The Hidden Effects of Algorithmic Recommendations

Alex Albright*

March 2026

Abstract

Algorithms provide human decision-makers with data-driven predictions, but they can also provide explicit recommendations. I demonstrate that 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 changing the cost of certain decisions. In this way, algorithms can affect human decisions through preferences as well as predictions.

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

I am grateful to Larry Katz, Winnie Van Dijk, Ed Glaeser, Megan Stevenson, Jennifer Doleac, Abbie Wozniak, and Andrew Goodman-Bacon for valuable feedback at key stages of this project. I also thank Alicia Modestino, Elior Cohen, Mike Mueller-Smith, Crystal Yang, Alma Cohen, Mandy Pallais, Louis Kaplow, Adam Soliman, as well as seminar and conference audiences at Harvard, LSU, Macalester, Williams, St. Olaf, the Minneapolis Fed, Clemson, the University of Minnesota, FGV EESP, Insper, PUC-Rio, SDSU, Northwestern, WEC Jr. (Northeastern), Wiser (Kansas City Fed), SOLE, and ASSAs for helpful comments. I am grateful to Daniel Sturtevant, Tara Blair, Christy May, and Kathy Schiflett for providing access to the data and for sharing institutional details about Kentucky Pretrial Services. I also thank James Holt for editorial support, Sara Brandel for assistance with data agreements, and Amisha Kambath for research assistance. This research was supported by a Stone PhD Fellowship from the Harvard Inequality & Social Policy Program (2018-2022) and a Considine Fellowship from the Olin Center at Harvard Law School (2018-2021). 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 mortgage lending (Bhutta, Hizmo and Ringo, 2025), algorithms predicting strokes are used in medical care (Abaluck et al., 2021), and algorithms predicting flight risk are used in immigration courts (Hausman, 2025).[1] Despite their prevalence, it is still the norm that humans (loan officers, doctors, immigration officers) – not algorithms – make the final decisions. Therefore, understanding how algorithms affect these systems requires understanding how algorithms affect human decision-making.

Most seminal work on this topic focuses on algorithms as providers of data-driven predictions (Kleinberg et al., 2018; Agrawal, Gans and Goldfarb, 2018; Mullainathan, 2025). However, many algorithms also issue explicit recommendations: underwriting algorithms recommend loan approval (Bhutta, Hizmo and Ringo, 2025), stroke risk algorithms recommend medical treatment (Abaluck et al., 2021), and immigration court algorithms recommend detention (Hausman, 2025). 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 this distinction, predictions and recommendations are usually conflated under the catch-all term "algorithm." As a result, the effects of "algorithms" can muddle the effects of predictions with the effects of recommendations. This paper demonstrates why this matters: algorithmic recommendations have independent effects on human decisions.

Isolating the effects of algorithmic recommendations is empirically challenging. Institutional details about how algorithms are developed and implemented are often opaque, and algorithmic predictions and recommendations are frequently introduced at the same time. I make progress by leveraging a unique setting in which algorithm design is transparent and algorithmic predictions already exist, but recommendations are introduced later.

My empirical setting covers bail decisions in Kentucky from 2011 to 2013. Bail decisions are high-stakes in the US criminal justice system: they set the conditions for defendants' release from jail after arrest. Judges can set money bail, which requires defendants to post money to secure release. This is consequential because money bail increases pretrial

[1] These examples are illustrative rather than exceptional. A 2023 report documents over 600 state government contracts for algorithmic decision systems across multiple policy domains (Fergusson, 2023). By the early 2020s, predictive tools were used by most US hospitals (Chang et al., 2025), child welfare agencies in at least 11 states (Loudenback, 2022), and nearly half of US counties for pretrial decisions (Lattimore et al., 2020).

detention (Albright, 2025), which in turn has been shown to increase conviction rates (Leslie and Pope, 2017), reduce formal employment (Dobbie, Goldin and Yang, 2018), and increase household insolvency (Slutzky and Xu, 2025).

During the study period, Kentucky bail judges received information about the case and defendant and had access to an algorithmic prediction of pretrial misconduct. In June 2011, a new policy, motivated by a desire to reduce incarceration costs, introduced algorithmic recommendations: cases with low or moderate predicted risk were now recommended for release without money bail ("lenient bail" recommendations). This policy change provides useful variation for causal inference: lenient recommendations applied only below the high risk cutoff and were introduced abruptly in June 2011.

Why might these recommendations change decision-maker behavior? The answer turns on whether recommendations matter independent of the predictive information they accompany. If recommendations operate only by communicating risk information, then the new recommendations should have no effect because judges already had access to the underlying risk predictions. But recommendations may do more than convey information: they can change the costs of different choices – for example, by reducing perceived accountability or lowering cognitive burden when choosing lenient bail. In that case, lenient bail should increase for the low risk and moderate risk cases that newly received a lenient recommendation.

To test this, I estimate the causal effects of algorithmic recommendations using 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 sharp risk score cutoff for lenient recommendations. After June 2011, cases with the highest moderate risk scores received lenient recommendations while similarly scoring cases with the lowest high risk scores did not. A simple regression discontinuity would identify the lenient recommendation effect if it were the only factor changing discontinuously at the threshold post-June 2011. However, other factors also changed discontinuously there. Since these confounding discontinuities existed pre-treatment as well, I use differences-in-discontinuities to recover the lenient recommendation effect. Unlike differences-in-

differences, this approach estimates effects for marginal cases near the moderate-high threshold.

I find that the 2011 policy change increased lenient decisions by about 55% 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. For the marginal moderate risk cases, the differences-in-discontinuities estimate is also large, at approximately 40%. These results suggest that algorithmic recommendations meaningfully change human decisions.

However, more empirical work is needed to isolate the causal effects of algorithmic recommendations. Although the calculation of risk was the same before and after the policy and risk levels were available in both periods, the policy also mandated that judges consider algorithmic predictions. If some judges ignored predictions beforehand, then the policy changed both the presence of algorithmic predictions *and* recommendations. Therefore, my initial differences-in-differences and differences-in-discontinuities results are necessarily an upper bound on the recommendation effect of interest.

I use two strategies to isolate the desired recommendation effect. First, I estimate the pre-policy use of risk levels by comparing regression discontinuities at the low-moderate threshold before and after the policy. Because risk levels change discontinuously at this threshold but recommendations do not, this approach allows me to infer how often judges consulted risk levels in the pre-period. I estimate that judges used risk levels in about 80% of cases. Adjusting my baseline estimates with this rate implies that lenient recommendations increased lenient bail decisions by 30-40% for marginal cases. These bounds rely on two assumptions, but intuitive violations would bias the implied recommendation effect upward, rendering them conservative.

Second, I examine a subset of cases for which the risk levels provide little new information: misdemeanor cases with zero risk factors. Intuitively, the risk level does not provide new information to judges for these cases because they always obviously have low risk scores (they have zero associated failures to appear, pending cases, convictions, etc.). Therefore, for these lowest-risk-scoring cases, the differences-in-differences estimate should capture recommendation effects alone. Consistent with meaningful recommendation effects, I find recommendations increase lenient bail by about 15 percentage points (or about 23%) for this group.

Taken together, the results show algorithmic recommendations have large, economically meaningful effects on human decisions – effects that persist after isolating them from

changes to predictive information. The results imply that recommendations operate by changing the costs of decisions: they may change perceived accountability or reduce cognitive load by providing a salient default. In this way, predictive algorithms can affect human decisions by updating preferences as well as predictions.

My paper contributes to the economics literature on algorithms and human decision-making. Early work typically contrasted human decisions without algorithms with hypothetical decisions made entirely by algorithms (Kleinberg et al., 2018; Jung et al., 2017; Berk, Sorenson and Barnes, 2016; Goel, Rao and Shroff, 2016), a comparison that rarely matches real policy environments where humans retain discretion. More recent studies examine settings in which humans use algorithms at their discretion (Hoffman, Kahn and Li, 2018; Stevenson, 2018; Gruber et al., 2020; Agarwal et al., 2023; Stevenson and Doleac, 2024; Grimon and Mills, 2025; Sloan, Naufal and Caspers, 2025). These empirical settings differ in whether algorithms supply predictions alone or also provide explicit recommendations. This distinction is policy-relevant because, as I show, algorithmic recommendations – distinct from the underlying predictions – independently affect human decisions.

This paper is among the first to estimate the causal effects of algorithmic recommendations. In doing so, I complement McLaughlin and Spiess (2025), who develop a model in which recommendations can directly shift preferences rather than merely update beliefs. My results provide empirical evidence of the independent effects they model.[2] I also complement Hausman (2025), who studies a change in algorithmic recommendations used by ICE for immigration detention and finds release rates fell by 50% when release recommendations were removed. However, because the ICE algorithm's predictions changed simultaneously, his design cannot separate prediction from recommendation effects. My paper advances the evidence by isolating the causal impact of recommendations.[3]

My findings also relate to the literature on defaults, framing, and multiphase decision-making. Research shows that defaults and recommendations influence decisions in contexts ranging from retirement savings (Madrian and Shea, 2001) to organ donation (Johnson and Goldstein, 2003). Closest to my context, Bushway, Owens and Piehl (2012) show that

[2]Several papers note that algorithms can shift decision-maker incentives (Stevenson and Doleac, 2024; Cowgill and Stevenson, 2020; Davenport, 2023), but McLaughlin and Spiess (2025) are unique in focusing on how algorithmic recommendations may do so. Relatedly, Almog et al. (2025) show that AI oversight increases the perceived cost of errors when human decisions can be publicly overruled.

[3]To study the effects of algorithmic recommendations, I leverage a 2011 Kentucky policy change first examined by Stevenson (2018). While she evaluated the policy's overall impact, I use the setting to isolate the effect of recommendations specifically, rather than the full policy bundle. This requires several novel steps: identifying a period when algorithmic predictions remained fixed while recommendations changed, reconstructing underlying risk scores from court records, and using data-driven strategies to address confounding changes in arrests and risk-level visibility.

sentencing guidelines have causal effects on judicial decisions, holding other case characteristics constant. Many high-stakes settings mirror the justice system in that they involve sequences of decisions rather than a single choice in isolation; my results show how upstream changes can affect downstream discretion in multiphase systems (Baron et al., 2024; Bohren, Hull and Imas, 2025).

The remainder of the paper proceeds as follows. Section 2 describes the Kentucky bail setting, the pretrial risk algorithm, and the introduction of algorithmic recommendations, and develops testable predictions. Section 3 describes the administrative court data. Section 4 presents the empirical results: I first document the effects of algorithmic recommendations using differences-in-differences and differences-in-discontinuities designs, and then develop additional tests and bounds to isolate recommendation effects from changes in the use of algorithmic predictions. Section 5 concludes.

2 Empirical Setting: Bail Decisions in Kentucky

2.1 Background on Bail Decisions and Pretrial Algorithms

I study how algorithmic recommendations impact bail decisions. Bail decisions are high-stakes decisions in the US criminal justice system that set the conditions for defendants' release from jail after arrest. Judges can set money bail, which requires defendants to post money to secure release. Money bail matters because it increases the likelihood of pretrial detention (Albright, 2025), which in turn can increase conviction rates (Leslie and Pope, 2017), reduce formal employment (Dobbie, Goldin and Yang, 2018), and increase household insolvency (Slutzky and Xu, 2025). Pretrial detention is common in the US, with roughly 470,000 unconvicted people held in jail each day (Zeng, 2023).

Bail decisions are made quickly – often within minutes – and the legal objective is well defined (Arnold, Dobbie and Yang, 2018). The legal objective is to set the least restrictive bail conditions needed to ensure court appearance and public safety (American Bar Association Criminal Justice Standards Committee, 2007). To assist with these decisions, many jurisdictions use algorithms designed to predict the risk of pretrial misconduct (failure to appear in court or rearrest) (Desmarais and Lowder, 2019).

Pretrial algorithms generate misconduct-risk predictions from individual and case characteristics, but jurisdictions vary widely in which algorithms they use. More than two dozen pretrial algorithms are used across the United States (Desmarais and Lowder, 2019). They differ in their inputs, the data used to train them, and whether they are transparent or

proprietary (Desmarais and Lowder, 2019).

The policy relevance of algorithmic recommendations is clear in the case of the Public Safety Assessment (PSA), the most widely used pretrial algorithm in the United States. The PSA operates in hundreds of localities, covering about 84 million people – nearly one-quarter of the US population (Advancing Pretrial Policy & Research, 2025*b*).[4] Although the PSA's underlying prediction model is identical across places, jurisdictions adopt different "Release Conditions Matrices" – recommendations that map PSA scores to specific pretrial conditions (Advancing Pretrial Policy & Research, 2025*a*). As a result, algorithmic predictions can be identical across localities but yield distinct algorithmic recommendations. This institutional feature illustrates a key point of my paper: algorithmic recommendations constitute a distinct design layer built on top of predictions.

2.2 Kentucky Institutional Details

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

2.2.1 The Algorithmic Predictions: KPRA Risk Levels

Between March 18, 2011, and June 30, 2013, Kentucky used one fixed algorithm to make misconduct predictions: the Kentucky Pretrial Risk Assessment (KPRA).[5] The KPRA was developed by Kentucky Pretrial Services by fitting a regression model on statewide administrative data to predict pretrial misconduct (failure to appear in court or rearrest) (Austin, Ocker and Bhati, 2010).

The KPRA operated as a mechanical checklist. Pretrial officers answered 12 binary questions related to defendant criminal history, charge information, and defendant personal history. Each "yes" or "no" response carried a predetermined point value. Officers summed

[4] I calculated this figure by summing the population counts reported for distinct PSA jurisdictions in Advancing Pretrial Policy & Research (2025*b*).

[5] Before March 18, 2011, Kentucky used an earlier version of the KPRA with different inputs, weights, and risk level thresholds (Austin, Ocker and Bhati, 2010). After June 30, 2013, the state adopted a different tool, the Public Safety Assessment (Laura and John Arnold Foundation, 2014). I restrict my study to the window of time from March 18, 2011, to June 30, 2013, so that the construction of algorithmic predictions remains constant.

these values to produce a risk score ranging from 0 to 24. Appendix Table A.1 reproduces the full scoring rubric from Austin, Ocker and Bhati (2010).

Risk scores were then mapped to three risk levels using fixed cutoffs: 0-5 (low), 6-13 (moderate), and 14-24 (high). These risk levels are what was available to judges rather than the underlying risk scores (Kentucky Courts, 2019). Therefore, these discrete risk levels are the algorithmic predictions in my setting.

Because both the KPRA scoring rules and the risk level cutoffs were unchanged over my study period, the algorithmic predictions remained stable. This feature is important for my empirical design: it ensures that any changes in judge decisions around June 2011 cannot be attributed to changes in the prediction model itself.

2.2.2 The Introduction of Algorithmic Recommendations

To address rising incarceration costs between 2000 and 2010, Kentucky enacted House Bill 463 (HB463) in June 2011 (Pew Center on the States, 2011; Kentucky General Assembly, 2011). The law added a new layer on top of the existing KPRA: *algorithmic recommendations* that recommended bail conditions based on the KPRA's risk levels.

Under HB463, cases with low or moderate risk levels were recommended release without money bail – which I refer to as "lenient bail" recommendations for brevity (Kentucky Courts, 2019). Cases classified as high risk received no such recommendation. Importantly, HB463 did not change the KPRA's inputs, scoring rules, or risk level cutoffs; it changed only how those existing predictions were normatively mapped into recommended actions.

This institutional feature is central to my identification strategy: the prediction algorithm remained fixed, but new recommendations were introduced in June 2011. Figure 1 illustrates the mapping from KPRA risk scores to risk levels and to recommended actions before and after HB463.

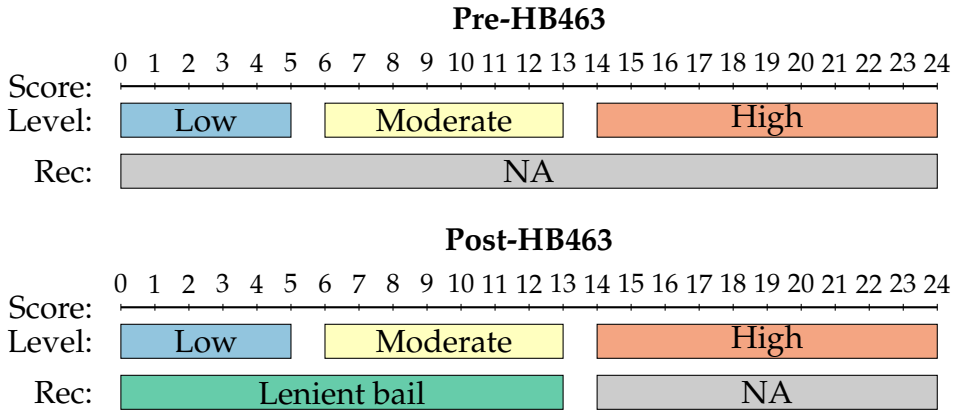
2.2.3 The Mechanics of Bail Decisions

Throughout the study period, bail decisions in Kentucky followed a consistent process. After booking, a pretrial services officer (an administrative court employee) interviewed the defendant, collected information, and computed the KPRA risk level.[6] Within 24 hours,

[6]Information in this subsection is based on 2019 phone interviews with the Executive Officer of Kentucky Pretrial Services and a Pretrial Officer who also served as a Risk Assessment Coordinator (hereafter, "KPS (2019)").

the officer relayed case information to a judge by phone, and the judge made a bail decision within minutes.[7]

Figure 1: Risk Scores, Risk Levels, and Bail Recommendations, Pre- and Post-HB463



Notes: This figure demonstrates the mapping from KPRA risk scores to risk levels and, subsequently, to bail recommendations. Both before and after HB463, scores of 0-5, 6-13, and 14-24 corresponded to low, moderate, and high risk levels, respectively. Before HB463, no risk level carried a bail recommendation. After HB463, low and moderate risk cases received a lenient bail recommendation, while high risk cases continued to receive no recommendation

The change introduced by HB463 concerned the algorithm-related content discussed during this phone call. KPRA risk levels were available to judges both before and after HB463. After HB463, two changes occurred. First, the new algorithmic recommendations were incorporated into the call: cases classified as low or moderate risk received a lenient bail recommendation, which officers communicated to judges during the call (Kentucky Courts, 2019). Second, KPRA risk levels – previously optional – became a required part of the judge-officer conversation.

Judges retained full discretion. Overriding a lenient bail recommendation required only a short justification communicated to the officer (e.g., stating "flight risk" on the phone).

2.3 Theoretical Framework and Testable Predictions

I describe two channels through which algorithmic recommendations may affect judicial decisions: by conveying predictive information or by directly changing decision-maker costs. These channels generate distinct empirical predictions in my setting. I summarize the main intuition below and present a formal model in Appendix A.3.

[7]The use of phone calls for bail decisions is specific to Kentucky; in-person bail hearings are more common in other US jurisdictions. Appendix A.2 provides additional context on how Kentucky's bail process compares to other US settings.

Theory 1: Recommendations only convey predictive information. Under this theory, recommendations matter only by conveying information about misconduct risk. Because judges already observe the underlying algorithmic risk level (KPRA level), recommendations do not add new predictive content. This theory therefore predicts no change in bail decisions following the introduction of recommendations.

Theory 2: Recommendations change costs. Under this theory, recommendations affect decisions through non-informational channels by altering the perceived costs of judicial choices. For example, following a recommendation generated by an external authority may provide institutional or procedural cover in the event of an adverse outcome (like pretrial misconduct). In addition, recommendations may act as defaults or soft constraints, making deviation from the recommended action more costly or effortful in general. In reduced form, these channels predict an increase in lenient bail among cases where the lenient recommendation applies (low and moderate risk cases).

3 Kentucky Administrative Court Data

I use administrative court data from Kentucky's Administrative Office of the Courts (AOC), which covers all criminal cases with felony- or misdemeanor-level charges in the state. Starting from the raw data shared by the AOC, I construct my dataset for this paper using the following steps.

i. Defining the appropriate observation level: The raw data consist 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 case, 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 restrict 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.[8] I restrict my data to this time frame so that the

[8]Before March 18, 2011, Kentucky used a different version of the KPRA with different inputs, weights,

construction of algorithmic predictions (KPRA risk levels) remains constant.

Second, I impose sample restrictions to rule out concerns that HB463 itself altered the composition of observed cases. A challenge in studying HB463 is that it was a large bill – approximately 150 pages and 110 sections – that implemented multiple reforms beyond the introduction of algorithmic bail recommendations (Kentucky General Assembly, 2011). A key empirical concern is therefore misattributing effects of concurrent policy changes to the bail recommendations.

A qualitative review of the bill, combined with interviews with practitioners, identifies a policing reform as the primary threat to identification (KPS, 2019). Specifically, HB463 amended existing law to require law enforcement officers to issue citations in lieu of arrest for certain misdemeanor offenses (White, 2011).[9] As a result, some misdemeanor offenses that previously led to arrest may no longer have entered the bail decision process after HB463.

To address this concern, I exclude from my sample cases that were newly subject to mandatory citation under HB463. This restriction removes cases whose post-reform absence could mechanically reflect changes in policing rather than changes in judge bail-setting behavior. I identify the relevant cases using the "Standard Operating Procedures" documentation from the Louisville Metro Police Department, which listed offenses affected by the new citation mandate (Louisville Metro Police Department, 2011). The final sample therefore consists only of offenses that were arrestable both before and after HB463.

iii. Constructing risk scores: The raw administrative data do not include the underlying KPRA risk scores; they include only the discretized KPRA risk levels (low, moderate, high). However, the data record all components used to construct the KPRA score. I therefore reconstruct the underlying risk scores using these components together with the published scoring weights described in Austin, Ocker and Bhati (2010).[10]

Final dataset: The resulting dataset consists of approximately 142,000 case-level observations covering 118,000 defendants and 423 judges. Each observation includes information on the defendant, the relevant charges, the initial bail decision, the bail judge, KPRA risk components, risk scores, and risk levels. As is standard in the pretrial context, the administrative data do not record the specific information considered by judges in individual bail

and risk level definitions (Austin, Ocker and Bhati, 2010). After June 30, 2013, Kentucky stopped using the KPRA because it adopted the Public Safety Assessment (Laura and John Arnold Foundation, 2014).

[9] As described in a memo from the Louisville Chief of Police, the new requirement preserved officer discretion to make physical arrests in some circumstances (White, 2011).

[10] Table A.1 reports the exact risk factors and point values used to calculate the KPRA score.

decisions.[11]

Table 1 reports descriptive summary statistics for the pre-period, before algorithmic recommendations were introduced. I report statistics for all cases and separately by KPRA risk level (low, moderate, and high). This breakdown is informative because, in the post-period, low and moderate risk cases receive a lenient bail recommendation, while high risk cases do not.

Table 1: Summary Statistics for the Pre-Period

	All	Low	Moderate	High
Sample size				
Observations	14,368	9,061	4,641	666
Case characteristics				
Share Misdemeanor Cases	0.60	0.65	0.52	0.47
Average Count of Charges	2.19	2.09	2.33	2.45
Defendant demographics				
Share Male Defendants	0.72	0.70	0.76	0.81
Share Black Defendants	0.17	0.14	0.21	0.25
Risk components				
Share No Verified Address	0.11	0.08	0.15	0.17
Share No Verified Support	0.57	0.50	0.68	0.77
Share Pending Charge	0.20	0	0.47	0.99
Share Failure to Appear (measure 1)	0.20	0.07	0.39	0.73
Share Failure to Appear (measure 2)	0.17	0.08	0.30	0.53
Share Prior Misdemeanor Conviction	0.70	0.59	0.89	0.99
Share Prior Felony Conviction	0.27	0.14	0.45	0.74
Share Prior Violent Conviction	0.23	0.13	0.38	0.59
Share Drug/Alcohol Abuse	0.12	0.03	0.23	0.46
Share Prior Felony Escape Charge	0.02	0	0.03	0.12
Share Felony Probation/Parole	0.1	0.04	0.18	0.31
Bail outcomes				
Share Money Bail	0.65	0.57	0.76	0.87
Share Lenient Bail	0.31	0.39	0.19	0.10
Share No Bond (Detained)	0.04	0.03	0.05	0.03
Median Bail Amount	$500	$250	$800	$1,500
Median Hours in Detention	24	17	49	107

Notes: This table reports descriptive statistics for cases in the pre-period (before HB463), shown for all cases and separately by KPRA risk level (low, moderate, and high). The table summarizes case characteristics, defendant demographics, risk components, and bail outcomes. Median bail amount is calculated including $0 observations for lenient bail.

[11] Administrative court data typically capture observable defendant and case characteristics but do not record verbatim communications or judicial deliberation (Dobbie, Goldin and Yang, 2018; Kleinberg et al., 2018).

Table 1 first summarizes case characteristics, showing that most cases are misdemeanors (60%) and involve just over two charges on average. It then reports defendant demographics: 72% cases involve male defendants, 83% involve white defendants, and 17% involve Black defendants.[12] The next table section reports on the frequency of the underlying risk components used to construct KPRA scores. As expected, higher risk levels are associated with substantially higher prevalence of adverse risk factors, which reflects the mechanical construction of the score.

The final table section summarizes bail and detention outcomes. In the pre-period, 65% cases received money bail, 31% received lenient bail, and only 4% resulted in detention without bond. The median bail amount was $500, and the median time spent in pretrial detention was 24 hours. Bail severity and time in detention increase with risk level.

Two features of Table 1 are particularly noteworthy. First, the vast majority of cases – approximately 90% – are classified as low or moderate risk. Second, despite this, only 32% of cases received lenient bail in the pre-period. This gap implies that the introduction of algorithmic recommendations corresponded to a substantially lower effective threshold for lenient bail relative to the pre-reform status quo.[13] These descriptive facts motivate the analysis that follows.

4 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

[12] Very few cases involve defendants who are not white or Black – only 0.3% of cases involve Asian defendants. I do not analyze ethnicity (Hispanic or non-Hispanic) because ethnicity is unknown for 27% of cases.

[13] Matching the pre-reform lenient bail rate would have entailed setting the lenient bail recommendation cutoff below a KPRA score of 4, rather than below 14.

these algorithmic predictions. Therefore, I take additional steps to test and adjust the estimated effects to align with the desired recommendation effects.

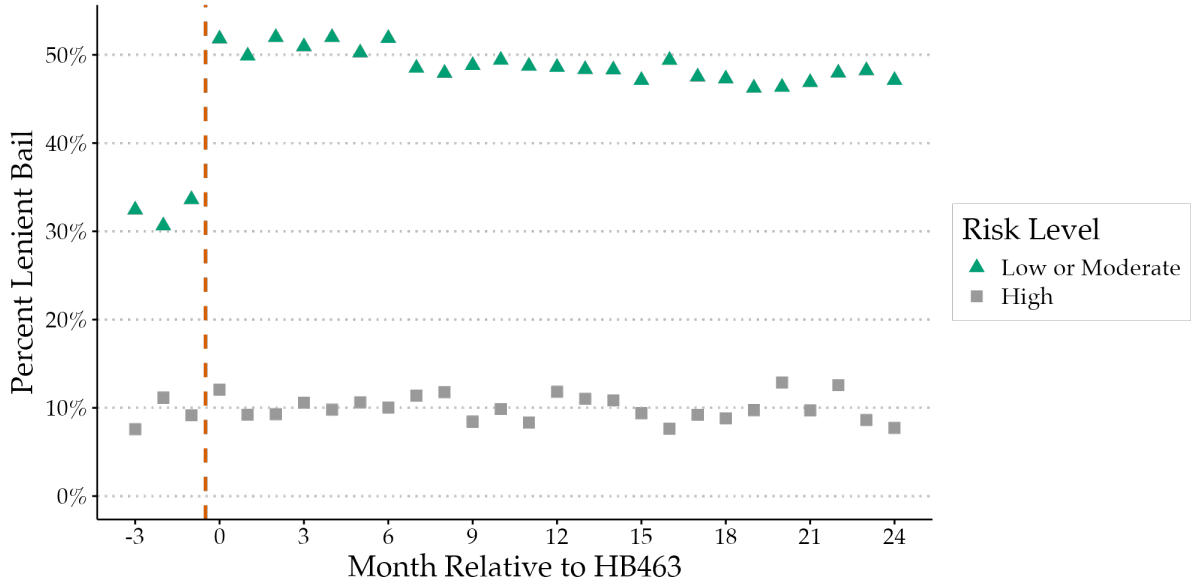
In Section 4.1, I demonstrate the straightforward (naive) differences-in-differences and differences-in-discontinuities results. In Section 4.2, I address concerns about potential confounding related to the use of risk levels with two different approaches. In the end, I find that algorithmic recommendations have independent causal effects on judge decisions, and these effects are robust to confounding concerns.

4.1 Naive Estimates

4.1.1 Differences-in-Differences Results

In my differences-in-differences framework, high risk cases serve as the control group because they never receive a lenient recommendation, while low and moderate risk cases are the treatment group because they newly receive a lenient recommendation under HB463. Figure 2 illustrates the rate of lenient bail for low or moderate risk cases and high risk cases over time.

Figure 2: Lenient Bail Rates by Risk Level over Time



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

Following the introduction of recommendations, lenient bail increases sharply for low and

moderate risk cases by roughly 15-20 percentage points. There is no similar increase for the high risk group.[14] The underlying assumption of using a differences-in-differences approach is parallel pre-trends. The raw visual evidence in Figure 2 is consistent with the parallel trends assumption.[15]

To formally estimate dynamic treatment effects and assess pre-trends, I estimate the following event-study specification:

$$\text{lenient}_{itj} = \sum_{m \neq -1} (\beta_m \times \mathbb{1}\{\text{score}_i < 14\}) + \mathbf{X}'_{ijt}\gamma + \varepsilon_{itj}, \tag{1}$$

where lenient_{itj} is an indicator for if the bail for case i at time t decided by judge j is lenient (no money bail) and $\mathbb{1}\{\text{score}_i < 14\}$ is an indicator for if the risk score for case i is below 14, which means 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 $\mathbf{X}_{ijt}$ that accounts 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 introduction of recommendations, 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.5 shows that results with zero detailed controls are nearly identical to those with controls based on all observed case variables.

To obtain a summary coefficient, I estimate pooled differences-in-differences coefficients and present these results across specifications in Table B.1. Pooling time periods, I find that algorithmic recommendations increased lenient bail by approximately 17 percentage points following the policy change, off of a baseline of 31%.[16] Therefore, the recommendations

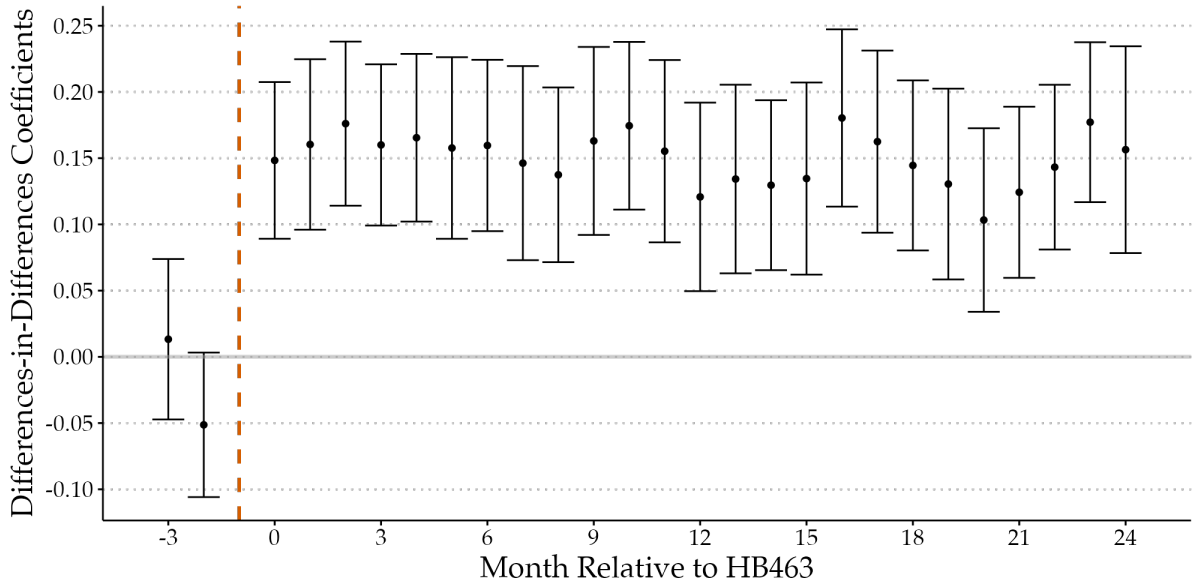
[14] I focus on the binary measure of lenient bail because it directly corresponds to the recommended action and summarizes the main empirical effects. Figure B.1 shows that the shift toward lenient bail reflects reductions in monetary bail amounts in the hundreds and thousands of dollars, and that there is no stark difference in the distribution for high risk cases. Figure B.2 shows no discontinuous change in outright detention across risk levels after recommendations are introduced. Finally, Figure B.3 shows that lenient bail is associated with very short pretrial detention (a median of 10 hours), whereas higher bail amounts are associated with longer detention durations.

[15] The number of pre-policy periods is limited because the construction of the KPRA risk score changed in March 2011 (see Appendix A.1). To ensure that risk scores and risk levels are comparable over time, I exclude observations from before March 2011 from the analysis sample, as described in Section 3. Appendix Figure B.4 shows that including earlier data without harmonizing the score or level construction also yields no evidence of pre-trends.

[16] Heterogeneity analyses are reported in Appendix B.2.

increased lenient bail by about 55%. These economically meaningful results are consistent with the theory that algorithmic recommendations change the costs of errors to decision-makers.

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.

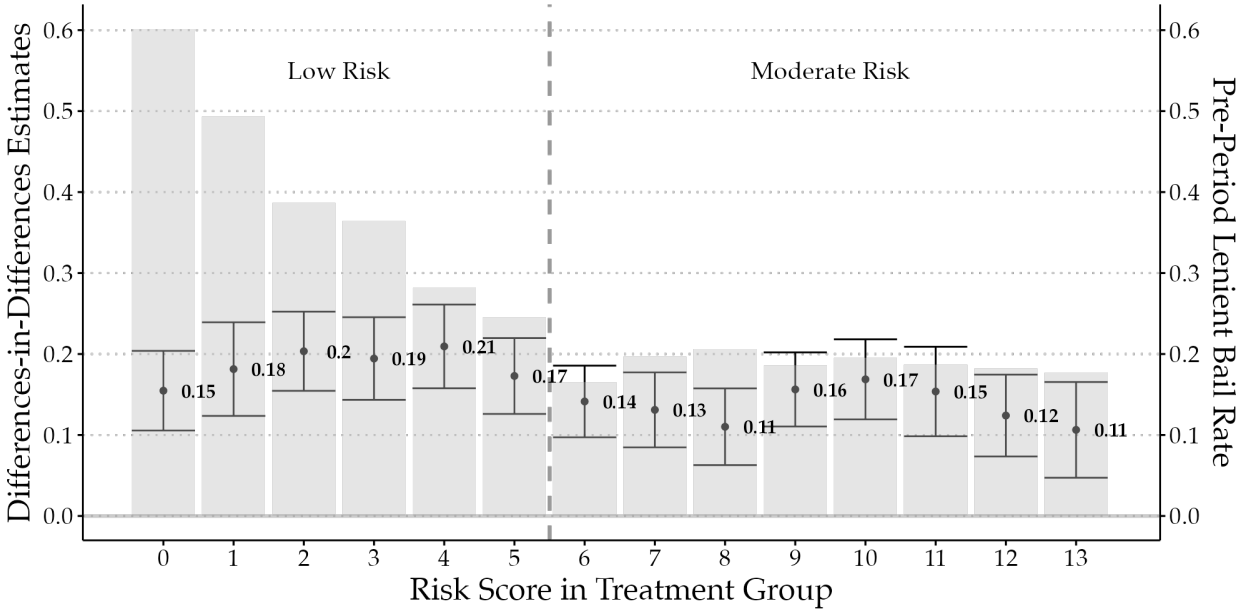
Recommendation Effects vs. Broad Preference Shift: A competing interpretation of the results is that HB463 shifted judges' underlying preferences toward leniency, independent of the introduction of algorithmic recommendations. Under such a broad preference-shift mechanism, one would generally expect either (i) increases in lenient bail across all risk levels, including high risk cases, or (ii) heterogeneity in effects across risk levels with judges concentrating their leniency for the lower risk levels. What would be difficult to reconcile with a broad preference shift is a pattern featuring no change in lenient bail for high risk cases combined with similar increases for both low and moderate risk cases.

The data do not support a broad preference-shift interpretation. Figure 2 shows that lenient bail rates for high risk cases remained essentially unchanged following HB463, while lenient bail increased sharply for cases that newly received a lenient recommendation (low and moderate risk cases). This contrast suggests that the new law did not induce a general shift in judicial preferences toward leniency for all cases, but instead generated changes aligned with the new recommendation.

Figure 4 further clarifies this distinction by examining heterogeneity within the recom-

mended region of the risk score distribution. I estimate pooled differences-in-differences coefficients separately for each risk score in the low and moderate risk range (scores 0–13).[17] Across all scores in this range, lenient bail increases by similar magnitudes, and the effects are statistically significant throughout the distribution. Notably, effect sizes do not change discontinuously at the low-moderate risk threshold.

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). The vertical gray dashed line marks the threshold between low and moderate risk. 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.

The similarity of these effect sizes across the low-moderate range is informative about mechanism. This pattern is consistent with recommendation effects rather than broad preference shift effects. Once a case falls into the region where lenient bail is recommended, following that recommendation becomes uniformly less costly, regardless of how far inside the recommended region the case lies.

By contrast, a broad shift in judicial preferences toward leniency following HB463 would

[17]The estimated coefficients across the distribution are similar regardless of specification and control choices, as demonstrated by Figure B.6.

not predict the combination of no effect for high risk cases and similar percentage-point increases throughout the low-moderate distribution. Reconciling these patterns with a broad preference-shift mechanism would require judges to apply any increased leniency only to low and moderate risk cases, and to apply it similarly across those groups. This alignment would be difficult to explain absent reliance on the recommendation rule itself.

4.1.2 Differences-in-Discontinuities Results

I also estimate recommendation effects using a complementary identification strategy that focuses on cases near the recommendation threshold. Throughout the sample period, judges could observe defendants' KPRA risk levels (low, moderate, high), which are a deterministic function of the underlying risk score. As of June 2011, cases with risk scores below 14 additionally receive an explicit lenient bail recommendation, while cases at or above this threshold do not.

A simple post-period regression discontinuity at the critical threshold would not isolate the effect of the recommendation for two reasons. First, judges observe risk levels and risk level itself changes discontinuously at the cutoff; therefore, this approach would confound the effect of predictive information with the recommendation.[18] Second, case features are not smooth through the cutoff (Figure B.7). In particular, 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, which violates standard regression discontinuity assumptions.

To address these challenges, I implement a differences-in-discontinuities design that compares changes in the size of the discontinuity at the threshold before and after June 2011. This approach differences out time-invariant discontinuous effects at the threshold, isolating the causal effect of the recommendation for marginal cases near the threshold. In other words, this design isolates the effect of adding a lenient recommendation, holding fixed the predictive information conveyed by algorithmic risk levels.

In a conventional regression discontinuity design, the 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 as long as those same discontinuities are present in both time periods (Grembi, Nannicini and Troiano, 2016). In other

[18] In the post-period, cases scored as 14 receive a *high risk* label and no recommendation, while cases scored as 13 receive a *moderate risk* level and a lenient recommendation.

words, the complier population (the population near the moderate-high risk threshold) must remain consistent across the two time periods.

This assumption is supported both institutionally and empirically. Institutionally, the construction of the KPRA score and the location of the relevant thresholds are fixed throughout the study time frame, and cases impacted by concurrent policy changes were removed from my sample (as described in Section 3). Empirically, Figure B.8 demonstrates that covariate patterns around the critical threshold were similar before and after HB463. In particular, 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.

For transparency, Figure 5 demonstrates the differences-in-discontinuities approach visually with raw data. 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.[19] 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).

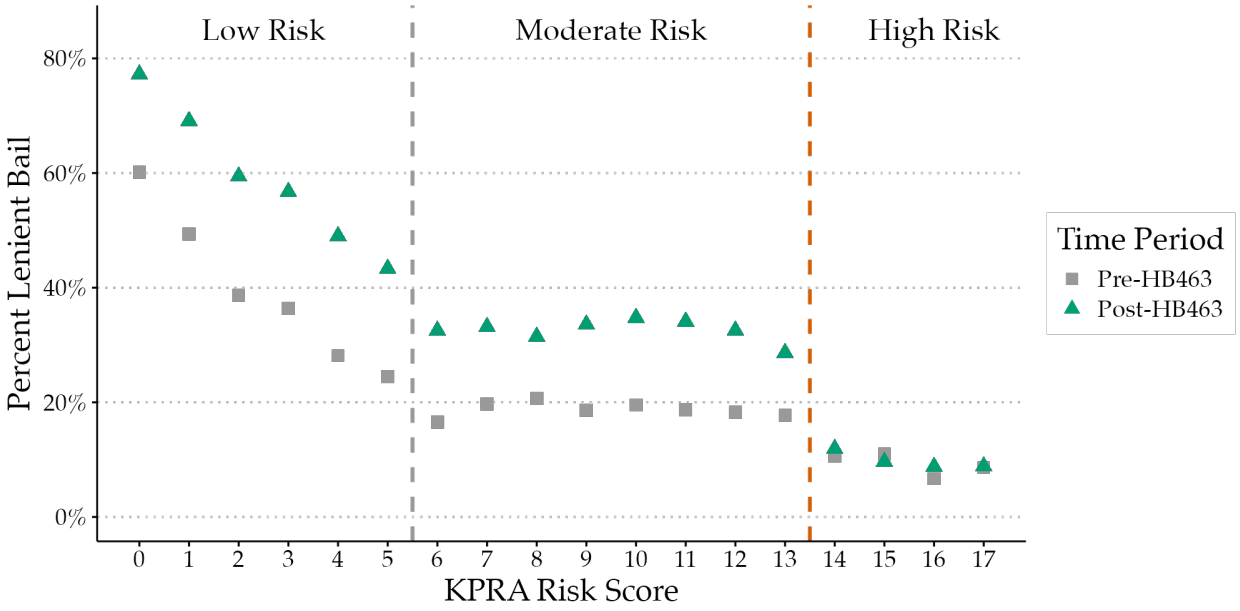
There were no changes in recommendations for high risk cases (points to the right of the orange dashed line) across time periods, but there were changes for low or moderate risk cases (points to the left of the orange dashed line). For cases that did not experience a change in recommendation, lenient bail rates are nearly identical in the pre- and post-periods. However, for cases that did experience a change in recommendations, lenient bail rates are meaningfully higher in the post-period.[20] 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

[19] 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 B.9 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.

[20] 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).

pre-period (before the introduction of recommendations).[21]

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 vertical gray dashed line marks the threshold between low and moderate risk, and the vertical 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.

To formally estimate the effect of the recommendation at the margin, I use the 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. For both the pre-period and post-period data, I separately estimate the effect of crossing the moderate-high threshold using nonparametric methods following Calonico, Cattaneo and Titiunik (2014) and Calonico, Cattaneo and Farrell (2020) for optimal bandwidth selection as well as robust bias-corrected confidence intervals and inference procedures.[22]

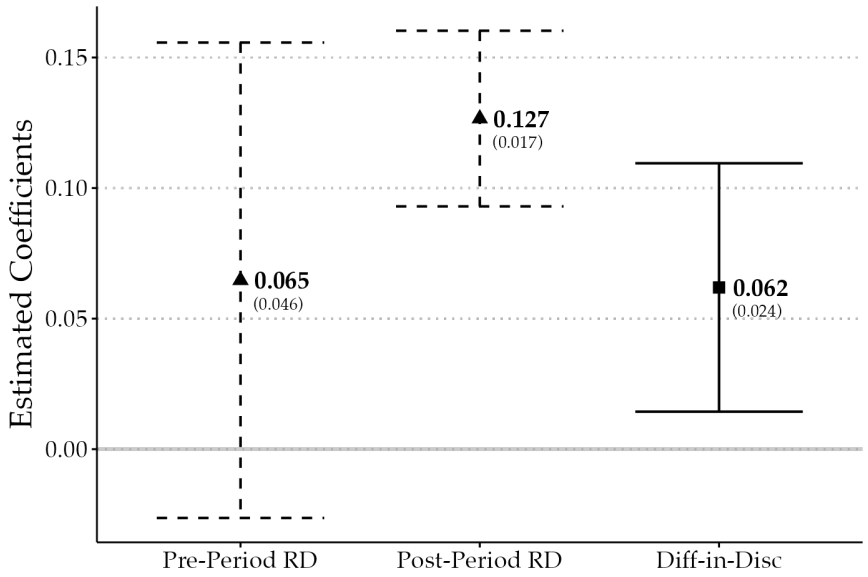
Figure 6 demonstrates the results. The regression discontinuity estimate in the post-period, which is the effect of changing risk levels, the effect of a prior felony conviction, *and* the

[21] As an alternative way to visualize dynamics, Figure B.10 plots an event-study-style series of regression discontinuity estimates at the moderate-high risk threshold, measured relative to the month immediately preceding the implementation of HB463. This figure shows that the discontinuity in lenient bail at the moderate-high threshold increases sharply at the time of HB463 and remains elevated thereafter.

[22] Because the recommendation applies to moderate risk scores, I report RD estimates as the effect of crossing the threshold from high to moderate risk (right to left). In practice, this corresponds to reporting the negative of the default right-minus-left estimand using the `rdrobust` package (Calonico et al., 2023).

effect of the recommendation, is 12.7 percentage points. Meanwhile, the estimate in the pre-period, which is the effect of changing risk levels and the felony conviction effect, is 6.5 percentage points.[23] The resulting differences-in-discontinuities estimate is 6.2 percentage points. 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).

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 solid black lines for the 95% confidence intervals. It is the difference between the two regression discontinuities.

The differences-in-discontinuities result is directionally consistent with the estimated differences-in-differences results. Their magnitudes differ because they study different populations of cases, and they rely on different underlying assumptions for identification. Both sets of results suggest that algorithmic recommendations have their own independent effects, consistent with the theory that recommendations change the costs of errors to

[23]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).

human decision-makers (Section 2.3).[24]

4.2 Testing and Adjusting the Naive Estimates

Both the differences-in-differences and differences-in-discontinuities strategies in Section 4.1 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 levels 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 that judges consider risk levels. Therefore, if some judges did not consider risk levels before the policy, then the policy changed the presence of algorithmic predictions and recommendations. This means the straightforward differences-in-differences and differences-in-discontinuities results from Section 4.1 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. In Section 4.2.1, I bound the recommendation effect using differences-in-discontinuities methods. In Section 4.2.2, I estimate recommendation effects with differences-in-differences for an intuitive subset of cases where confounding should be limited. These two approaches are distinct in their assumptions and methods, but both continue to provide evidence that algorithmic recommendations have independent causal effects on decision-making.

4.2.1 Bounding Recommendation Effects using Differences-in-Discontinuities

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 high risk for marginal cases), and the effect of increased prior felony conviction (F):[25]

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

[25] 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 by $0.2F'$,

$RD^{post}_{mh} = R + L_{mh} + F$.

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): $RD^{pre}_{mh} = L_{mh} + F$. Therefore, the difference between the two (the differences-in-discontinuities coefficient) isolates the desired recommendation effect (R):

$$DiDC_{mh} = RD^{post}_{mh} - RD^{pre}_{mh} = R.$$

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: $RD^{post*}_{mh} = R + L_{mh} + F$. 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 did not consult risk levels before HB463 respond to risk levels similarly to those who did (Assumption 1), then I can write the pre-period regression discontinuity estimate as $RD^{pre*}_{mh} = \omega L_{mh} + F$. 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}:

$$DiDC^{*}_{mh} = RD^{post*}_{mh} - RD^{pre*}_{mh} = R + (1 - \omega)L_{mh}.$$

Since $\omega \geq 0$, the differences-in-discontinuities estimate from Section 4.1.2 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. Cases also experience a discontinuous change in their risk level at the low-moderate threshold, as they do at the medium-high threshold. However, importantly, there is no change in 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

where F' is the sharp effect of moving from 0% to 100% of cases with prior felony convictions.

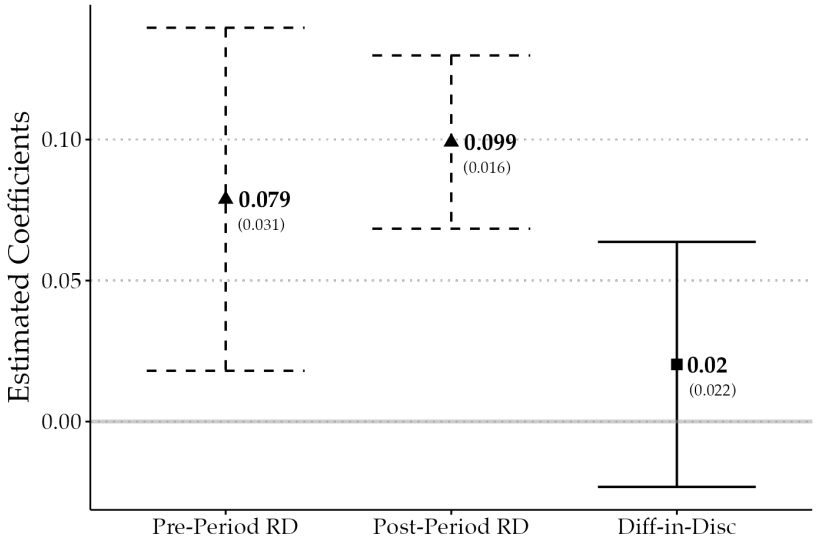
labeled low risk rather than moderate risk for marginal cases: $RD_{lm}^{post*} = L_{lm}$.

If I assume that judges consult risk levels at the same rates near the low-moderate and moderate-high thresholds before HB463 (Assumption 2), then I can write the pre-period regression discontinuity as $RD_{lm}^{pre*} = \omega L_{lm}$.[26] Accordingly, the resulting differences-in-discontinuities estimate at the low-moderate threshold is

$$DiDC_{lm}^{*} = RD_{lm}^{post*} - RD_{lm}^{pre*} = (1 - \omega)L_{lm}.$$

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: 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 solid black lines for the 95% confidence intervals. It is the difference between the two regression discontinuities.

[26]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 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. The fact that the differences-in-discontinuities estimate is near 0 while L_{lm} is not suggests that ω is close to 1 and confounding is limited.

Bounds using point estimates: I can use the point 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 $DiDC_{mh}^{*} = R + (1 - \omega)L_{mh}$. Plugging in the estimated differences-in-discontinuity coefficient, ω, and rearranging terms yields the expression $R = 0.062 - (0.20)L_{mh}$.

Therefore, R depends on the magnitude of L_{mh}, which I can also bound based on previous expressions and estimates. Recall that $RD_{mh}^{pre*} = \omega L_{mh} + F$. 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 that $L_{mh} \in [0, 0.081]$. Thus, the possible values of the recommendation effect R can be written as follows:

$$R = 0.062 - 0.20\, L_{mh}, \qquad \text{where } L_{mh} \in [0,\ 0.081].$$

This expression for R and range for L_{mh} implies that 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.

Evaluating the required assumptions: The above bounding exercise relies on two assumptions about judicial decision-making. Assumption 1 is that judges who did not consult risk levels before HB463 would respond to risk level information similarly to those who did. Assumption 2 is that before HB463, judges consulted risk levels at similar rates near the low-moderate and moderate-high thresholds.

Neither assumption can be directly tested because I do not observe when individual judges consult risk levels. Nevertheless, it is useful to consider the likely direction of any bias if the assumptions are violated. In both cases, intuitive violations would bias my estimates toward *larger* recommendation effects, rendering the bounds conservative.

For Assumption 1, a natural concern is that judges who did not voluntarily consult risk levels pre-HB463 were more skeptical of algorithmic tools and thus less responsive to the risk information than early adopters were. If these "late adopters" indeed place less weight on risk levels, then the contamination term $(1-\omega)L_{mh}$ used in constructing the bounds is overstated. Accounting for this would reduce the estimated contamination term and shift my recommendation effect bounds upward.

For Assumption 2, judges might be more likely to consult risk levels when cases appear higher risk (because of reputational concerns). This would imply $\omega_{mh} > \omega_{lm}$. Since I estimate ω using the low-moderate threshold, this again means that $(1-\omega)L_{mh}$ is overstated: more judges were already consulting risk levels near the moderate-high threshold in the pre-period. Accounting for this violation would similarly shift the recommendation effect bounds upward.

Robustness to uncertainty in ω: While the point estimate approach provides useful bounds for the recommendation effect, these bounds depend critically on ω, which itself is estimated with uncertainty. I now assess how uncertainty in ω affects the recommendation effect.

To characterize the uncertainty in ω, I conduct a Monte Carlo simulation with 10,000 draws. Since $\omega = RD_{lm}^{pre*}/RD_{lm}^{post*}$, I treat RD_{lm}^{pre*} and RD_{lm}^{post*} as normally distributed random variables based on the estimated means and standard errors shown in Figure 7. For each simulation draw, I sample values from these distributions and calculate the implied ω as the ratio of the two. This approach allows me to propagate the uncertainty in the regression discontinuity estimates through to an empirical distribution for the parameter ω. Figure 8 shows the resulting distribution of ω with gray bars, which centers around 0.80.[27]

My goal is to determine what values of ω would be necessary to eliminate the recommendation effect (i.e., to yield $R = 0$), and then assess the likelihood of such values given the simulated distribution. Based on prior equations, the isolated recommendation effect R

[27]Note that this approach generates values below 0 and above 1 even though $\omega \in [0,1]$ conceptually (because it is a share). I can limit ω values to the $[0,1]$ range, and the results in the following exercise are unchanged. See Figure B.11 for a version of this graph with ω values limited to the $[0,1]$ range.

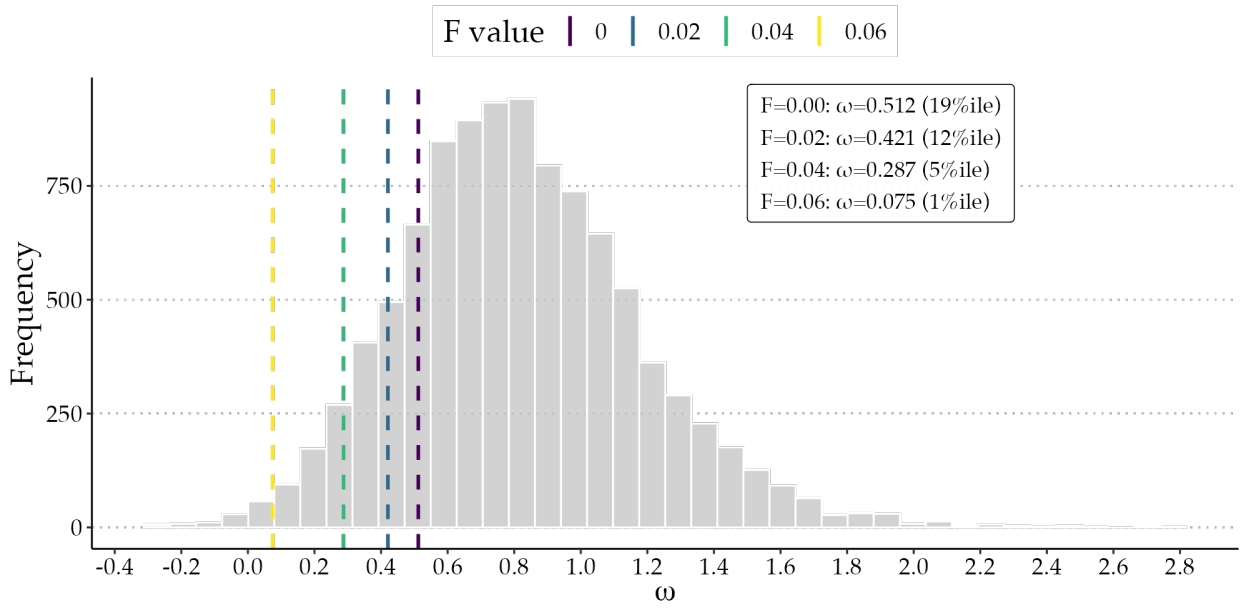
can be written as

$$R = DiDC^{*}_{mh} - (1-\omega)(\frac{RD^{pre*}_{mh} - F}{\omega}).$$

Plugging in the point estimates for $DiDC^{*}_{mh}$ and RD^{pre*}_{mh} from earlier results, I can determine what values of ω are needed for R to equal zero. This depends on the value of F, the effect of increased prior felony conviction. Since $RD^{pre*}_{mh} = \omega L_{mh} + F$ and I assume $F \geq 0$ and $L_{mh} \geq 0$ (since these factors should only make judges more strict), then it must be the case that $RD^{pre*}_{mh} \geq F$. If I use the point estimate for RD^{pre*}_{mh}, this implies $F \in [0, 0.065]$.

Therefore, I solve for the critical ω values that render $R = 0$ for different values of $F \in [0, 0.065]$. I plot these critical ω values as vertical dashed lines in Figure 8, with different colors corresponding to different values of F. This figure demonstrates which values of ω would yield a null recommendation effect, and the likelihood of those values given the empirical distribution of ω.

Figure 8: Omega Values Required for a Recommendation Effect of 0



Notes: This figure is a histogram of values of ω generated from a Monte Carlo simulation with 10,000 draws. $\omega = RD^{pre*}_{lm}/RD^{post*}_{lm}$ such that RD^{pre*}_{lm} and RD^{post*}_{lm} are normally distributed random variables based on their estimated means and standard errors. The vertical dashed lines then denote which values of ω are necessary to generate a recommendation effect of 0. There are different colors of dashed lines for different values of F, the effect of increased felony convictions.

The ω values that yield a null recommendation effect depend on F. If F is near its upper bound at 0.06, then $\omega = 0.075$ would be required to yield $R = 0$. Under this scenario, the recommendation effect is null only if judges consulted risk levels in just 7.5% of pre-HB463

cases. This value falls at the 1st percentile of the empirical distribution of ω, which makes it highly unlikely. However, if F is lower at 0.02, then $\omega = 0.42$ would yield $R = 0$, meaning the recommendation effect is null if judges consulted risk levels in 42% of pre-HB463 cases. While this ω value is more likely than 0.075, it still falls at only the 12th percentile of the empirical distribution, suggesting it is relatively unlikely given the data.

Overall, for risk level saliency to fully account for the estimated effects, ω would need to fall between the 1st and 19th percentiles of its empirical distribution, depending on the value of F.[28] While this range of scenarios cannot be definitively ruled out, they represent unlikely outcomes relative to the ω distribution.

4.2.2 Recommendation Effects for Lowest-Risk-Scoring Cases

My second approach to addressing potential confounding uses a simple intuitive strategy that does not require the two behavioral assumptions needed in Section 4.2.1. The key idea is to test whether recommendations matter for a subset of cases where the expected effect of risk levels is very small. If confounding from risk level visibility is minimal, then the differences-in-differences result for this group should primarily capture the effect of algorithmic recommendations.

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.). For these cases, the formal "low risk" label provides little additional predictive information beyond what judges already observe from case characteristics. While judges may still believe there is some probability of misconduct, they can already infer low risk from the observable combination of misdemeanor charges and clean criminal histories during the bail phone call. The key assumption therefore is that hearing "low risk" does not substantially update their beliefs about misdemeanor defendants they already know have zero risk factors. Since $L \approx 0$ intuitively for this group of cases, the differences-in-differences estimate should approximate a recommendation effect without requiring assumptions about (1) whether judges consult risk levels at similar rates across thresholds or (2) whether "early-adopter" and "late-adopter" judges respond similarly to risk level information.

Misdemeanor cases with zero associated risk factors are 7% of cases in the data. Figure B.12 illustrates the rate of lenient bail for these cases in contrast to the rate of lenient bail

[28] The upper bound (19th percentile) occurs only when $F = 0$, which corresponds to the extreme case of no effect of a prior felony convictions on decisions. For a more plausible value such as $F \in [0.02, 0.06]$, ω would need to fall between the 1st and 12th percentiles of its empirical distribution.

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, lenient bail rates increase by 10-15 percentage points for this group 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 4.1.1. Figure B.13 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.14 shows that results with no detailed controls are nearly identical to those with controls based on all observed case variables.

Table 2: Differences-in-Differences Results across Specifications (Treated Group: Lowest Risk Cases)

	Dependent variable: I(lenient bail)		
I(score<14) x Post	0.149***	0.150***	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	0.554	0.554	0.490
Adjusted R^2	0.543	0.543	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 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.

Table 2 shows the pooled differences-in-differences results across different sets of controls. The estimated coefficients in this case are around 15 percentage points, a similar magnitude to the estimates for the entire sample in Table B.1 (about 17 percentage points). However, the relative effects are smaller because the baseline lenient bail rates are higher for this low risk sample. For this specific subsample, the 15 percentage point increase in lenient bail is a 23% increase relative to the baseline of 66%. These results demonstrate recommendation effects survive in a subsample of cases where potential confounding is necessarily minimal.

Overall, I demonstrate that the estimated effects of algorithmic recommendation are robust to concerns about confounding using two complementary approaches. Section 4.2.1 derives bounds on recommendation effects using differences-in-discontinuities, implying a 30-40% increase in lenient bail for marginal cases. Section 4.2.2 instead focuses on cases where confounding is intuitively minimal, estimating a 23% increase in lenient bail for the lowest-risk-scoring cases. Despite relying on different assumptions and empirical strategies, both approaches provide consistent evidence that algorithmic recommendations have independent causal effects on human decisions.

5 Conclusion

This paper studies how predictive algorithms affect human decisions. Most existing work focuses on algorithms as providers of data-driven predictions (Kleinberg et al., 2018; Agrawal, Gans and Goldfarb, 2018; Mullainathan, 2025). I demonstrate that algorithms can matter in another important way: they 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 in a unique setting in the US criminal justice system. In this setting, the introduction of algorithm-based lenient recommendations increased judges' use of lenient bail by 30-40% for marginal cases.

These economically meaningful effects are not attributable to changes in predictive information available to decision-makers. Instead, the evidence is consistent with recommendations altering the perceived costs of different choices. For example, recommendations may change perceived accountability for certain decisions or introduce cognitive costs to deviating from a recommended action. In this way, algorithms can impact decision-maker preferences as well as predictions.

More broadly, these findings highlight that algorithms can influence not only how decisions are allocated across individuals, but also the overall composition of decisions. When decision-makers and algorithm designers differ in how they weight the costs of errors, algorithmic recommendations can serve as a tool for aligning decision-making with policy objectives (McLaughlin and Spiess, 2025).[29] In this sense, algorithmic recommendations

[29] Alternative policy tools, such as explicitly codifying the costs associated with different decision errors, could also affect judicial behavior. Comparing the relative effectiveness of these approaches is left for future work.

can be viewed as a form of what Cowgill and Stevenson (2020) describe as "algorithmic social engineering": recommendations are derived from algorithmic predictions but designed to advance certain normative or policy goals.

References

Abaluck, Jason, Leila Agha, David C Chan Jr, Daniel Singer, and Diana Zhu. 2021. "Fixing Misallocation with Guidelines: Awareness vs. Adherence." National Bureau of Economic Research Working Paper 27467.

Advancing Pretrial Policy & Research. 2025*a*. "How the PSA Works." https://www.advancingpretrial.org/how-the-psa-works/. Accessed December 10, 2025.

Advancing Pretrial Policy & Research. 2025*b*. "PSA Map." https://www.advancingpretrial.org/psa-map/. Accessed December 10, 2025.

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

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence.* Harvard Business Review Press.

Albright, Alex. 2025. "No Money Bail, No Problems? Trade-Offs in a Pretrial Automatic Release Program." Preprint, SocArXiv, March 31. https://doi.org/10.31235/osf.io/42pbz_v2.

Almog, David, Romain Gauriot, Lionel Page, and Daniel Martin. 2025. "AI Oversight and Human Mistakes: Evidence from Centre Court." Preprint, arXiv, February 12. https://doi.org/10.48550/arXiv.2401.16754.

American Bar Association Criminal Justice Standards Committee. 2007. *ABA Standards for Criminal Justice: Pretrial Release.* 3rd ed. American Bar Association.

Arnold, David, Will Dobbie, and Crystal S 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." *Bureau of Justice Statistics*.

Baron, E Jason, Joseph J Doyle Jr, Natalia Emanuel, Peter Hull, and Joseph Ryan. 2024. "Discrimination in Multiphase Systems: Evidence from Child Protection." *Quarterly Journal of Economics*, 139(3): 1611–1664.

Berk, Richard A, Susan B Sorenson, and Geoffrey Barnes. 2016. "Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions." *Journal of Empirical Legal Studies*, 13(1): 94–115.

Bhutta, Neil, Aurel Hizmo, and Daniel Ringo. 2025. "How Much Does Racial Bias Affect Mortgage Lending? Evidence from Human and Algorithmic Credit Decisions." *Journal of Finance*, 80(3): 1463–1496.

Bohren, J Aislinn, Peter Hull, and Alex Imas. 2025. "Systemic Discrimination: Theory and Measurement." *Quarterly Journal of Economics*, 140(3): 1743–1799.

Bushway, Shawn D, Emily G 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, Max Farrell, and Rocío Titiunik. 2023. "`rdrobust`: Robust Data-Driven Statistical Inference in Regression-Discontinuity Designs." R package documentation.

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

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

Chang, Wei, Priscilla Owusu-Mensah, Jordan Everson, and Chelsea Richwine. 2025. "Hospital Trends in the Use, Evaluation, and Governance of Predictive AI, 2023–2024." U.S. Department of Health and Human Services, Office of the Assistant Secretary for Technology Policy Data Brief 80.

Cornell Law School Legal Information Institute. 2022. "Bondsman." *Wex*, https://www.law.cornell.edu/wex/bondsman. Accessed February 18, 2026.

Cowgill, Bo. 2018. "The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities." Unpublished Manuscript.

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

Davenport, Diag. 2023. "Discriminatory Discretion: Theory and Evidence from Use of Pretrial Algorithms." Unpublished Manuscript.

Desmarais, Sarah L., and Evan M. Lowder. 2019. "Pretrial Risk Assessment Tools: A Primer for Judges, Prosecutors, and Defense Attorneys." Issue Brief, Safety and Justice Challenge.

Dobbie, Will, Jacob Goldin, and Crystal S Yang. 2018. "The Effects of Pre-Trial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges." *American Economic Review*, 108(2): 201–240.

Feigenberg, Benjamin, and Conrad Miller. 2021. "Racial Divisions and Criminal Justice: Evidence from Southern State Courts." *American Economic Journal: Economic Policy*, 13(2): 207–240.

Fergusson, Grant. 2023. "Outsourced and Automated: How AI Companies Have Taken Over Government Decision-Making." Report, Electronic Privacy Information Center. Accessed: 2026-02-04.

Fung, Katherine. 2021. "Darrell Brooks Should Not Have Been Released on Low Bail, Milwaukee DA Admits." *Newsweek*. November 22. https://www.newsweek.com/darrell-brooks-should-not-have-been-released-low-bail-milwaukee-da-admits-1652059.

Goel, Sharad, Justin M Rao, and Ravi Shroff. 2016. "Personalized Risk Assessments in the Criminal Justice System." *American Economic Review*, 106(5): 119–123.

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

Grimon, Marie-Pascale, and Christopher Mills. 2025. "Better Together? A Field Experiment on Human-Algorithm Interaction in Child Protection." Preprint, arXiv. https://doi.org/10.48550/arXiv.2502.08501.

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

Hausman, David K. 2025. "Risk Assessment as Policy in Immigration Detention Decisions." *Journal of Law and Economics*, 68(1): 103–119.

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li. 2018. "Discretion in Hiring." *Quarterly Journal of Economics*, 133(2): 765–800.

Johnson, Eric J, and Daniel Goldstein. 2003. "Do Defaults Save Lives?" *Science*, 302(5649): 1338–1339.

Jung, Jongbin, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G Goldstein. 2017. "Simple Rules for Complex Decisions." Preprint, arXiv. https://doi.org/10.48550/arXiv.1702.04690.

Kentucky Courts. 2019. "Virtual Tour of Kentucky Pretrial Services." http://courts.ky.gov/courtprograms/pretrialservices/Pages/virtualtour.aspx. Accessed 2019.

Kentucky General Assembly. 2011. "House Bill 463." *2011 Regular Session*, The Public Safety and Offender Accountability Act.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics*, 133(1): 237–293.

Lattimore, Pamela K, Stephen Tueller, Alison Levin-Rector, and Amanda Witwer. 2020. "The Prevalence of Local Criminal Justice Practices." *Federal Probation*, 84(1): 28–37.

Laura and John Arnold Foundation. 2014. "Results from the First Six Months of the Public Safety Assessment – Court in Kentucky." Report.

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

Loudenback, Jeremy. 2022. "The Foster Care System Turns to Big Data: Promising or Profiling?" *The Imprint*. https://imprintnews.org/child-welfare-2/the-foster-care-system-turns-to-big-data-promising-or-profiling/62359. Accessed: February 4, 2026.

Louisville Metro Police Department. 2011. "Standard Operating Procedures, SOP Number 10.1: Arrests—Enforcement." Louisville Metro Police Department, Effective February 18, 2004; previously revised May 9, 2011; revised June 12, 2011.

Madrian, Brigitte C, and Dennis F Shea. 2001. "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior." *Quarterly Journal of Economics*, 116(4): 1149–1187.

McLaughlin, Bryce, and Jann Spiess. 2025. "Algorithmic Assistance with Recommendation-Dependent Preferences." Preprint, arXiv. https://doi.org/10.48550/arXiv.2208.07626.

Mullainathan, Sendhil. 2025. "Economics in the Age of Algorithms." *AEA Papers and Proceedings*, 115: 1–23.

Pew Center on the States. 2011. "2011 Kentucky Reforms Cut Recidivism, Costs: Broad Bill Enacts Evidence-Based Strategies." Issue Brief.

Sloan, CarlyWill, George Naufal, and Heather Caspers. 2025. "The Effect of Risk Assessment Scores on Judicial Behavior and Defendant Outcomes." *Journal of Human Resources*, 60(5): 1778–1810.

Slutzky, Pablo, and Sheng-Jun Xu. 2025. "The Financial Consequences of Pretrial Detention." *The Review of Financial Studies*, 38(11): 3329–3373.

Stevenson, Megan. 2018. "Assessing Risk Assessment in Action." *Minnesota Law Review*, 103: 303–384.

Stevenson, Megan T, and Jennifer L Doleac. 2024. "Algorithmic Risk Assessment in the Hands of Humans." *American Economic Journal: Economic Policy*, 16(4): 382–414.

White, Robert C. 2011. "Re: SOP 10.1, Enforcement – Revised General Order #11-013." Memorandum from Chief of Louisville Metro Police Department, June 2.

Zeng, Zhen. 2023. "Jail Inmates in 2022 – Statistical Tables." Report, Bureau of Justice Statistics.

Main Appendix

A.1 Kentucky Pretrial Risk Assessment (KPRA)

After March 2011: Table A.1 summarizes the construction of the KPRA score. The tool comprised 12 "yes"/"no" questions, with each response assigned a fixed point value. Pretrial officers summed the points across all 12 items to produce a total risk score ranging from 0 (lowest risk) to 24 (highest risk). Officers then translated these scores into three risk levels that were reported to judges: scores of 0-5 were classified as "low risk," 6-13 as "moderate risk," and 14-24 as "high risk."

Table A.1: KPRA Factors and Weights (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).

Before March 2011: Table A.2 summarizes how the KPRA was constructed prior to March 18, 2011. Compared with the version used in the main analysis, this earlier KPRA included an additional item (Item 0, concerning references) and assigned different point values to seven of the questions. The mapping from scores to risk levels also differed: scores of 0-5 were classified as "low risk," 6-12 as "moderate risk," and 13-23 as "high risk." Because this earlier KPRA differs in its risk factors, weights, and risk-level definitions, I restrict the study period to ensure that algorithmic predictions remain constant.

Table A.2: KPRA Factors and Weights (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 Kentucky Compared with Other US Bail Settings

The Kentucky bail setting has several institutional features that distinguish it from other US jurisdictions. First, whereas pretrial services are administered locally in many states – requiring county-level data collection – Kentucky operates a single statewide pretrial services agency serving all 120 counties. This structure allows me to use uniform administrative data for the entire state, which I describe in Section 3.

Second, initial bail decisions in Kentucky are made through phone conversations between pretrial officers and judges rather than during in-person bail hearings, which are more common elsewhere in the United States.[30] Because bail decisions are made over the phone, defendants are not present.

Third, police in Kentucky have full authority to file charges, meaning there is no prosecutorial review prior to the judge's initial bail decision. Consequently, judges' bail decisions are not conditioned on prosecutors' actions.

Finally, Kentucky prohibits commercial bail bonds and, as of 2022, is one of four US states with such a ban (Cornell Law School Legal Information Institute, 2022). Defendants who cannot afford money bail therefore cannot rely on private bail bonds agents for release.[31]

[30] Kentucky has used phone-based pretrial decision-making since 1976 because geographic dispersion within the state would make routine in-person hearings costly in terms of travel time (KPS, 2019).

[31] If money bail has not been posted within 24 hours, the pretrial officer notifies the court, and the judge may reconsider the bail conditions. If detention continues, bail is typically reconsidered at the first appearance.

A.3 Theoretical Model and Testable Predictions

This appendix presents a simple model illustrating two channels through which algorithmic recommendations may affect judicial decisions. First, recommendations may convey predictive information. Second, recommendations may change the perceived losses associated with judicial decisions.[32] These channels generate distinct empirical predictions.

Judges observe a vector of case characteristics X, an algorithmic risk level r^A, and—after HB463—an algorithmic recommendation R. Judges choose bail to minimize expected loss. All expectations and loss components are evaluated by the judge conditional on the information available at the time of the decision, with conditioning on X and r^A made explicit only where it plays a substantive role.

A.3.1 Status Quo (No Recommendations)

Judges choose between *harsh bail* ($b = h$, money bail) and *lenient bail* ($b = l$, no money bail). Let $d \in \{0, 1\}$ denote whether pretrial detention occurs, and let $m \in \{0, 1\}$ denote whether misconduct occurs following release.

In the absence of recommendations, judge J minimizes

$$\mathbb{E}^J \left[\ell(b, m, d) \mid X, r^A \right],$$

where $\ell(\cdot)$ summarizes losses that depend on realized outcomes.

Under harsh bail $(b = h)$, the judge does not directly control whether the defendant is detained – detention occurs if the defendant does not post their money bail. As a result, judges form beliefs about the probability of detention $Pr^J(d = 1 \mid b = h)$, and detention generates loss $\ell(d = 1 \mid b = h) > 0$.

If the defendant is released despite harsh bail, subsequent misconduct does not generate additional loss because, conditional on choosing harsh bail, adverse outcomes do not signal that the judge's decision was overly lenient. Thus,

$$\mathbb{E}^J[\ell(h, m, d)] = Pr^J(d = 1 \mid b = h)\, \ell(d = 1 \mid b = h).$$

[32] Throughout this appendix, I use the term "loss" to denote a general penalty associated with a judicial decision or its outcomes. This object corresponds to the decision-maker costs discussed in the main text.

Under lenient bail ($b = l$), detention is impossible. If misconduct occurs, the judge incurs outcome-based loss $\ell(m = 1 \mid b = l) > 0$. Judges form beliefs about the probability of misconduct based on case observables X and the algorithmic risk level r^A, which is a coarsened mapping of the algorithm's prediction $Pr^A(m = 1 \mid b = l, X)$. Let these beliefs be denoted by $Pr^J(m = 1 \mid b = l, X, r^A)$. Expected outcome-based loss under lenient bail is

$$\mathbb{E}^J[\ell(l, m, d)] = Pr^J(m = 1 \mid b = l, X, r^A)\, \ell(m = 1 \mid b = l).$$

Accordingly, judges choose lenient bail whenever

$$\frac{Pr^J(m = 1 \mid b = l, X, r^A)}{Pr^J(d = 1 \mid b = h)} \leq \frac{\ell(d = 1 \mid b = h)}{\ell(m = 1 \mid b = l)},$$

and harsh bail otherwise.

A.3.2 Adding Algorithmic Recommendations

After HB463, judges also receive an algorithmic recommendation

$$R = \mathbf{1}\{r^A \in \{\text{low, moderate}\}\}, \tag{2}$$

where $R = 1$ denotes a recommendation to set lenient bail and $R = 0$ denotes no recommendation. Because R is a deterministic function of risk level r^A, it is measurable with respect to r^A.

With recommendations, judges minimize expected loss of the form

$$\mathbb{E}^J\left[\ell(b, m, d; R) \mid X, r^A\right] + \phi(b; R),$$

where $\ell(\cdot; R)$ captures outcome-based losses (which may depend on R) and $\phi(b; R)$ captures decision-based losses that depend on the choice and the recommendation, regardless of realized outcomes.

A.3.3 Two Theories and Their Testable Predictions

Theory 1: Recommendations only convey predictive information. If recommendations affect decisions only by conveying predictive information, then (i) outcome-based and

decision-based losses are unchanged, and (ii) beliefs satisfy

$$Pr^J(m = 1 \mid b = l, X, r^A, R) = Pr^J(m = 1 \mid b = l, X, r^A).$$

Since judges already observe r^A, recommendations have no effect on the decision rule. This theory predicts no effect of recommendations on bail decisions.

Theory 2: Recommendations change losses. Alternatively, recommendations may change losses through two channels.

(i) Cover / accountability channel (outcome-based): Following a recommendation may reduce outcome-based losses when misconduct occurs after lenient bail:

$$\ell(m = 1 \mid b = l; R = 1) < \ell(m = 1 \mid b = l; R = 0) = \ell(m = 1 \mid b = l).$$

Intuitively, a lenient recommendation may provide reputational cover for judges. If misconduct occurs following lenient bail, judges can attribute their decision to compliance with a formal recommendation rather than individual discretion.[33]

(ii) Default / compliance channel (decision-based): Recommendations may also act as defaults or soft constraints, creating decision-based losses from deviating from the recommended action, regardless of whether misconduct occurs. Under this channel,

$$\phi(h; R = 1) > \phi(l; R = 1), \qquad \phi(b; R = 0) = 0 \text{ for } b \in \{h, l\}.$$

Intuitively, it is more cognitively costly for judges to set harsh bail when lenient bail is the recommended action.

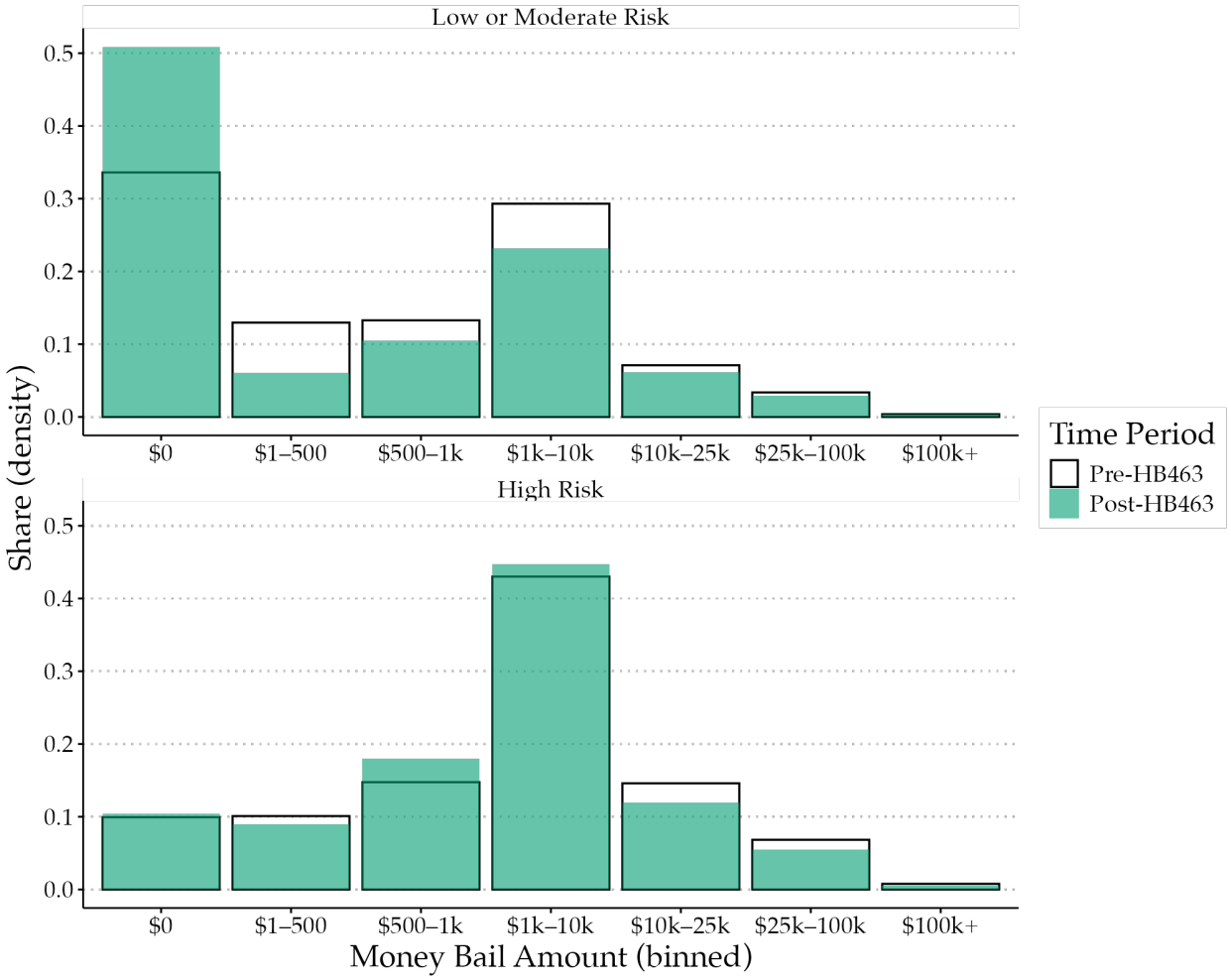
Under either mechanism, the recommendations reduce judges' expected losses from setting lenient bail. Accordingly, this theory predicts an increase in lenient bail for low and moderate risk cases, driven either by reduced perceived downside risk (cover) or by increased costs of deviating from the recommended action (default).

[33] Conversely, deviating from recommendations may increase perceived scrutiny. For example, a district attorney faced calls for removal after setting low bail for a defendant who later committed a lethal crime, in part because the decision was viewed as inconsistent with the defendant's algorithmic risk assessment (Fung, 2021).

Supplemental Appendix

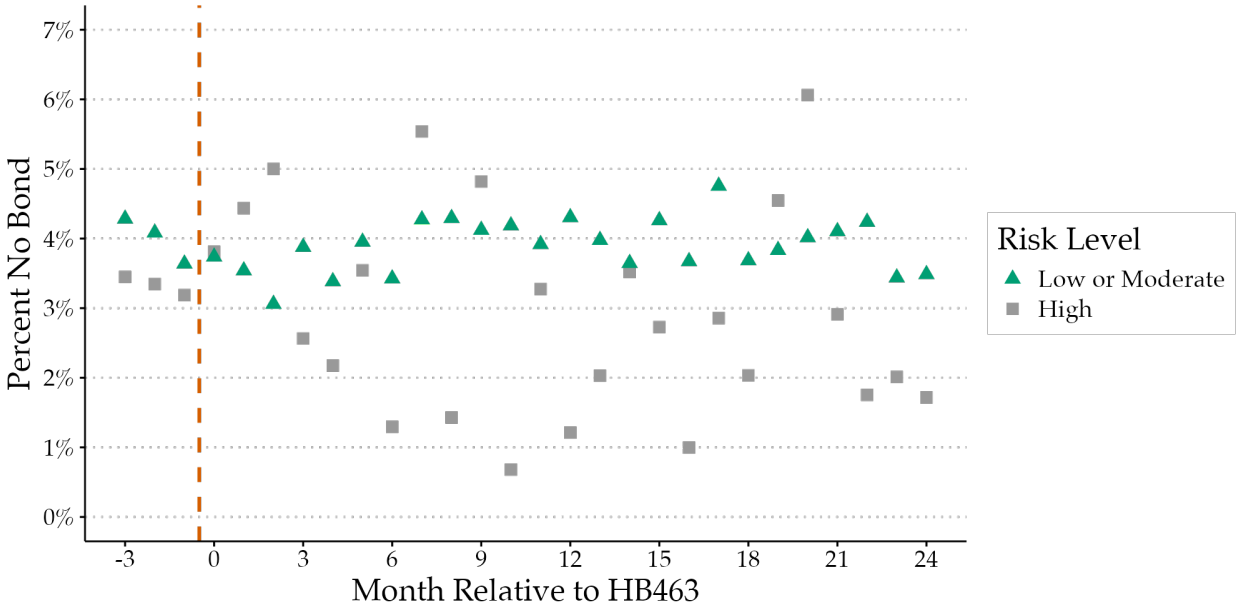
B.1 Supplementary Figures and Tables

Figure B.1: Bail Amounts by Risk Level over Time



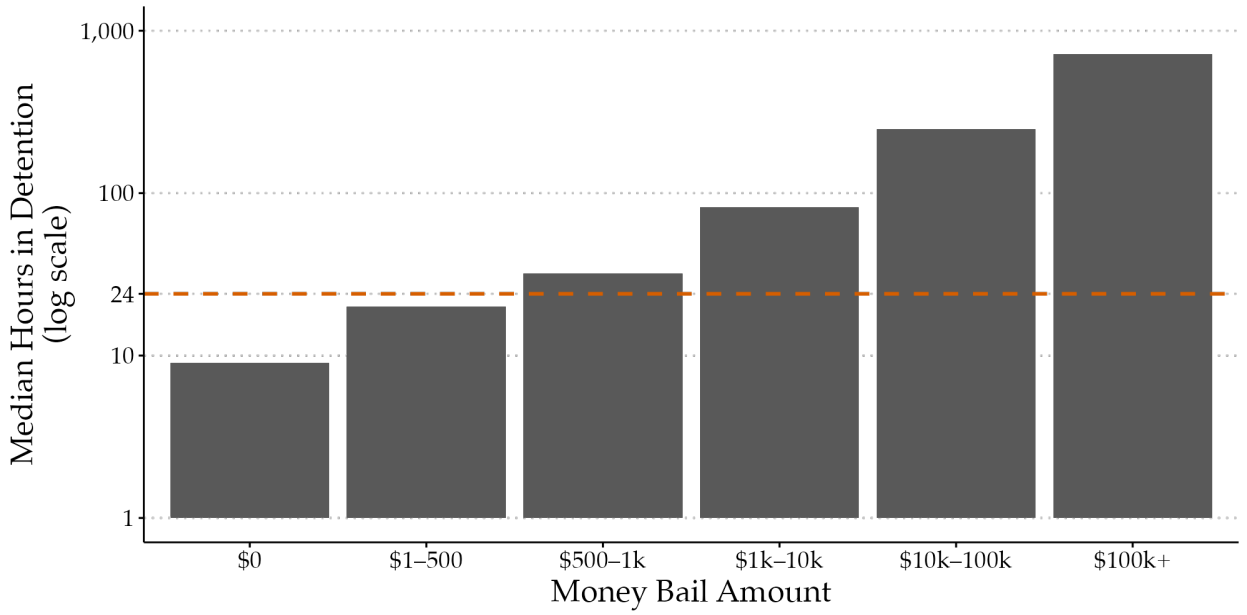
Notes: This figure shows the distribution of continuous bail amounts for low or moderate risk cases and for high risk cases, separately for the pre- and post-HB463 periods. A bail amount of $0 corresponds to lenient bail. Observations with no bond are excluded; these cases are rare.

Figure B.2: No Bond Rates by Risk Level over Time



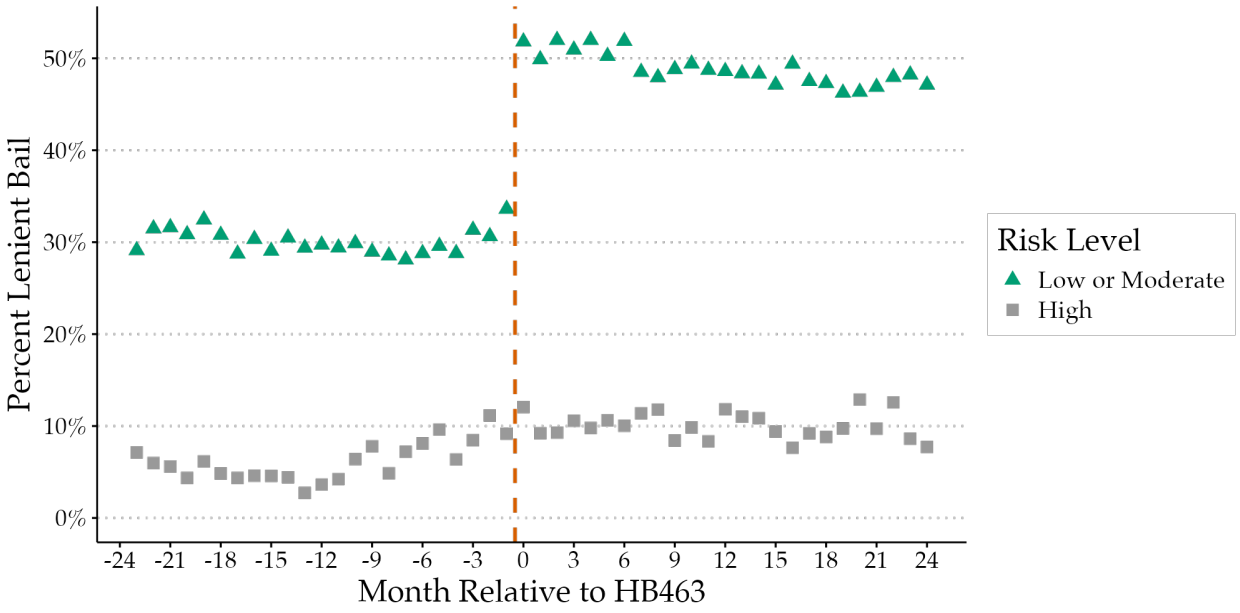
Notes: This figure shows the percentage of cases receiving no bond (outright detention) over time, split by risk level groups. Months are indexed relative to the introduction of algorithmic recommendations. Cases with a low or moderate risk level (risk scores below 14) are shown as green triangles, while cases with a high risk level (risk scores at or above 14) are shown as gray squares. The orange dashed line shows when HB463 went into effect.

Figure B.3: Median Pretrial Detention Time by Bail Amounts



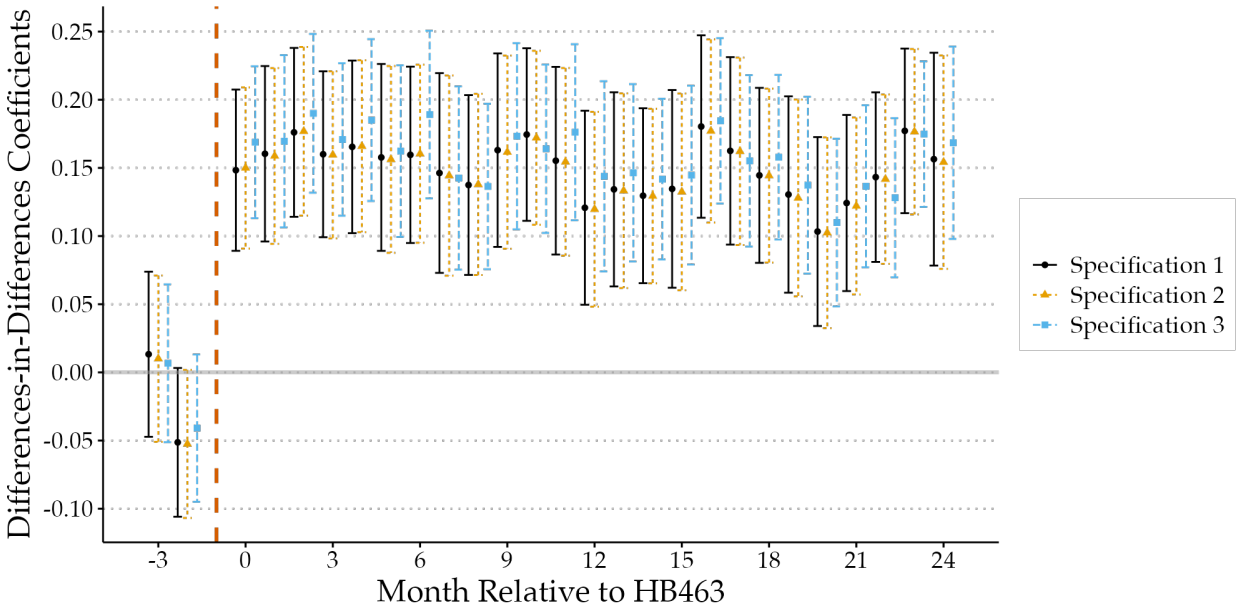
Notes: This figure shows the median number of hours in pretrial detention by bail amount bin (in dollars). The y-axis is plotted on a log10 scale to accommodate the right-skewed distribution of detention hours. The dashed orange line marks 24 hours (one day) of pretrial detention.

Figure B.4: Lenient Bail Rates by Risk Level over Time (Longer Pre-Period)



Notes: This figure shows the percentage of cases receiving lenient bail over time, split by risk level groups. Months are indexed relative to the introduction of algorithmic recommendations. Cases with a low or moderate risk level (risk scores below 14) are shown as green triangles, while cases with a high risk level (risk scores at or above 14) are shown as gray squares. The orange dotted line shows when HB463 went into effect. The construction of the KPRA risk score slightly changed in March 2011 (see Appendix A.1). To ensure that risk scores and risk levels are comparable over time, I exclude observations before March 2011 from the main analysis sample, as described in Section 3.

Figure B.5: Dynamic Differences-in-Differences Estimates across Specifications



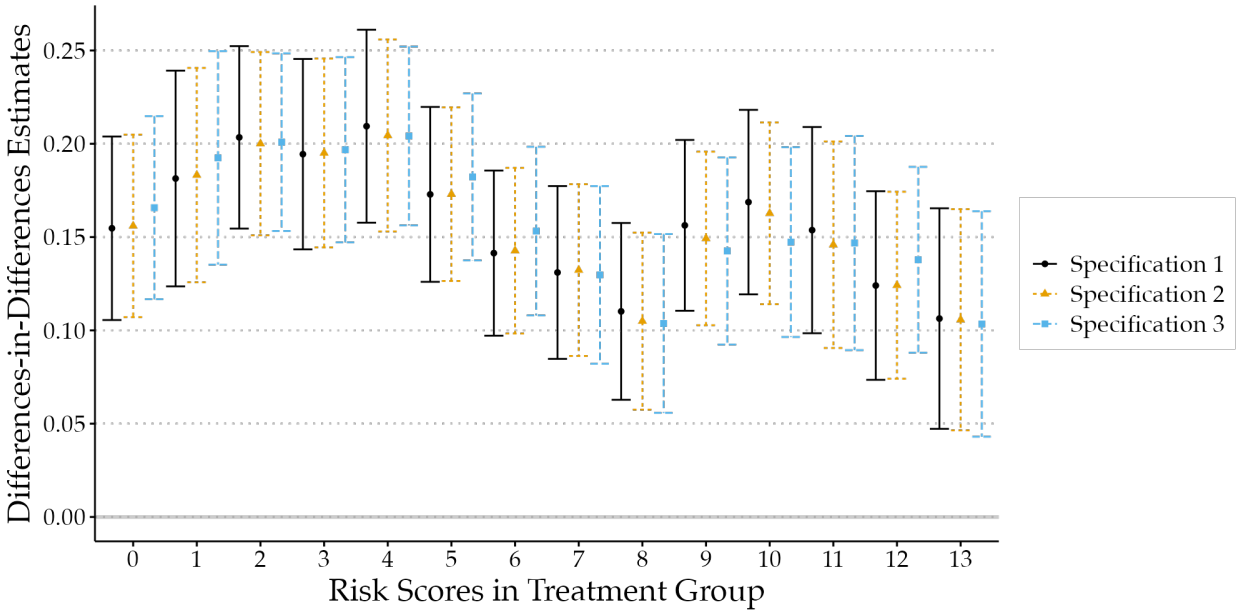
Notes: This figure shows the difference-in-differences coefficients for months relative to recommendation introduction across specifications. The outcome variable is the binary variable for lenient bail. The control group consists of cases with high risk scores, and the treated group consists of cases 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.

Table B.1: Differences-in-Differences Results across Specifications

	Dependent variable: I(lenient bail)		
I(score<14) x Post	0.167***	0.167***	0.172***
	(0.023)	(0.022)	(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	0.289	0.285	0.133
Adjusted R^2	0.285	0.282	0.132
Note:		*p<0.1; **p<0.05; ***p<0.01	

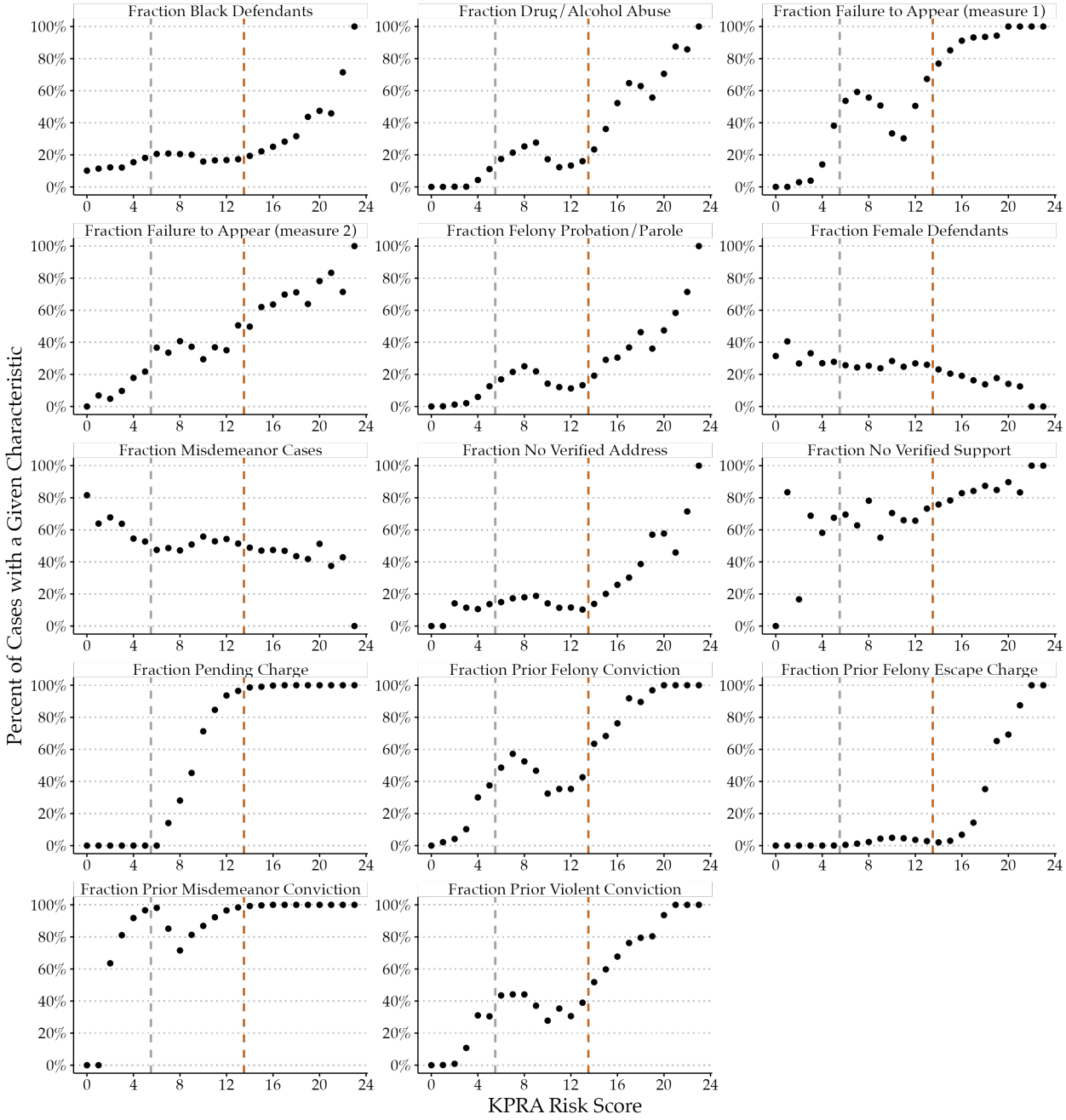
Notes: This table displays estimated differences-in-difference coefficients in specifications with lenient bail as the dependent variable. The control group consists of cases with high risk levels, and the treated group consists of cases with low or moderate risk levels. The table shows results across 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 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 B.6: Pooled Differences-in-Differences Estimates across Risk Score Values and Specifications



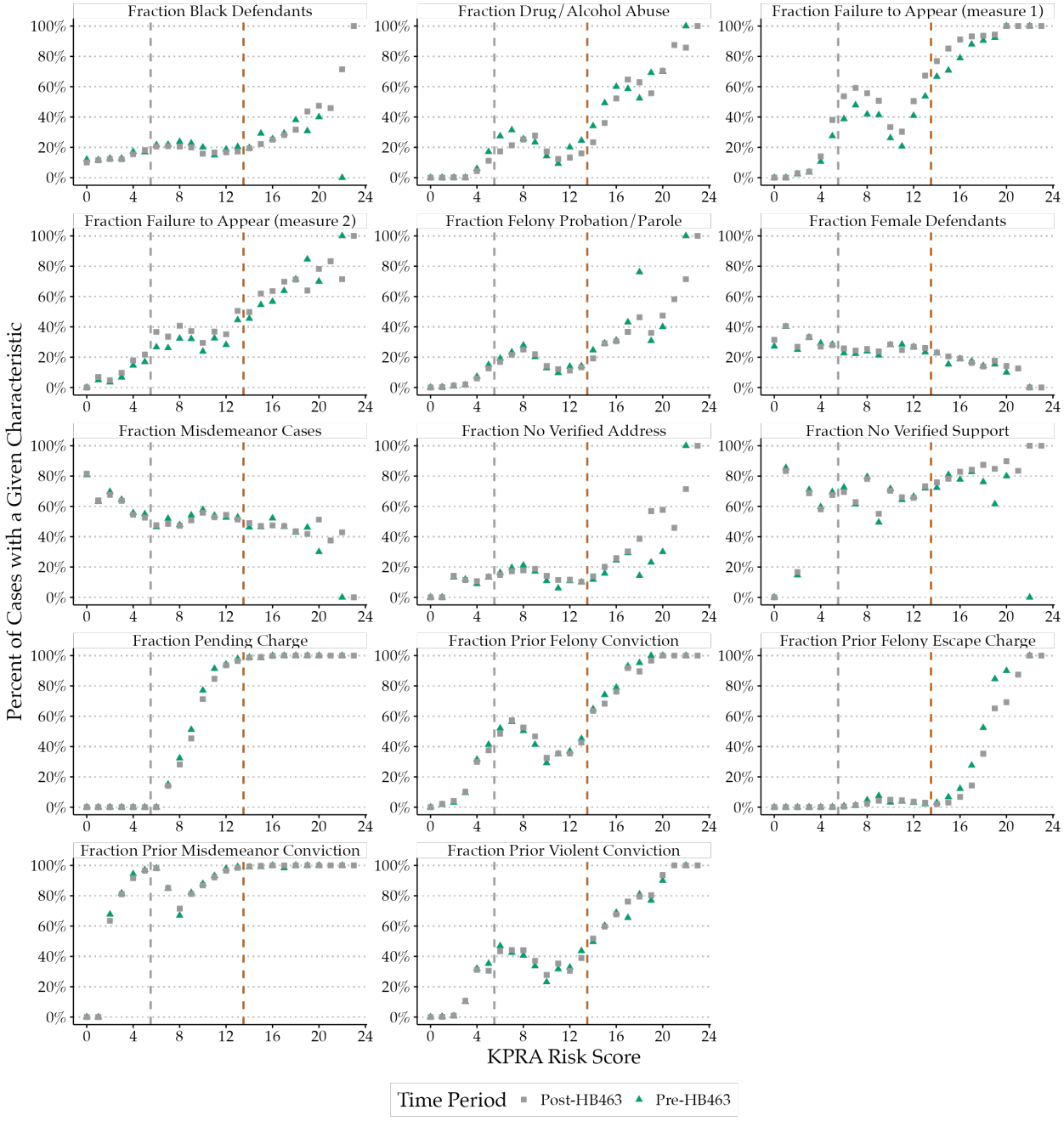
Notes: This figure shows the pooled difference-in-differences coefficients across different treatment groups based on risk scores and across different specifications. The outcome variable is the binary variable for lenient bail. The control group consists of cases with high risk scores, and the treated 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.7: Defendant and Case Covariates over Risk Score Distribution (Post-Period)



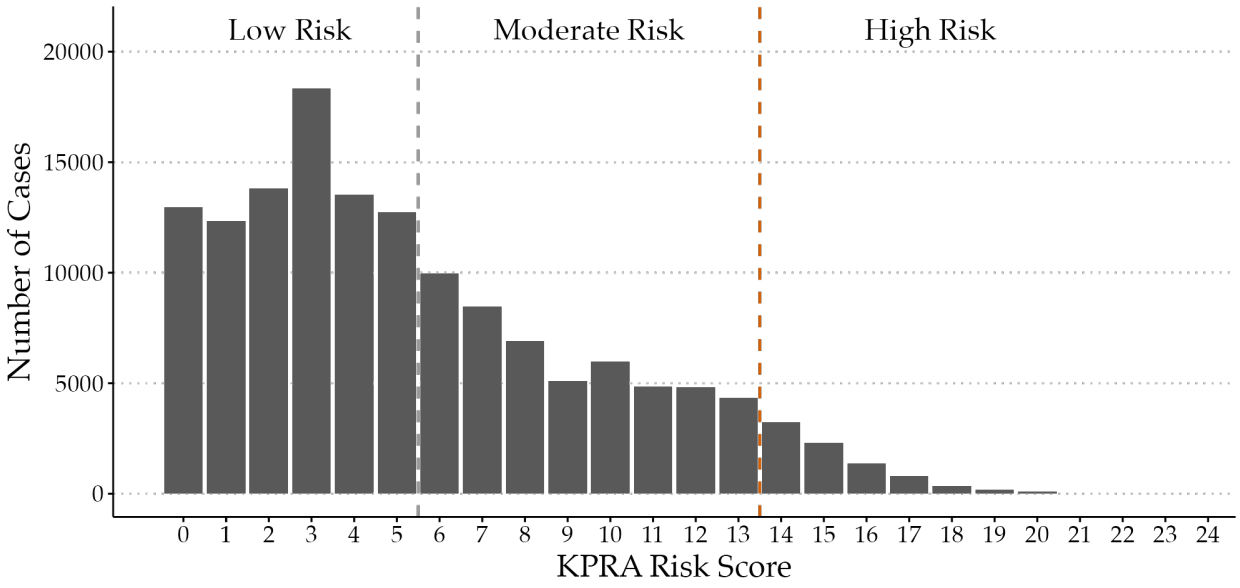
Notes: This figure shows the average defendant and case covariates for each discrete case risk score using data from the post-period. The dashed lines indicate the cutoffs between risk levels. The orange line is the threshold between "moderate" and "high" risk, and the gray line is the threshold between "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 B.8: Defendant and Case Covariates over Risk Score Distribution and Time Periods



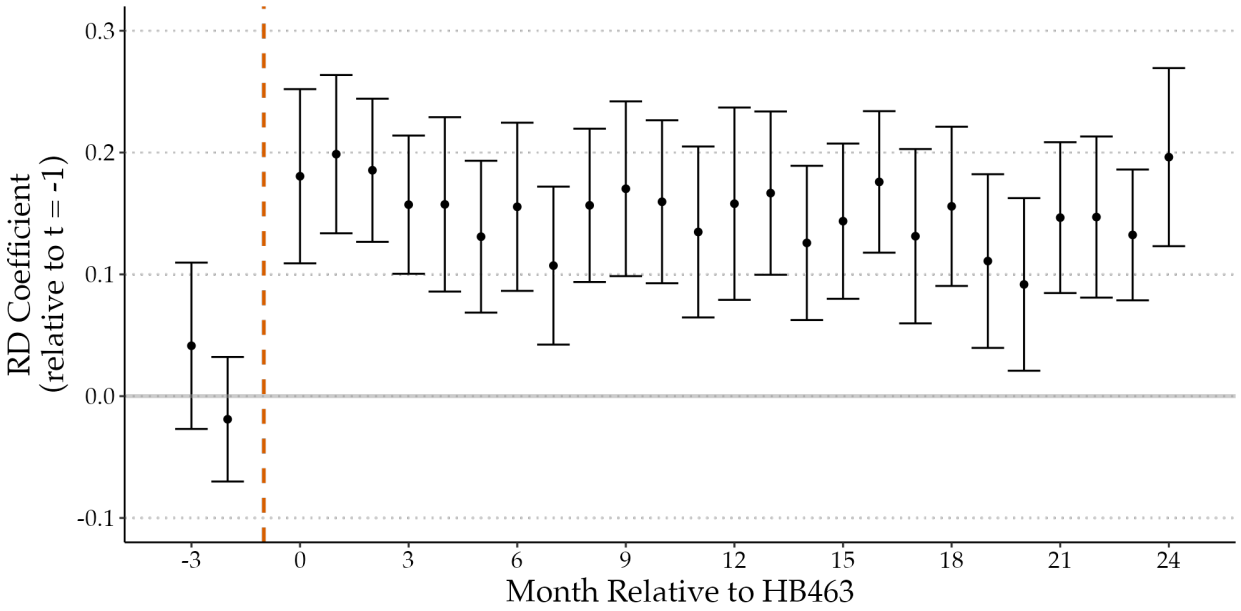
Notes: This figure shows each discrete case risk score's average defendant and case covariates. The gray rectangles show the averages before HB463, while the green triangles show the averages after HB463. The dashed lines indicate the cutoffs between risk levels. The orange line is the threshold between "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 B.9: The Risk Score Distribution



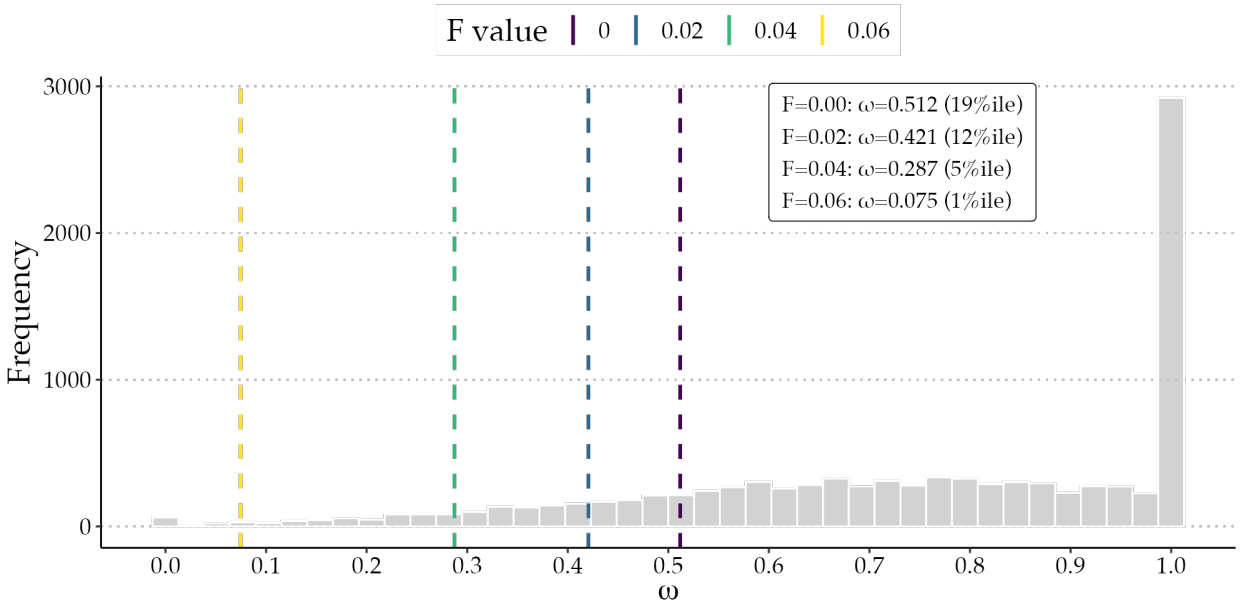
Notes: This histogram demonstrates the number of cases across the full risk score distribution. The dashed lines indicate the cutoffs between risk levels. The orange line is the threshold between "moderate" and "high" risk, and the gray line is the threshold between "low" and "moderate" risk. Scores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

Figure B.10: Regression Discontinuities over Time (Moderate-High Threshold)



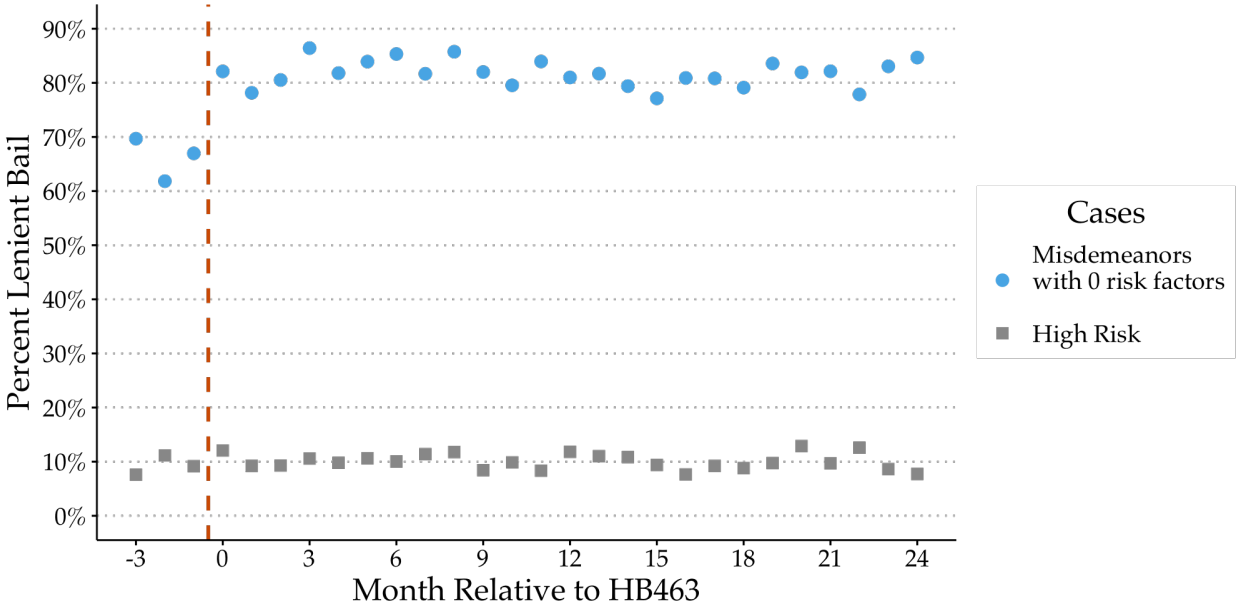
Notes: This figure presents an event-study-style regression discontinuity at the moderate-high risk threshold, estimated using a pooled sample within a fixed bandwidth around the cutoff. The bandwidth is fixed based on the optimal local-linear bandwidth for the full sample following Calonico, Cattaneo and Farrell (2020). The specification includes month fixed effects and interactions between the left-of-cutoff indicator and event-time (relative-month) indicators, and standard errors are clustered by judge. Each point shows the discontinuity at the cutoff in a given month, relative to the discontinuity in the month immediately preceding the implementation of HB463.

Figure B.11: Omega Values Required for a Recommendation Effect of 0



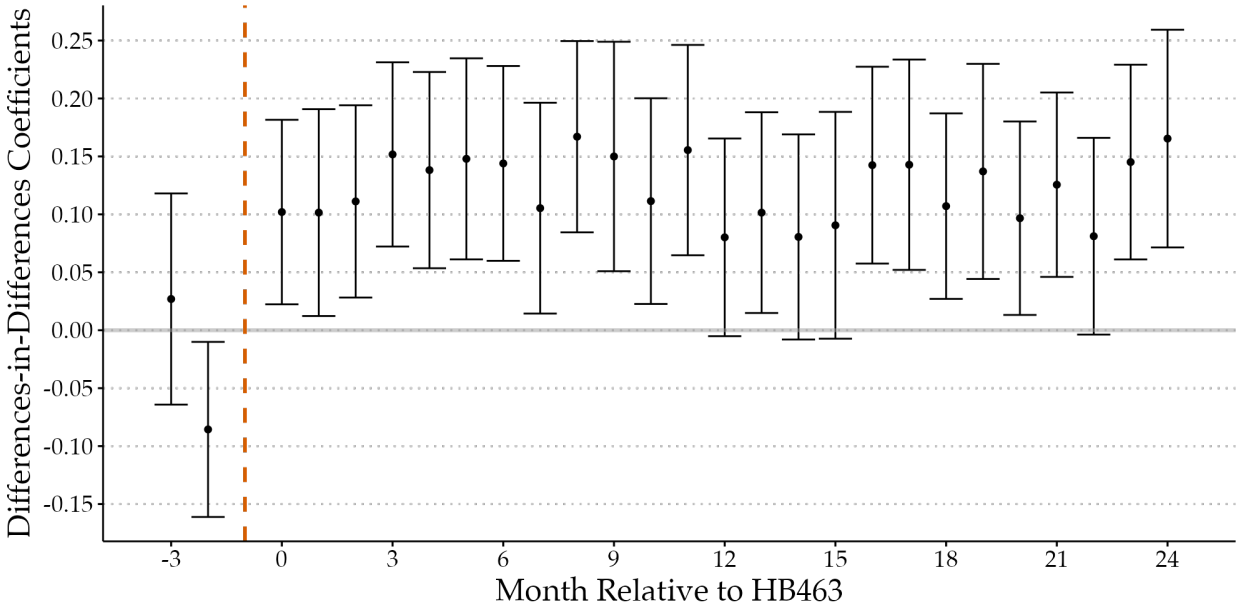
Notes: This figure is a histogram of values of ω generated from a Monte Carlo simulation with 10,000 draws. This is the same graph as Figure 8 but with ω values limited to the [0,1] range. In this case, $\omega = \min\{1, \max\{0, RD_{lm}^{\text{pre}*}/RD_{lm}^{\text{post}*}\}\}$ such that RD_{lm}^{pre*} and RD_{lm}^{post*} are normally distributed random variables based on their estimated means and standard errors. The vertical dashed lines then denote which values of ω are necessary to generate a recommendation effect of 0. There are different colors of dashed lines for different values of F, the effect of increased felony convictions.

Figure B.12: Lenient Bail Rates by Case Type over Time



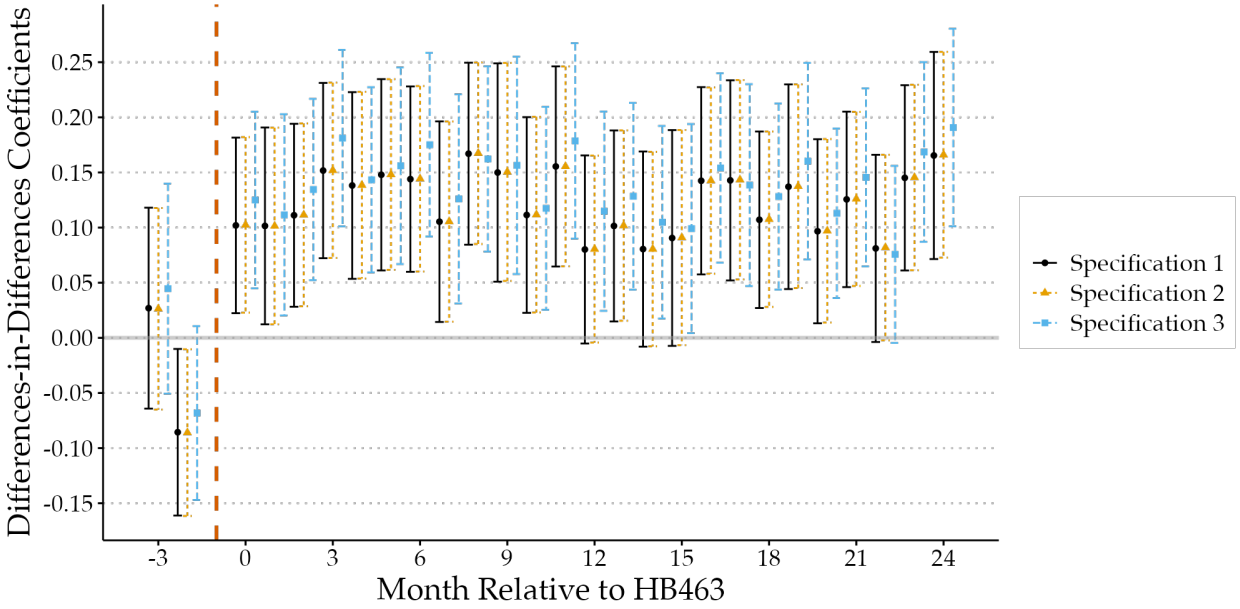
Notes: This figure shows the rate of lenient bail over months by risk score groups. Months are indexed relative to the introduction of algorithmic recommendations. Misdemeanor cases with risk scores of 0 are shown as blue circles, while cases with high risk scores are shown as gray squares. The orange dotted line shows when HB463 went into effect.

Figure B.13: 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.

Figure B.14: 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.

B.2 Heterogeneity of Recommendation Effects

This appendix examines heterogeneity in judges' responses to algorithmic recommendations across case, defendant, and judge characteristics.

B.2.1 Heterogeneity by Case and Defendant Characteristics

I first examine whether recommendation effects vary by salient case and defendant characteristics: defendant race (white or Black), defendant gender (male or female), case severity (felony or misdemeanor), and prior violent conviction (yes or no). I assess these dimensions of heterogeneity within the same differences-in-differences framework used in the main analysis.

Recall that in the differences-in-differences design, cases with low and moderate risk scores are treated because they newly receive a lenient recommendation under HB463, while high risk cases serve as a control group. Figure B.15 plots lenient bail rates over time by risk level and by each characteristic of interest. Low and moderate risk cases are shown in green and high risk cases in gray, with shapes (triangles and squares) distinguishing characteristics.

The raw trends shown in Figure B.15 suggest that changes in lenient bail following the introduction of recommendations vary across defendant characteristics in the pooled data. However, these patterns are descriptive and may reflect differences in which judges handle different types of cases.

To assess heterogeneity more formally, I estimate a series of triple-differences specifications of the form:

$$\begin{aligned}\text{lenient}_{itj} = \beta_1 \left(\mathbb{1}\{\text{score}_i < 14\} \times \text{Post}_t\right) + \beta_2 \left(\mathbb{1}\{\text{score}_i < 14\} \times Z_i\right) \\ + \beta_3 \left(\text{Post}_t \times \text{Black}_i\right) + \beta_4 \left(\mathbb{1}\{\text{score}_i < 14\} \times \text{Post}_t \times Z_i\right) + \mathbf{X}'_{ijt}\gamma + \varepsilon_{itj}, \quad (2)\end{aligned}$$

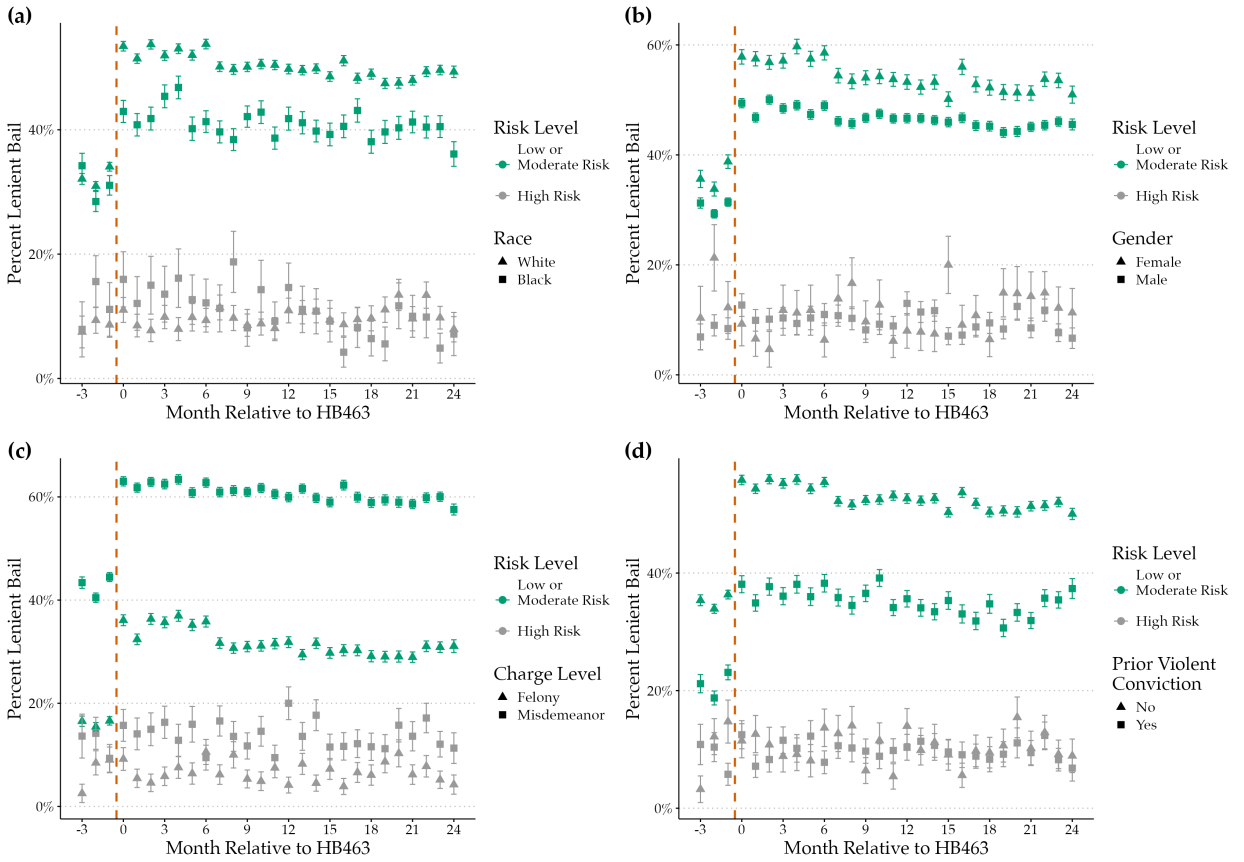
where Z_i denotes the case or defendant characteristic of interest and $\mathbf{X}_{ijt}$ includes the same controls as in main specification 1. The coefficient of interest, β_4, captures heterogeneity in the recommendation effect along the dimension indexed by Z_i.[34]

Table B.2 reports results from these pooled triple-differences models. In these specifications, recommendation effects appear to differ across some defendant characteristics. In

[34] As usual, 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 cluster standard errors by judge.

particular, pooled estimates suggest smaller effects for Black defendants and male defendants, while effects do not differ meaningfully by current charge level. Recommendation effects also appear somewhat smaller for defendants with prior violent convictions.

Figure B.15: Lenient Bail Rates over Time by Risk Level and Case/Defendant Characteristics



Notes: This figure shows the rate of lenient bail over time by risk score level as well as case and defendant characteristics of interest. Months are indexed relative to the introduction of algorithmic recommendations. The orange dashed line shows when HB463 went into effect. I illustrate cases with low or moderate risk in green and cases with high risk in gray. I use shapes (triangles and squares) to differentiate in terms of case and defendant characteristics. Panel (a) splits by defendant race, panel (b) splits by gender, panel (c) splits by charge level, and panel (d) splits by prior violent conviction.

One concern with this analysis is that judges are not randomly assigned to cases. Instead, judges tend to see systematically different case and defendant populations based on geography and court assignment. As a result, pooled heterogeneity estimates may reflect differences across judges rather than differences in how the same judge responds to recommendations across cases.

To assess this possibility, I re-estimate the triple-differences models allowing for flexible judge fixed effects that vary by period and by risk group (low or moderate vs. high

risk). This approach absorbs all judge-level differences in baseline leniency and overall responsiveness to the reform, isolating within-judge heterogeneity.

Table B.3 reports the results of these specifications. Once flexible judge fixed effects are included, most of the previously significant heterogeneity by race and gender attenuates substantially and is no longer statistically distinguishable from zero at conventional levels. This pattern suggests that much of the pooled heterogeneity reflects differences across judges in responses rather than differential responses within judges. In contrast, heterogeneity by prior violent conviction remains similar in magnitude and marginally statistically significant, indicating that judges may be less responsive to lenient recommendations when defendants have prior violent records.

Taken together, these results suggest that while recommendation effects vary across defendants in the pooled data, much of this variation reflects which judges respond more strongly to recommendations rather than how individual judges condition on defendant characteristics. The remaining heterogeneity by prior violent conviction is consistent with an interpretation that recommendations provide less effective cover in cases where potential misconduct is more salient.

Table B.2: Triple Differences Models across Case/Defendant Characteristics

	Dependent variable: I(lenient bail)			
	(1)	(2)	(3)	(4)
I(score<14) x Post	0.180***	0.248***	0.165***	0.186***
	(0.022)	(0.040)	(0.026)	(0.028)
I(score<14) x Black	0.033			
	(0.033)			
Post x Black	0.004			
	(0.035)			
I(score<14) x Post x Black	−0.083**			
	(0.037)			
I(score<14) x Male		0.046		
		(0.038)		
Post x Male		0.088**		
		(0.035)		
I(score<14) x Post x Male		−0.102***		
		(0.038)		
I(score<14) x Misdemeanor			0.146***	
			(0.026)	
Post x Misdemeanor			0.022	
			(0.028)	
I(score<14) x Post x Misdemeanor			−0.004	
			(0.028)	
I(score<14) x I(Prior Vio Convic)				0.050**
				(0.024)
Post x I(Prior Vio Convic)				0.019
				(0.022)
I(score<14) x Post x I(Prior Vio Convic)				−0.053**
				(0.025)
Avg Dep Var (Pre, Low/Mod, Z=0)	0.325	0.363	0.428	0.353
Time/Score FEs	Y	Y	Y	Y
Num. Judge FEs	1	1	1	1
Charge/county/demographic controls	Y	Y	Y	Y
Risk component controls	Y	Y	Y	Y
Observations	142,089	142,089	142,089	142,089
R^2	0.289	0.289	0.290	0.289
Adjusted R^2	0.286	0.286	0.287	0.286

Notes: This table reports triple differences results when taking the default pooled difference-in-difference approach and interacting with case/defendant characteristics of interest. The outcome variable is the binary variable for lenient bail. As in the main differences-in-differences approach, this specification includes controls for day of week, month-year, exact risk score, top charge level and class, defendant demographics (race, gender), judge, county, and all risk score components listed in Table A.1 except for verified address and support. I cluster standard errors by judge. The reported dependent variable mean is the pre-period mean for low or moderate risk defendants in the omitted category for the interaction in that column. *p<0.1; **p<0.05; ***p<0.01.

Table B.3: Triple Differences Models across Case/Defendant Characteristics with Flexible Judge FEs

	Dependent variable: I(lenient bail)			
	(1)	(2)	(3)	(4)
I(score<14) x Black	−0.006			
	(0.033)			
Post x Black	−0.002			
	(0.031)			
I(score<14) x Post x Black	−0.017			
	(0.034)			
I(score<14) x Male		0.007		
		(0.042)		
Post x Male		0.047		
		(0.040)		
I(score<14) x Post x Male		−0.055		
		(0.043)		
I(score<14) x Misdemeanor			0.147***	
			(0.025)	
Post x Misdemeanor			0.027	
			(0.024)	
I(score<14) x Post x Misdemeanor			−0.025	
			(0.025)	
I(score<14) x I(Prior Vio Convic)				0.034*
				(0.020)
Post x I(Prior Vio Convic)				0.013
				(0.020)
I(score<14) x Post x I(Prior Vio Convic)				−0.040*
				(0.022)
Avg Dep Var (Pre, Low/Mod, Z=0)	0.325	0.363	0.428	0.353
Time/Score FEs	Y	Y	Y	Y
Num. Judge FEs	4	4	4	4
Charge/county/demographic controls	Y	Y	Y	Y
Risk component controls	Y	Y	Y	Y
Observations	142,089	142,089	142,089	142,089
R^2	0.300	0.299	0.300	0.299
Adjusted R^2	0.284	0.283	0.284	0.283

Notes: This table reports triple differences results when taking the default pooled difference-in-difference approach and interacting with case/defendant characteristics of interest. In this version, I allow for flexible judge fixed effects. The outcome variable is the binary variable for lenient bail. As in the main differences-in-differences approach, this specification includes controls for day of week, month-year, exact risk score, top charge level and class, defendant demographics (race, gender), judge, county, and all risk score components listed in Table A.1 except for verified address and support. I cluster standard errors by judge. The reported dependent variable mean is the pre-period mean for low or moderate risk defendants in the omitted category for the interaction in that column. *p<0.1; **p<0.05; ***p<0.01.

B.2.2 Heterogeneity by Judge Characteristics

I next examine heterogeneity in responses to recommendations across judges. The full sample includes 423 judges, though many judges make few bail decisions. To obtain reliable judge-level estimates, I focus on the subset of judges who decided at least 50 low or moderate risk cases both before and after HB463. This yields 110 judges, whose decisions together account for approximately 76% of all low and moderate risk case decisions.

Focusing on low and moderate risk cases, I estimate a model with judge-by-period fixed effects:

$$\text{lenient}_{ijt} = \mathbf{X}'_{ijt}\gamma \; + \; \alpha_{j \times Post} \; + \; \varepsilon_{ijt}, \tag{3}$$

where the term $\gamma_{j \times Post}$ denotes judge-by-period fixed effects, allowing each judge to have a distinct intercept in the pre- and post-HB463 periods.[35]

I then compute, for each judge j, the change in the fitted probability of lenient bail between the pre- and post-HB463 periods:

$$\Delta_j = \mathbb{E}\left[\widehat{\text{lenient}}_{ijt} \mid j, \text{Post}\right] - \mathbb{E}\left[\widehat{\text{lenient}}_{ijt} \mid j, \text{Pre}\right],$$

where $\widehat{\text{lenient}}_{ijt}$ denotes the fitted value from equation (3). The quantity Δ_j measures how much judge j's propensity to choose lenient bail changed after HB463, relative to their own pre-reform behavior, holding case characteristics and time effects constant.

I then examine which judge- and county-level characteristics are correlated with these estimated responses in a descriptive regression framework. I collect information on judge demographics (race, gender, years of experience), election competition (whether the judge or any judge in the district faced a contested election in 2010, and the size of the electorate), and county characteristics (population, rural status, and crime rates).[36]

Table B.4 reports results from regressions of Δ_j on these characteristics. Judges who see a higher share of Black defendants exhibit significantly smaller increases in lenient bail following the reform. Seeing a defendant population that is 10 percentage points more

[35] As usual, lenient_{itj} is an indicator for if the bail for case i at time t decided by judge j is lenient (no money bail), $\mathbf{X}_{ijt}$ includes the same controls as in main specification (1), and I cluster standard errors by judge.

[36] Judge demographics and years of experience were collected from publicly available biographical information and conversations with clerks and Administrative Office of the Courts staff. Information on judicial elections – including whether a race was contested and the number of voters – comes from public data on the Kentucky 2010 general election results. County population and rural status are based on the 2010 Census, and county-level crime rates are drawn from the FBI's Uniform Crime Reports. (Since judges may make decisions for more than one county, I focus on the county where each judge made most of their decisions.) I successfully collected complete judge-level information for 93 of the 110 judges included in this analysis.

Black is associated with increasing lenient bail rates by 4.3–5.4 fewer percentage points. This relationship is robust to controlling for judge demographics, election exposure, pre-reform misconduct rates, and county-level crime characteristics. Other judge characteristics – such as experience, gender, race, and election-related variables – are not systematically associated with responsiveness to recommendations.[37]

These results are descriptive but consistent with an interpretation in which recommendations alter the perceived costs of errors in a way that varies across judges and environments. One possible interpretation is that recommendations provide less effective cover in settings where potential scrutiny of lenient decisions is higher. This interpretation is broadly consistent with evidence in related work showing links between racial heterogeneity and punishment severity (Feigenberg and Miller, 2021). Overall, these findings suggest that recommendation effects may operate unevenly across institutional contexts, even when the recommendations themselves are uniform.

[37] Moreover, the share of Black defendants alone explains a substantial portion of the variation in judge responses ($R^2 = 0.22$), and the inclusion of additional judge and county characteristics does not increase the adjusted R^2.

Table B.4: Explaining Judge Responsiveness to Lenient Recommendations

	Dependent Variable = Judge x Post FE				
	(1)	(2)	(3)	(4)	(5)
Share Black Defendants	−0.435***	−0.443***	−0.494***	−0.468**	−0.543***
	(0.086)	(0.090)	(0.154)	(0.178)	(0.188)
Black Judge		0.034	0.054	0.052	0.063
		(0.071)	(0.074)	(0.076)	(0.077)
Woman Judge		−0.020	−0.015	−0.017	−0.020
		(0.025)	(0.028)	(0.029)	(0.029)
Years as Judge		−0.002	−0.003	−0.003	−0.002
		(0.002)	(0.002)	(0.002)	(0.002)
Contested in 2010			−0.045	−0.048	−0.040
			(0.035)	(0.037)	(0.037)
Any Contest in District in 2010			0.009	0.008	−0.005
			(0.038)	(0.039)	(0.039)
log(Election Voters)			0.007	0.018	0.006
			(0.024)	(0.043)	(0.043)
log(County Population)				−0.012	−0.034
				(0.044)	(0.046)
Rural County				−0.004	−0.014
				(0.046)	(0.048)
Total Crime Rate					−0.0001
					(0.00005)
Total Index Crime Rate					0.010
					(0.008)
Property Crime Rate					−0.010
					(0.008)
Violent Crime Rate					−0.009
					(0.008)
Constant	0.254***	0.281***	0.232	0.262	0.588*
	(0.019)	(0.027)	(0.213)	(0.302)	(0.338)
N	93	93	93	93	93
R^2	0.219	0.240	0.256	0.257	0.313
Adjusted R^2	0.211	0.205	0.195	0.177	0.200

Notes: This table shows the estimated coefficients from regressing Δ_j estimates, as estimated in equation (3), on judge- and county-level characteristics. Judge-level characteristics include share of Black defendants seen by the judge, whether the judge is Black, whether the judge is a woman, number of years of experience as a judge (as of 2011), whether the judge was contested in their 2010 election, whether the judge was elected in a district where any judge faced a contest in 2010, number of voters in the judge's election, the judge's pre-HB463 rearrest rate, and the judge's pre-HB463 failure to appear rate. County-level characteristics refer to where a judge made the most decisions in the data (because judges can make decisions in multiple counties). County-level characteristics include county population, whether the county is rural (under 50,000 people), total crime rate, total index crime rate, property crime rate, and violent crime rate. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.