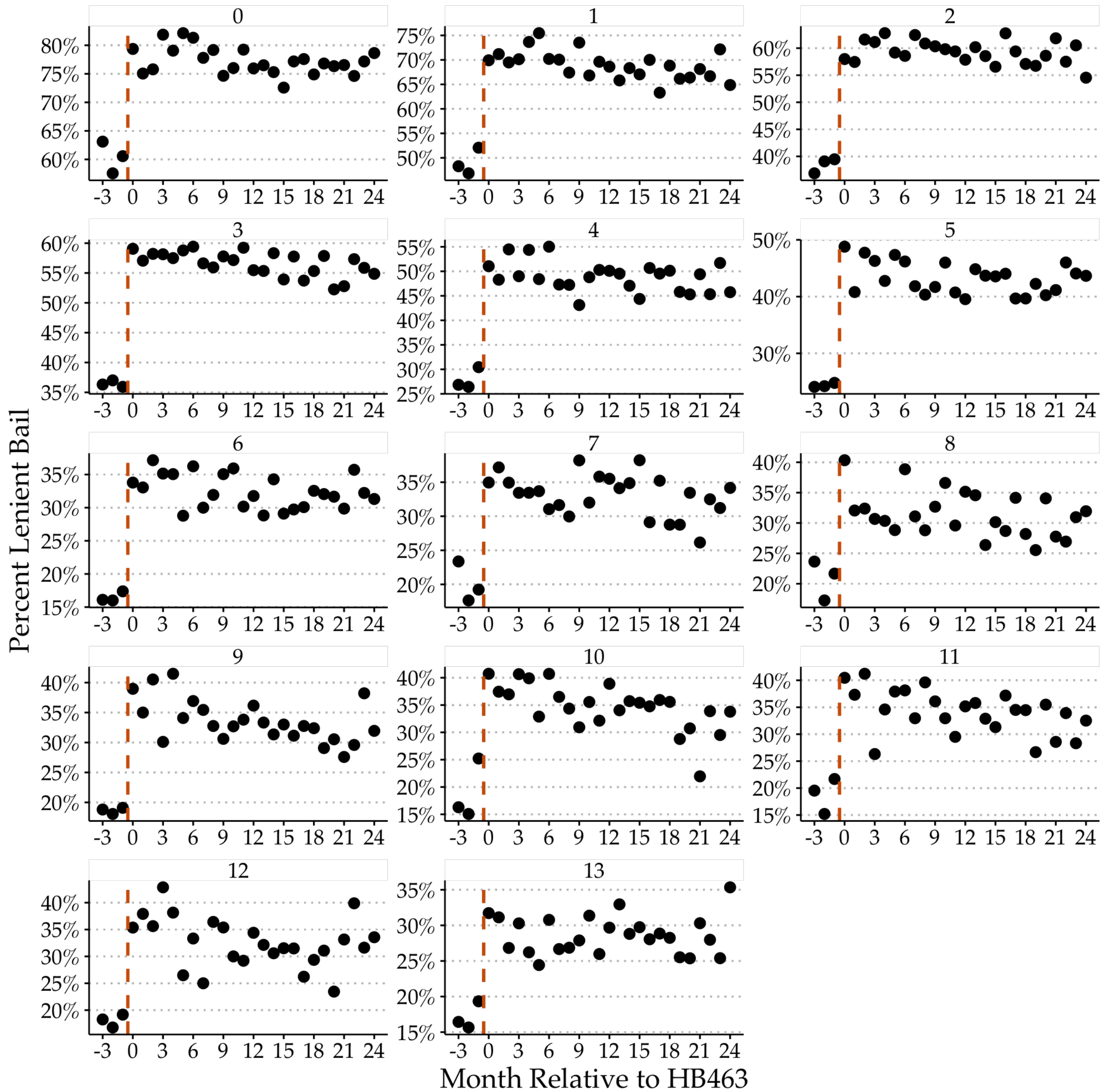
Online Appendix

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



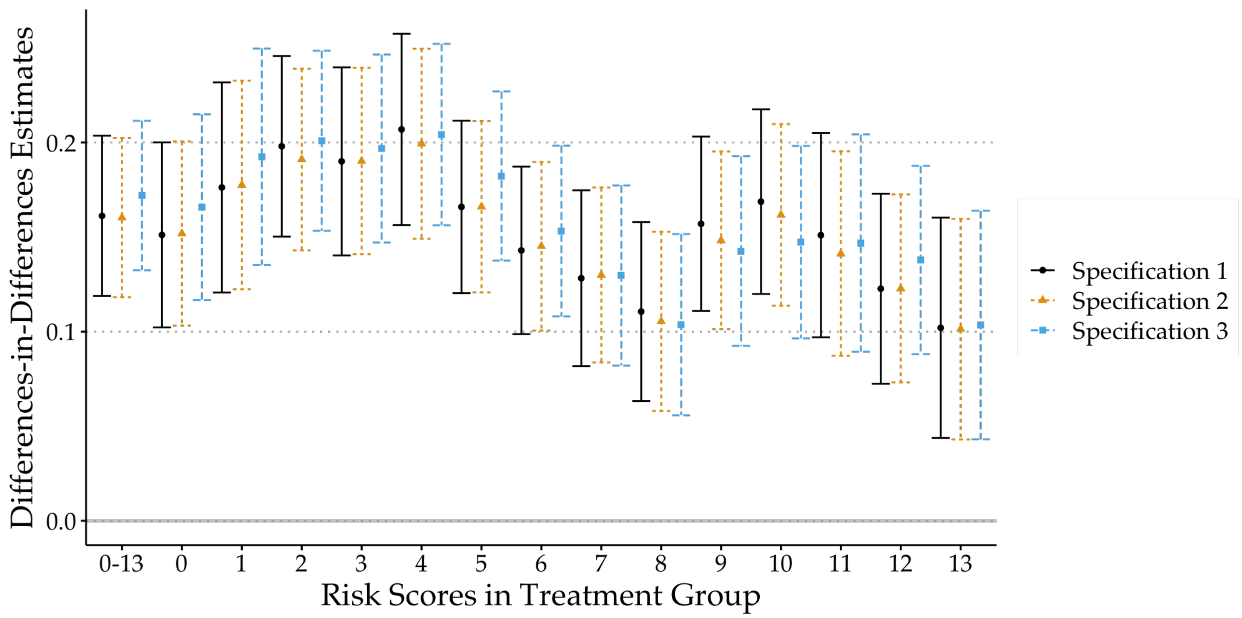
Notes: This figure shows the difference-in-differences coefficients for months relative to recommendation introduction across specifications. The outcome variable is the binary variable for lenient bail. The control group consists of cases with high risk scores, and the treated group consists of cases 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 3, 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 B.2: Lenient Bail Rates by Risk Score over Time



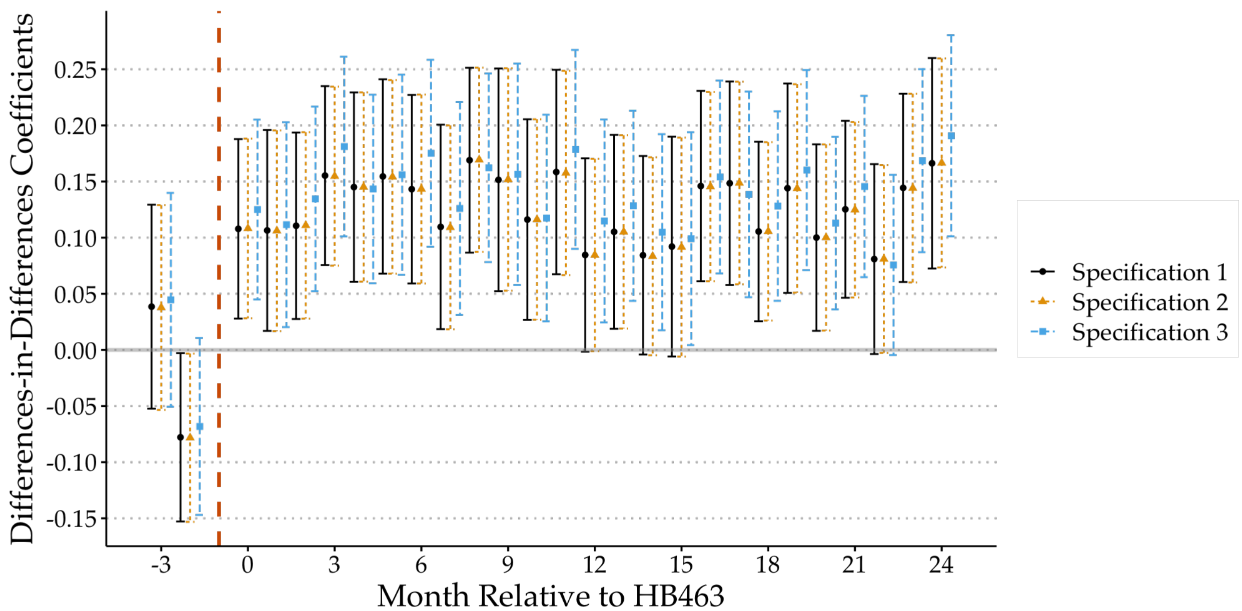
Notes: This figure shows the rate of lenient bail over months by risk scores for low and moderate risk cases. Months are indexed relative to the introduction of algorithmic recommendations. The orange dotted line shows when HB463 went into effect. Each plot is for a different discrete value of risk score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

Figure B.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 outcome variable is the binary variable for lenient bail. 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.

Figure B.4: 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 outcome variable is the binary variable for lenient bail. 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 3, 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.