The Case Assignment System as a Technology: Algorithmic Allocation and Its Impact on Court

Rui Zuo*

September 15, 2025
Click here for the latest version.

Abstract

This paper exploits a nationwide reform in China where courts replaced manual case assignment with algorithmic systems: either random or machine learning (ML)-based. Using 66 million court documents from 2014-2021, I estimate effects on judge-case matching and performance. Random assignment reduces path dependence but weakens expertise matching, leading to modest performance declines. ML assignment improves matching and reduces path dependence while maintaining performance levels. Using a subsample of dangerous driving cases, I find both systems increase sentencing variance and leniency. These findings demonstrate that seemingly neutral administrative technologies embed consequential trade-offs requiring careful design.

Court decisions permeate almost all aspects of society and daily life, from family matters and social norms to the enforcement of the rule of law, making court performance a fundamental determinant of societal well-being. Understanding what drives variation in court decisions and performance has therefore attracted sustained interest across fields. Increasing evidence suggests that judges' characteristics: their ideology, identity, experience, and training, interact with case characteristics to create measurable variation in judicial outcomes[1]. If this judge-case matching generates differences in court performance, a natural follow-up question arises: What about the systems that create these matches? Every court must decide who hears which case, yet the case assignment process is often treated as a neutral infrastructure that is fixed, procedural, and outside the scope of analysis. However, if judge-case assignment influences court performance, as a growing body of research suggests, then the system generating these assignments may itself influence performance.

This paper asks: Do case assignment systems affect case assignment patterns, court performance, and sentencing disparity? I approach this question by studying a nationwide reform in China, where courts transitioned from manual assignment to two types of algorithmic systems: random assignment or machine learning-based assignment. Using 66 million court documents from 2014-2021, I estimate how different assignment technologies on (1) assignment patterns measured by judge-case expertise matching and the alignment between judge characteristics and case complexity; (2) court performance captured through case processing efficiency and verdict quality; and (3) sentencing disparity decomposed to isolate the judge behavioral changes from the change in case composition.

*University of Texas at Austin; Massachusetts Institute of Technology; ruizuo11@mit.edu.

[1] An extensive literature examines how judge characteristics, including ideology, gender, race, personality traits, and even external factors like home football team performance, influence judicial decisions (Eren and Mocan, 2018; Shayo and Zussman, 2011; Boyd et al., 2010; Harris and Sen, 2019; Glynn and Sen, 2015).

Each case assignment approach, manual, random, and ML-based, presents distinct strengths and weaknesses. Manual assignment, for example, allows for nuanced, human-observed case- and judge-specific information, potentially leading to tailored pairings. However, this reliance on humans is susceptible to inconsistency or discretion. Random assignment, as a rule-based system, effectively minimizes such discretion. Yet, by its design, it disregards specific judge and case characteristics, which can result in inappropriate or inefficient judge-case pairings (e.g., assigning family law cases to judges who primarily handle patent cases). Theoretical models suggest that even when judges are impartial, random assignment may reduce judicial effort and increase errors, as judges lose interest when assigned cases outside their areas of expertise or preference (Hübert, 2021). ML-based assignment, a data-driven alternative, seeks to optimize pairings by learning from historical outcome data. This approach, however, might inherit the discretions present in its training data and fail to account for unquantified information that human decision-makers might consider. Given this coexistence of strengths and weaknesses, the net effect of each system on assignment patterns, court performance, and sentencing disparity is not self-evident.

This paper overcomes a key empirical challenge in studying assignment system effects by finding a setting that enables large-scale observation of both institutional change and court outcomes. The challenge has two dimensions. First, identifying when and how assignment systems change: OECD countries with comprehensive court records have used random assignment for decades, providing no variation to study, while developing countries with flexible assignment systems often lack documentation of when and how changes occur. I solve this by leveraging China's nationwide court reform, where I can pinpoint the exact timing and type of assignment system changes through government procurement contracts. Second, observing assignment outcomes and court performance: I overcome the typical data limitations by using 66 million publicly available court documents that reveal both who gets assigned to which cases (from judge names) and detailed case outcomes (from written facts and verdicts).

I exploit the staggered rollout of algorithmic assignment systems across Chinese provinces from 2014 to 2020. Courts in eight provinces adopted random assignment, while courts in 18 provinces adopted ML-based assignment, creating plausibly exogenous variation in treatment timing driven by budget and bidding process rather than court performance. Each system has distinct characteristics: manual assignment relies on senior judges selecting suitable matches using detailed information, but allowing discretion; random assignment eliminates discretion but disregards judge-case characteristics; ML-based assignment optimizes using historical data but may inherit training biases. Using generalized difference-in-differences following Callaway and Sant'Anna (2021), I compare outcomes in treated provinces with not-yet-treated provinces to identify causal effects of these different assignment technologies.

This study shows that case assignment systems are far from a neutral infrastructure; they impact case assignment patterns, court performance metrics, and sentencing outcomes in the justice system. The transition from manual to random case assignment reduced path dependence: the complexity of cases assigned to a judge became less dependent on the complexity of their past caseload. However, this procedural change did not improve the matching of judge expertise to case type (e.g., a family law judge was not more likely to be assigned a family law case). This outcome is anticipated, as random assignment, by its rule-based nature, disregards judge-specific expertise and case characteristics.

This shift in assignment patterns correlated with a slight reduction in court performance. We observe a decrease in the official performance index (-0.054 standard deviations) and in a measure of verdict quality (-0.114 standard deviations). These declines appear driven by an increase in appeal rates (approximately six additional appealed cases per court division per quarter) and decreases in verdict length (around 47 fewer words) and the number of legal provisions citing (0.096 fewer). Analysis of a subsample of dangerous driving cases reveals that judges become more lenient under random assignment, reducing probation length and penalties for probation violations, while simultaneously increasing sentencing variance across similar cases. This pattern suggests that random assignment generates inconsistent rather than principled leniency.

The transition from manual to machine learning (ML)-based case assignment yielded distinct changes in assignment patterns. Similar to random assignment, ML systems reduced path dependence in case complexity. Additionally, under ML assignment, case complexity no longer correlated with judge experience. This procedural change, however, increases the expertise-matching ratio for case assignments (i.e., cases were more likely to be assigned to judges with relevant specialization).

Despite these modifications to assignment patterns—including improved expertise matching—this did not translate into detectable gains in overall court performance. Empirical evidence indicates no significant change in either the official performance indicators or the verdict quality index. Furthermore, no significant changes were observed in the components of these indices. Analysis of dangerous driving cases reveals that ML assignment also produces more lenient sentencing, particularly in reducing penalties for probation violations, along with increased sentencing variance across similar cases. However, unlike random assignment, this leniency coincides with maintained court performance, suggesting more calculated judicial discretion.

These findings are robust to multiple checks, including alternative sample constructions (e.g., excluding courts with less public case disclosure), alternative weighting methods for the index components, and the inclusion of additional controls for judge and court characteristics.

This paper contributes to the empirical literature on judicial decision-making by examining how case assignment systems influence court outcomes. A large body of research demonstrates that heterogeneity in judicial decisions (such as bailing, paroling and sentencing) creates variation in case outcomes, with much of this work relying on random case assignment as a source of exogenous variation to identify causal effects of different judges (Kling, 2006; Agan et al., 2023; Aizer and Doyle Jr, 2015; Bhuller et al., 2020; Eren and Mocan, 2021)[2]. This approach treats assignment systems as neutral administrative infrastructure that generates useful variation without independently affecting outcomes. However, a smaller literature has begun to examine assignment systems themselves, primarily through qualitative comparisons of assignment practices across courts (Gramckow

[2]For comprehensive reviews of this literature, see Frandsen et al. (2023) and Chyn et al. (2024).

et al., 2016; Fabri and Langbroek, 2007; Macfarlane, 2023; Jin, 2020; Chilton and Levy, 2015). My paper advances this emerging area by providing the first quantitative comparison of different assignment technologies—manual, random, and machine learning-based systems—and their effects on court performance and public trust.

Second, this study contributes to the growing literature on algorithmic governance in public sector settings. As government entities increasingly use algorithms for decision-making, research has examined both custom-built and existing systems, typically comparing algorithmic performance to human decision-makers. The closest context to my setting is algorithmic use in judicial processes, where algorithms primarily assist with binary decisions such as bail, parole, and sentencing guidelines (Arnold et al., 2025; Angelova et al., 2023, 2024; Berk, 2017; Berk et al., 2016; Kleinberg et al., 2018; Stevenson and Doleac, 2024). These studies find that algorithms often outperform humans in consistency and impartiality (Ludwig et al., 2024). Similar findings appear in other algorithmic governance applications, including policing misconduct, tax auditing, and child protection (Stoddard et al., 2024; Kleinberg et al., 2015; Battaglini et al., 2024; Rittenhouse et al., 2023). My paper contributes to this literature by examining algorithms in a different task: rather than making binary decisions ("yes or no"), I study algorithms in work allocation tasks ("who gets what"). This represents a shift from algorithmic decision-making to algorithmic management that requires allocating different tasks to different workers. I find that well-designed algorithms can perform at least as effectively as human assigners, while different algorithmic approaches produce different outcomes.

Third, this paper contributes to the algorithmic management literature, where algorithms take on managerial responsibilities by assigning tasks to human workers. The algorithmic management literature primarily consists of qualitative studies focusing on gig platforms, such as food delivery or ride-sharing, where algorithms match workers to tasks in low-skill settings without human assignment as a control group (Rosenblat and Stark, 2016; Becker et al., 2023; Lee et al., 2015; Cram et al., 2020; Jarrahi et al., 2023). One related study by Hoffman et al. (2018) shows that algorithms can make better hiring decisions than humans, in a low-skill setting. My paper contributes quantitative evidence using causal inference to identify the effects of algorithmic versus human assignment in a high-stakes professional environment. This work extends algorithmic management beyond routine tasks to complex work allocation.

These findings offer insights for court reform globally. As judicial systems worldwide increasingly transition towards algorithmic case assignment, policymakers must recognize that these are not neutral administrative tools, but technologies with complex effects on court operations, performance metrics, and sentencing outcomes. The choice of system requires careful consideration of inherent trade-offs. Random assignment offers transparency and eliminates human discretion in case allocation, but reduces expertise matching, diminishes court performance, and produces inconsistent leniency in sentencing. ML-based assignment, while less transparent in its decision-making process, improves expertise matching and enables more principled judicial discretion, though without necessarily improving overall performance metrics. The broader lesson is that seemingly neutral administrative technologies embody value trade-offs that require explicit policy choices rather than purely technical solutions. As institutions increasingly rely on algorithms for resource allocation, understanding these trade-offs becomes essential for designing systems that balance efficiency, fairness, and transparency.

The remainder of the paper is structured as follows: Section 1 provides institutional background on Chinese courts and the assignment reform. Section 3 describes the data and variable construction. Section 2 outlines the empirical strategy. Section 4 describes results on assignment patterns. Section 5 presents results on court performance and robustness checks. Section 6 reports results on sentencing disparity using a subset of cases. Section 7 concludes.

1 Institutional Background

In Chinese courts, judges are the primary determinant of sentencing outcomes beyond case characteristics themselves. Without active jury participation or binding precedents, Chinese courts rely exclusively on judicial decision-making. Prior to 2014, courts used manual case assignment. Following the Supreme People's Court's publication of "Several Opinions on Improving Judicial Accountability System of the People's Courts," courts began transitioning to algorithmic assignment systems to reduce judge shopping[3]. Courts subsequently adopted either random or ML-based assignment approaches, each with distinct advantages and limitations. The following subsections detail these assignment procedures, their mechanisms, and the rationale for their implementation.

1.1 Case Assignment Procedure

As in Figure A1, each judge works in one division only and handles cases within this division. After the plaintiffs file the case, the case-filing division forwards it to the relevant division based on the topic of the case. Within each division, case is either assigned manually or via algorithm. All divisions within a court, and all courts within a province, use the same assignment system. They follow a top-down approach: the provincial court purchases the assignment system for all courts within the province.

1.2 Manual, Random and ML

Manual Under manual assignment, court staff assign cases based on perceived judge-case fit, allowing consideration of extensive information about judicial capabilities, performance, and communication skills. While this discretion can produce better matches, it creates opportunities for judge shopping and allows inappropriate factors—such as personal connections or demographics—to influence assignments. These concerns were documented by judges within the system, including commentary from the Shanghai Second High Court[4]. Manual assignment remained China's predominant method through 2014. Internationally, manual assignment persisted in the U.S. until

[3] Supreme People's Court. "Several Opinions on Improving Judicial Accountability System of the People's Courts." http://gongbao.court.gov.cn/Details/58f02f7ad96f8dcb0e75b8c7e08999.html

[4] Wang Zhigang. Exploration and Practice of Case Assignment System Reform. People's Court Daily, 2016-03-02.

the mid-1990s and remained common in some OECD countries as of 2007, according to (Fabri and Langbroek, 2007).

Random Random assignment matches cases to judges based on predefined rules without considering case- or judge-specific characteristics. Typically, incoming cases are assigned to available judges with the lowest current caseloads. This design eliminates human discretion but may produce suboptimal matches. For instance, a family law specialist might be poorly suited for a commercial loan case, just as a junior judge might be inappropriate for a highly complex case. China adopted random assignment around 2014, following the Supreme People's Court guidance to move away from manual assignment. Internationally, the U.S. implemented random case assignment as early as 1995, setting a global trend. By 2016, most countries used some form of random assignment, though with considerable variation[5].

ML-based assignment combines elements of both manual and random systems. An ML algorithm is trained on observable judge characteristics, case features, and performance metrics, then matches judges to incoming cases by optimizing predicted outcomes based on these characteristics[6]. While this approach eliminates direct human discretion in assignment decisions, it may still reflect biases embedded in the training data. Additionally, ML systems lack access to the full range of judge characteristics considered in manual assignment, potentially resulting in suboptimal matches. The ML-based assignment system is implemented in China only, introduced over the past decade.

1.3 Timing and Choice of Assignment System

As shown in Figure A2, among the 26 provinces sampled, courts in 8 provinces adopted random assignment while 18 implemented ML-based systems[7]. Geographical proximity to algorithm developers likely influenced these choices: northern provinces tended toward ML systems due to a Beijing-based ML company's presence. By 2021, most Chinese courts had transitioned from manual to either random or ML assignment[8]. Adoption timing varied significantly at the provincial level (Figure A3), influenced by budget constraints, procurement processes, and leadership preferences[9]. Provincial wealth showed no clear correlation with timing: Beijing, for instance, adopted systems relatively late. The variation in implementation timing provides the identification strategy for this study's empirical analysis. The data section will provide further details on the definitions and timing of random and ML adoption.

[5]Please see Good Practices for Courts Report from World Bank for details: `https://documents1.worldbank.org/curated/en/465991473859097902/pdf/108234-WP-GoodPracticesforCourtsReport-PUBLIC-ABSTRACT-EMAILED.pdf`.

[6]While software companies provide details about the algorithms used (e.g., Random Forest, XGBoost), the exact parameters are proprietary. This paper focuses on the outcomes rather than replicating the algorithms.

[7]Five provinces did not disclose their assignment systems and were therefore excluded from the sample.

[8]This study focuses on provinces that implemented new systems between 2014 and 2020, based on data accessibility.

[9]Although anecdotal evidence suggests variation in timing could trickle down to the prefecture level due to technical challenges in installing systems, provincial-level variation is used for consistency.

2 Empirical Strategy

The identification strategy uses the exogenous variation in the timing of the implementation of algorithmic case assignment systems across Chinese courts from 2014 to 2020[10]. I use a staggered Difference-in-Differences (DiD) framework by comparing outcomes between treated and control groups before vs. after the treated court divisions' adoption of the random or ML-based assignment. Because all court divisions are eventually treated, I use not-yet-treated court divisions as the control group for already treated court divisions at each point in time. Here, random and ML-based systems serve as separate treatments to avoid selection bias in assignment types. To be specific, I compare court divisions within each treatment type: not-yet-random court divisions to already random court divisions, and not-yet-ML court divisions to already ML court divisions[11]. The following subsection provides a detailed explanation of my DiD framework. I will first discuss why my setting does not satisfy the assumptions required for unbiased estimation using the traditional Two-Way Fixed Effects (TWFE) method. Next, I will justify the choice of Callaway and Sant'Anna (2021) approach and demonstrate how my setting meets the necessary assumptions. Finally, I will outline the procedure for estimating average treatment effects and event studies using CSDiD, tailored to the context of this study.

2.1 Choice of Generalized DiD

The standard solution for analyzing staggered adoption is the Two-Way Fixed Effects (TWFE) approach. However, TWFE relies on two key assumptions to provide unbiased results: static treatment effects and homogeneous treatment effects across groups, as highlighted in recent literature (Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021; Sun and Abraham, 2021)[12]. In my setting, these assumptions are likely not met. First, TWFE's use of already-treated groups as controls assumes static treatment effects where the treatment impact must remain constant over time. This is a challenge in the context of case assignment, where judges gain experience as they accumulate cases under the new system[13]. Second, TWFE requires homogenous treatment effects across different groups, which is unlikely given that court divisions in various provinces handle cases with

[10] As discussed in Section 1, the timing of policy adoption was influenced by a range of factors, including uncertainties in procurement processes, and budget constraints. Currently, there is no evidence suggesting that these factors correlate with pre-policy outcome variables.

[11] There is limited evidence showing selection into treatment. Pre-treatment case assignment patterns between random and ML court divisions are similar (see Section 4). The decision to adopt random or ML is partially influenced by factors like geographical proximity to assignment system developers. For instance, ML court divisions are often concentrated in northern China, where the only ML assignment developer is based. To be cautious, I limit the control group to not-yet-treated court divisions within the same treatment type.

[12] Sun and Abraham (2021) demonstrates that heterogeneous treatment effects can create pre-trends, complicating interpretation. Additionally, Goodman-Bacon (2021) shows that TWFE only assigns positive weights when treatment effects remain constant; varying effects can skew results towards groups with larger sample sizes, especially those treated mid-study.

[13] If new case assignment systems improve performance, their benefits could build as judges grow more adept. Conversely, if these systems hinder performance, the effects might reduce as judges adjust or worsen if mismatches increase over time.

distinct characteristics and have diverse judicial personnel, leading to heterogeneous effects[14]. If each court division fine-tunes the system, which is an option, this can further introduce heterogeneity. To address these concerns, I use the Difference-in-Differences (DiD) method developed by Callaway and Sant'Anna (2021), which accommodates multiple time periods and variation in treatment timing.

The reliability of CSDiD depends on several key assumptions: Irreversibility of Treatment: Once a province adopts a case assignment reform, it remains treated. This is consistent with the context of case assignment systems in China, where provinces rarely revert to manual assignment after transitioning to random or ML-based systems[15].

Random Sampling: Each observation should be randomly drawn from the population of interest. My dataset consists of panel data covering court performance, judicial effort, and case characteristics for every division within all court divisions over the specified timeframe (6 quarters pre-policy and 8 quarters post-policy). This dataset includes the universe of publicly disclosed cases, ensuring representative coverage.

Limited Treatment Anticipation: The assumption is that provinces do not change their assignment patterns in anticipation of reform implementation. This assumption appears reasonable, as adoption timing was driven by a mix of factors such as budget, procurement process, and preferences of provincial leaders—rather than pre-existing trends in performance or effort.

Conditional Parallel Trends: The identification strategy requires that, conditional on observable covariates, treated and not-yet-treated court divisions would have followed similar trends in the absence of the reform. The event studies presented in Section 5 shows the absence of pre-trends. Not-yet-treated court divisions serve as the control group.

Common Support: There should be sufficient overlap in characteristics between treated and control groups, ensuring that treated and control court divisions have comparable profiles. This condition is met, as the control group consists of court divisions in not-yet-treated provinces. Given the substantial variability in court performance, judge characteristics, and case features within a single province, the earlier treated court divisions are likely to find suitable comparisons among the later treated court divisions.

In the not-yet-treated as comparison group, the group time average treatment effect (ATT) is:

$$ATT(g,t) = \mathbb{E}\left[Y_t - Y_{g-1} \mid G_g = 1\right] - \mathbb{E}\left[Y_t - Y_{g-1} \mid D_t = 0, G_g = 0\right] \tag{1}$$

where $g \in (2014\text{ Q3}, 2018\text{ Q4})$ is the quarter the court divisions was first treated (for example, 2016 Q2 corresponds to the court divisions first treated in second quarter of 2016, and so on), $t \in (2014\text{ Q1}, 2018\text{ Q4})$ is a quarter, G_g is a binary variable that equals one if a unit is first treated in quarter g and zero otherwise. D_t is a binary variable that equals one if a unit is treated in quarter

[14] In staggered designs involving multiple groups and time periods, assuming homogenous treatment effects is often unrealistic (De Chaisemartin and d' Haultfoeuille, 2020).

[15] Even if reversals occur, they are limited to specific small divisions within court divisions and not on a large scale. However, provinces may upgrade or switch to new assignment systems after the expiration of the current contracts, which typically last for two years. During my study's post-policy period (eight quarters), the treatment is stable.

t. Here, the first term shows the difference of the outcome variable of the treated court divisions between period t and the period before first-treated, $g-1$. The second term shows the difference between the court divisions not-yet-treated by quarter t and themselves one period before period $g-1$. The ATT is a weighted average of the group time ATT:

$$\theta = \frac{1}{\kappa}\sum_{g\in G}\sum_{t}\mathbb{1}\{t \geq g\}ATT(g,t)P(G=g \mid G \leq 2020\text{ Q4}) \tag{2}$$

where $ATT(g,t)$ is defined in Equation 1, and $\kappa = \sum_{g\in G}\sum_{t}\mathbb{1}\{t \geq g\}P(G=g \mid G \leq 2020\text{ Q4})$. κ ensures that the weights on $ATT(g,t)$ in the second term sum up to one. This weighting method ensures positive weights and gives a larger weight to groups with more court divisions. The event study specification is:

$$\delta_e = \sum_{g}\sum_{t}\mathbb{1}\{t-g=e\}P(G=g \mid t-g=e)ATT(g,t) \tag{3}$$

where δ_e represents the average treatment effect e periods after adoption, calculated across all groups that participated in the treatment for exactly e periods.

3 Data

This study uses government procurement data to identify treatment timing and type, and 66 million case documents to analyze case features, judge characteristics, and case outcomes. Below, I detail the sources of data, data processing methods, and key summary statistics.

3.1 Case Assignment System Adoption Data: China Government Procurement Online (中国政府采购网)

I collected contracts between provincial high courts and technology firms to identify when and what type of algorithmic case assignment system was adopted. The adoption date is set as the start date of the contract, rounded to the first day of the corresponding quarter for standardization. This timeline is cross-referenced with local policy documents and news reports to ensure accuracy. Notably, all divisions within a court and courts within a province follow the same adoption schedule.

To determine the type of assignment system, I checked the technical report from the contracted firms' websites. Based on the algorithms mentioned, I classified each court into random or ML-based assignment categories. Figure A3 shows the quarterly distribution of courts adopting random and ML systems, indicating a relatively even spread over time.

3.2 Case Documents: China Judgment Online (中国裁判文书网)

China Judgment Online (CJO) is a centralized platform, launched by the SPC in 2013, that hosts publicly accessible judicial documents. According to SPC guidelines, all judgment documents must be uploaded unless exempted for legal reasons[16][17]. As of January 2023, CJO contained 120 million case records.

3.2.1 Sample Selection

I focus on 26 provinces where the timing and type of automated system adoption are confirmed. Some provinces, while known to have adopted an algorithm by 2020, lack precise adoption details and are therefore excluded from the analysis. The study period spans from January 1, 2014, to December 31, 2020, covering the full rollout of automation. My sample includes civil, business, criminal, and administrative cases, comprising 85% of the total caseload—resulting in a final dataset of 66 million cases[18].

3.2.2 Raw Case Document

Each CJO document consists of four sections: title, facts, verdict, and closing. These sections provide details on case characteristics, judge characteristics, and case outcomes, which are extracted as follows: Title: The title includes the case name, unique case ID, and court ID. From this, I extract the court name, case ID, filing year, and case type (e.g., civil, business).

Fact: The facts section covers litigant details and case background. I extract litigant names, the number of litigants, their identity (individual or corporate), the monetary amount in dispute, cause of action (COA), and related case IDs.

Verdict: The verdict section contains judicial reasoning and the final decision. Key extracted elements are the verdict word count, legal provisions cited, and provisions written in the decision.

Closing: This section records judge(s) and clerk(s)' names and the decision date. I use judge names, case type, and court ID to construct judge-level panel data and court-division-quarter panel data, enabling analyses of both assignment patterns and case outcomes[19][20].

3.2.3 Variables for Case Assignment Patterns and Case Outcomes

To assess case assignment patterns (Section 4) and case outcomes (Section 5, Section 6), I use summary indices for case complexity, judge/court-division performance, and judge/court-division

[16] This document can be found at http://gongbao.court.gov.cn/Details/d0e837bbafb75a8863b4d4c407d694.html

[17] Some cases are restricted from public access, such as divorce or juvenile cases, but the platform still displays basic case details and reasons for non-disclosure. This allows calculation of transparency rates for different courts. For robustness, I use courts with disclosure rates above 85% and 90% as alternative samples, and results remain consistent.

[18] Law enforcement cases, accounting for 15% of total caseloads, are excluded as they do not involve judge decision-making but rather execution of court rulings, which falls outside the scope of traditional judicial reasoning.

[19] Judges rarely move between courts in China, and the combination of judge name and court ID serves as a reliable identifier due to the unique nature of Chinese names.

[20] Court divisions are not explicitly specified in raw documents. I proxy divisions using case types and COAs, as case processing standards vary across divisions, e.g., small loans vs. anti-trust. This granularity allows for normalization of outcome measures within divisions.

effort, reducing dimensionality and offering a clearer overview. Following Anderson (2008), summary indices are beneficial for minimizing over-testing risks and providing a comprehensive view of general effects. I aggregate variables by (1) standardizing signs so that higher values indicate better outcomes, (2) normalizing based on pre-treatment standard deviations, and (3) weighting by the inverse of the covariance matrix[21].

The case complexity index, inspired by Gramckow et al. (2016), includes the number of litigants, the percentage of corporate litigants, disputed monetary amounts, and the number of COAs, indicating the workload involved[22]. For the performance index, I use the number of cases processed, appeal rate, and reversal rate. This selection of performance metrics follows SPC's 2011 official guideline on assessing performance and Kahn and Li (2019)[23]. For the judicial effort index, I draw from verdict word count, provisions cited from existing law, and provisions written in the decision, guided by the SPC' s standards and (Liu et al., 2022; Liu, 2018)[24].

3.2.4 Summary Statistics

Table A1 and A2 compare pre-treatment averages for courts using random and ML systems. Columns (1) and (2) present averages for random and ML courts, respectively, while column (3) reports the differences. Random courts show higher caseloads and slightly lower appeal and reversal rates, implying better performance. They also produce longer and more complex verdicts. Panel C reveals higher caseload inputs and fewer judges in random courts, which handle more civil cases than ML courts. These differences hint at potential selection bias, reinforcing the need to treat not-yet-treated courts within the same category as controls for each treatment type.

4 Who Gets What Case?

While existing research has focused on proving or disproving randomness in case assignment, the actual patterns of case allocation in non-random systems remain largely unexplored. Previous studies, typically relying on field interviews or assignment rule comparisons, have been limited to small-scale variations across courts or confined to theoretical frameworks (Gramckow et al., 2016; Fabri and Langbroek, 2007; Macfarlane, 2023; Jin, 2020; Chilton and Levy, 2015). The empirical reality of case assignment patterns, especially among Chinese courts has remained a black box. However, both theoretical and qualitative research suggest these allocation patterns as crucial determinants of court performance, potentially the primary mechanism through which assignment systems influence outcomes (Hübert, 2021; Macfarlane, 2023). Understanding how

[21] This method down-weights highly correlated variables and emphasizes variables with unique information. I also conducted robustness checks using equal weights, yielding consistent results.

[22] Admittedly, the complexity index cannot fully capture the real workload. Some cases might appear straightforward initially, but appear more complex during the hearing process. The complexity index aims to reflect the court's initial assessment of case difficulty, which influences assignment decisions. Therefore, I limit the index to pre-assignment variables, avoiding any bias introduced by judges' characteristics that might change the perceived complexity after the case is assigned.

[23] This document can be found at `http://www.court.gov.cn/zixun-xiangqing-2298.html`.

[24] Please check for details: `http://gongbao.court.gov.cn/Details/25a9b4684d384ea16f78e276f14f13.html`.

cases are distributed among judges is therefore essential before assessing performance. This section focuses on judge-case expertise matching as the primary dimension of interest. Appendix B provides additional analysis of how case complexity correlates with judge demographics and case history.

Expertise matching refers to the alignment between a judge's specialization and the nature of assigned cases. When family law judges handle divorce cases or patent judges oversee infringement disputes, they can process facts more efficiently and apply relevant precedents more effectively, potentially improving both case outcomes and judicial effort.

This analysis assesses how the judge expertise changes following the transition to random and ML assignment. I measure judge expertise using relative specialization within each court division. For each judge-case pair, I calculate the proportion of cases with the same cause of action that the judge has previously handled, then compute the average (or median) across all cases in that division during a given quarter. For example, an expertise score of 60% indicates that cases in a particular division are assigned to judges who have, on average, previously handled 60% of cases with the same cause of action. This measure captures observable specialization patterns, though more subtle judge preferences or informal expertise remain unobservable.

I use the same generalized difference-in-differences approach described in section 2, with the outcome variable defined as the average expertise matching ratio within each court division-quarter. All other specifications, including controls and clustering, remain identical to the main analysis.

Figure A5 presents the results, with the left panel showing random assignment court divisions and the right panel showing ML divisions. Random assignment shows no significant improvement in expertise matching compared to manual assignment. Accounting for pre-trends, judge-case matching decreases slightly but insignificantly (-0.02). The pre-treatment average is [NUMBER]. This finding aligns with the design of random assignment: by construction, it disregards judge characteristics and case features that human assigners might use to create expertise-based matches.

In contrast, ML-based assignment significantly increases expertise matching by 0.02, representing approximately a 4.2% improvement over the baseline matching rate of 0.470. This improvement emerges gradually over time, consistent with the algorithm learning optimal matching patterns from historical data. The ML system appears to identify and replicate successful judge-case pairings that human assigners previously made, while potentially discovering combinations that humans might have overlooked.

5 Result

The adoption of random and ML judge-case assignment systems produces markedly different effects on court performance and effort. Court divisions that transitioned to random assignment experienced significant declines in both metrics, with performance decreasing by 0.054 and effort by 0.114 on the standardized indices. In contrast, divisions adopting ML-based systems showed no statistically significant changes in either performance or effort, maintaining baseline levels. These effects remain consistent across all division types. The following subsections provide detailed analysis of each outcome.

5.1 Random Courts

Main Results Figure A6 presents event study results for random assignment effects. The left panel illustrates estimated effects on the performance index, while the right panel displays results for the effort index. Both diagrams demonstrate that estimates for the five pre-treatment periods are close to zero and statistically insignificant, showing no significant pre-treatment trends.

Following random assignment implementation, the performance index shows a delayed but significant decline beginning in the third quarter, with this negative trend persisting thereafter[25]. Table A3, Panel A, presents point estimates for the performance index (column 1), normalized outcome variables (columns 2-4), and raw outcome variables (columns 5-7). Following policy intervention, the performance index declines by 0.054, statistically significant at the 1% level. This decline is primarily driven by a 0.110 standard deviation decrease in the normalized non-appeal rate (significant at the 10% level), which corresponds to approximately six additional appeals per court division per quarter (significant at the 5% level)[26].

In contrast, the effort index shows an immediate post-policy decline that stabilizes around zero in subsequent quarters[27]. This index, measuring judicial input, reflects the adaptation phase judges undergo.

Table A3, Panel B, quantifies the decline in the effort index, showing a reduction of 0.114 standard deviations (significant at the 1% level) driven primarily by decreases in verdict word count and provisions cited. The normalized verdict word count decreases by 0.097 standard deviations, equivalent to approximately 46.533 fewer characters per case per quarter (significant at the 1% level). Similarly, the normalized number of provisions cited decreases by 0.097 standard deviations, corresponding to 0.096 fewer provisions per case per quarter (also significant at the 1% level).

Heterogeneity by Divisions Figure A8 reveals that random assignment reduces efficiency across divisions, with increased effort failing to yield corresponding performance gains. The left panel presents heterogeneous effects on the performance index across five divisions. Most effects are either negative or statistically insignificant compared to the pre-period, with the civil division, which handles the highest caseload, experiencing a significant decline of 0.13 in the standardized performance index. The right panel shows effects on effort by division. While the economic-related

[25] This delayed effect is expected given the nature of the performance index, which includes processed caseloads, appeal rates, and reversal rates. Appeal and reversal rates change gradually as litigants require time to evaluate the costs and benefits of appealing first-instance decisions. Moreover, reversals by higher courts occur only when substantial errors are identified in original judgments, a process that requires considerable time for review and assessment.

[26] While the performance index provides a general measure of court output, appeal and reversal rates can be influenced by multiple factors beyond judicial quality. Litigants might appeal more frequently if they perceive higher courts as less corrupt following policy implementation, or conversely, if they view newly assigned judges as less experienced or professional. Despite these limitations, the performance index serves as a pragmatic measure of court efficiency, analogous to key performance indicators in business settings, where higher appeal rates signal increased costs and resources required for case resolution.

[27] The initial decrease in effort is consistent with a learning curve as judges adapt to cases assigned under the new random system. Initially, judges may produce shorter verdicts while becoming acquainted with unfamiliar case types, but they develop familiarity and adjust accordingly over time.

crimes division, which handles high-stakes cases, records a slight increase of 0.1 in the standardized effort index, other divisions exhibit no significant changes in effort. These results demonstrate that random assignment generally decreases efficiency across divisions, as divisions with similar effort levels show lower performance, while those with increased effort do not achieve corresponding performance improvements.

5.2 ML Courts

Main Results Figure A7 presents event study results for ML assignment effects, with the left panel showing the performance index and the right panel displaying the effort index. Both panels demonstrate estimates that remain close to zero across all quarters, with no significant pre-treatment trends. The ML case assignment system produces no significant changes in judicial performance or effort following implementation. These results indicate that ML assignment maintains baseline performance and effort levels.

Table A4, Panel A, provides detailed point estimates for the performance index and its component variables. The overall performance index increases slightly by 0.025, but this effect is not statistically significant. Similarly, the normalized component variables, including cases processed, non-appeal rates, and reversal rates, show no significant changes, showing that ML assignment maintains existing performance levels.

Similarly, Table A4, Panel B, shows no significant changes in the effort index or any of its components. The normalized values for verdict word count, provisions written and cited remain unchanged, as do the raw counts for these variables. This indicates that ML assignment does not appear to change judicial effort levels, either in the overall index or in the individual elements that composes it.

Heterogeneity by Divisions The heterogeneous analysis by division, presented in Figure A9, underscores the absence of significant effects across different types of cases. Both the performance and effort indices remain close to zero across all divisions, including economic-related crime, ordinary crime, business, civil, and civilian vs. government cases, highlighting the uniform impact (or lack thereof) of ML assignment. These results collectively suggest that ML-based case assignment does not substantially change the efficiency within each division, as the input and output levels stay the same.

In summary, random and ML assignment systems exert different effects on court operations. Random assignment leads to decreased performance, driven by increased appeals and reduced verdict length and legal provisions cited. Heterogeneity analysis reveals declining efficiency across divisions: similar effort levels yield worse performance, while increased effort fails to improve performance outcomes. In contrast, ML assignment produces no significant changes in either performance or effort indices. None of the component variables show significant changes compared to the pre-policy period, and this pattern remains consistent across all divisions.

5.3 Robustness Checks

I conduct several robustness checks on the main results, which remain consistent across alternative weights, alternative samples, and additional controls.

First, I assign alternative weights to the outcome indices. The performance and effort indices are initially constructed as weighted averages using the inverse covariance matrix, which assigns higher weights to less correlated variables and reduces redundant information. However, this approach is sensitive to the specific data structure. To assess robustness, I construct alternative indices using equal weights for each component variable. The effects on both performance and effort indices remain consistent across random and ML court divisions under equal weight.

Second, I apply alternative sample restrictions by limiting the analysis to courts with case disclosure rates of at least 85% and 90%, respectively. As discussed in the data section, certain case types, such as juvenile or divorce cases, are legally withheld from public disclosure, potentially affecting courts with lower disclosure rates. Restricting the sample to courts with higher disclosure rates helps ensure that incomplete data does not bias the results. I find that the effects on both performance and effort indices remain robust to these alternative disclosure thresholds in both random and ML court divisions.

Third, I include additional controls for case and judge characteristics to verify that the results are not driven by changes in case composition or judge profiles. The estimated effects remain similar with these additional controls, confirming the robustness of the findings.

6 Sentencing Disparity: A Case Study of Dangerous Driving Cases

Having shown that random and ML assignment systems generate distinct performance outcomes, we next examine whether these systems differentially affect judicial decision-making at the case level. While measuring "judicial fairness" or "justice" directly remains challenging due to ongoing definitional debates and context-dependent interpretations, we can examine sentencing disparities as a more tractable proxy: specifically, whether cases with similar observable characteristics receive consistent treatment.

The literature on sentencing disparities typically focuses on demographic and ideology differences, examining either variation across judges with different backgrounds or differential treatment of defendants by race, gender, or other characteristics (Abrams et al., 2012; Cohen and Yang, 2019; Harris, 2024)[28]. However, given limited demographic variation in my sample and my interest in obtaining a comprehensive assessment of algorithmic impact, I adopt a broader definition: sentencing disparity as the degree to which cases with similar observable facts receive different outcomes.

This measure necessarily overestimates true disparity because case documents cannot capture all relevant factors that might justify differential treatment. Increased measured disparity could reflect either problematic arbitrary decision-making or beneficial judicial discretion responding to

[28]For a comprehensive review of this literature on judicial bias and demographic effects in sentencing, see Harris and Sen (2019).

unrecorded case-specific circumstances. I therefore interpret my disparity measures in conjunction with our performance findings to distinguish between these alternative explanations.

6.1 Dangerous Driving Offense (DDO) Cases

I measure sentencing disparities using dangerous driving offense (DDO) cases as our empirical setting. DDO cases, defined under Article 133-1 of China's Criminal Law, include four categories of vehicular offenses: reckless racing, drunk driving, severe violations in passenger transport, and improper transport of hazardous materials. These cases provide an ideal setting for studying sentencing disparity for several reasons.

First, DDO cases represent the most prevalent category in Chinese criminal courts, with the dataset containing 1.2 million cases, accounting for 28.26% of ordinary criminal cases processed during our study period. Second, DDO cases are highly standardized in terms of both legal reasoning and evidence requirement. Unlike complex criminal cases involving subjective assessments of intent or credibility, DDO cases typically rely on objective evidence such as blood alcohol concentration, traffic accident reports, and documented prior offenses. Third, sentencing outcomes are relatively straightforward: judges determine custody decisions (including probation eligibility), probation length, and fine amounts. All these features enables quantitative analysis.

I analyze a 10% stratified sample by court and year-quarter, yielding approximately [NUMBER] observations. Case characteristics and outcomes were extracted from legal documents using DeepSeek, a Chinese large language model, capturing key sentencing factors including blood alcohol content, accident severity, whether fled from the scene, confession behavior, prior criminal record, and vehicle license. Outcome variables include custody decisions, probation terms, and fine amounts.

This standardized setting allows us to apply decomposition methods to isolate changes in sentencing patterns after case assignment system switch from changes due to case composition differences.

6.2 DiNardo, Fortin, and Lemieux (DFL) Decomposition

To examine sentencing disparities comprehensively rather than focusing on specific demographic dimensions, I apply the DiNardo, Fortin, and Lemieux (DFL) decomposition method, widely used in labor economics to decompose wage gaps into components attributable to differences in observable characteristics versus differences in how those characteristics are valued in the labor market. Applied to judicial settings, this method allows us to distinguish between sentencing changes due to evolving case composition and changes due to changed judicial decision-making processes.

The DFL method decomposes the overall change in any distributional statistic between pre- and post-algorithm periods into two components: a composition effect capturing changes attributable to different case characteristics, and a structure effect reflecting changes in how courts translate case characteristics into sentencing outcomes. This approach proves particularly valuable because it can simultaneously analyze both mean changes (indicating systematic shifts in leniency) and variance changes (indicating alterations in judicial discretion).

Formally, let Group A represent the pre-algorithm period and Group B the post-algorithm period. For any distributional statistic W (such as mean or variance), the overall change is:

$$\Delta W_o = W(F_{Y_B|D_B}) - W(F_{Y_A|D_A}) \tag{4}$$

where $F_{Y_g|D_g}$ represents the observed distribution of sentencing outcomes Y for cases in group g. The DFL decomposition constructs a counterfactual distribution F_Y^C representing what sentencing outcomes would have been if pre-algorithm cases had faced post-algorithm case characteristics while retaining pre-algorithm sentencing standards:

$$F_Y^C(y) = \ F_{Y_A|X_A}(y|X)dF_{X_B}(X) \tag{5}$$

This counterfactual is estimated through reweighting, where each pre-algorithm case receives weight $\theta(X) = [Pr(D_B = 1|X)/Pr(D_B = 1)]/[Pr(D_A = 1|X)/Pr(D_A = 1)]$, computed from a logistic regression of post-algorithm indicators on case characteristics X. The overall change then decomposes as:

$$\Delta W_o = \Delta W_S + \Delta W_X \tag{6}$$

where $\Delta W_S = W(F_{Y_B|D_B}) - W(F_Y^C)$ captures the structure effect (changes in judicial decision-making) and $\Delta W_X = W(F_Y^C) - W(F_{Y_A|D_A})$ captures the composition effect (changes in case characteristics)[29].

In my application, X includes extracted case characteristics (blood alcohol content, traffic accident involvement, cooperation behavior, prior record, and license status) while Y represents sentencing outcomes (custody decisions, probation length, and fine amounts). The structure effect in mean outcomes indicates systematic changes in leniency, while the structure effect in variance measures changes in judge discretion. I bootstrap each decomposition 100 times to obtain standard errors robust to estimation uncertainty.

6.3 Result: More Lenient, More Discretion

Table A5 presents decomposition results for mean sentencing outcomes. Panel A shows that random assignment courts exhibit systematic leniency across multiple dimensions in the raw sentencing outcome. However, after controlling for differences in observable case characteristics through the DFL decomposition, the structure effects shows significant decreases in total custody terms (-0.190 month) while immediate custody shows no significant change. Similarly, after removing composition effects, the probability of receiving probation remains unchanged, but probation length decreases significantly (-0.014 month). This pattern suggests that random assignment primarily affects con-

[29]For a more comprehensive introduction to DFL reweighting and other decomposition method, see Fortin et al. (2011)

tingent rather than immediate punishments: judges impose the same immediate consequences but reduce penalties for probation violations. Panel B shows ML assignment courts display different patterns. The structure effects, show significant decreases in total custody (-0.121 month) with no change in immediate custody, similar to random courts. However, the structure effect on probation patterns differ: ML courts show no significant changes in either probation probability or duration. This suggests judges after ML assignment focus specifically on reducing contingent punishment.

Table A6 examines changes in sentencing variance. Panel A reveals that while, in observed difference, random assignment courts show decreased variance (suggesting more uniform outcomes), after controlling for case composition, structure effects are consistently positive and significant across all sentencing dimensions, indicating judges exercise substantially more discretion based on unobservable case factors. Panel B shows ML assignment courts exhibit mixed observed difference in variance but consistently large positive structure effects, similar to random courts. Judges in both systems exercise substantially more discretion based on unobservable case factors than they did before algorithmic assignment, though this discretionary decision-making may differ in quality.

6.4 Interpretation: Calculated versus Arbitrary Leniency

The structure effects reveal distinct patterns when combined with our performance findings. Random assignment generates ”arbitrary leniency” where judges’ behavioral changes, independent of case mix, include more lenient sentences and increased discretion but result in worse performance metrics. ML assignment enables ”calculated leniency” where similar judicial behavioral changes occur after controlling for case composition, but with maintained performance quality.

This distinction mirrors findings in Angelova et al. (2023), who demonstrate that both high- and low-performing judges utilize private information in bail decisions, but high-performing judges apply this information more effectively. In this context, ML assignment appears to facilitate more effective use of judge discretion while random assignment may enable discretionary decision-making that undermines sentencing quality.

Two limitations should be acknowledged when interpreting these results. First, the DFL decomposition method provides an accounting framework rather than strict causal identification. While we can decompose observed changes into composition and structure effects, establishing that these structure effects definitively result from algorithmic assignment requires additional identifying assumptions.

Second, this analysis focuses on a single category of criminal cases, potentially limiting generalizability to more complex legal proceedings. However, DDO cases provide a conservative test of algorithmic effects. If algorithms influence judicial behavior even in these standardized cases, the effects are likely magnified in more complex legal contexts requiring greater judicial discretion.

7 Conclusion

This paper examines how the transition from manual to automated case assignment systems affects case allocation, court performance, and sentencing disparities, leveraging China's staggered adoption of random and machine learning (ML) assignment systems since 2014. Our findings suggest that random assignment, while reducing path dependence, does not increase judge-case expertise matching and diminishes overall court performance. Analysis of a subsample of criminal cases reveals increased leniency and greater sentencing disparities under random assignment, controlling for case characteristics. This suggests that reduced performance coincides with more variable judicial discretion that does not translate into improved outcomes. In contrast, ML-based assignment systems reduce path dependence and improve expertise matching while maintaining performance levels. Although these systems also exhibit increased leniency and sentencing disparities, the discretion appears more calculated.

These results bring out a paradox in workload allocation: different conceptualization of impartiality conflict each another. Random assignment ensures equal treatment of judges but produces more variable and lower-quality decisions. ML assignment generates more consistent decisions and maintains performance but treats judges differentially based on their expertise profiles. Neither approach is inherently superior in terms of impartiality, as they optimize different normative values: random assignment prioritizes equal treatment of the actors, while ML assignment prioritizes consistent outcomes.

This tension extends beyond courts to any institutional setting requiring trade-offs between equal treatment of workers and consistency of outcomes. The conventional algorithmic bias often assumes that neutral processes yield fair outcomes, but this paper suggests this relationship is more nuanced. Sometimes processes that appear "biased" toward expertise matching produce more consistent and higher-quality decisions than random allocation.

My findings suggest two possible paths for future research. First, from a normative perspective, better designed algorithms might allow systems that preserve expertise matching while incorporating sufficient arbitrariness and explainability. This direction requires balancing competing objectives within a single allocation mechanism. Second, from a positive perspective, when these priorities cannot be reconciled within a single algorithm, we must examine the long-term labor market effects of different work allocation approaches. In the court contexts, for instance, algorithmic assignment may change judges' human capital accumulation. If algorithms systematically change the types and complexity of cases that judges handle, judge's expertise profile and career path might differ, a consequence that may not be immediate but could surface over time.

References

Abrams, David S, Marianne Bertrand, and Sendhil Mullainathan, "Do judges vary in their treatment of race?," *The Journal of Legal Studies*, 2012, *41* (2), 347–383.

Agan, Amanda, Jennifer L Doleac, and Anna Harvey, "Misdemeanor prosecution," *The Quarterly Journal of Economics*, 2023, *138* (3), 1453–1505.

Aizer, Anna and Joseph J Doyle Jr, "Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges," *The Quarterly Journal of Economics*, 2015, *130* (2), 759–803.

Anderson, Michael L, "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American statistical Association*, 2008, *103* (484), 1481–1495.

Angelova, Victoria, Will Dobbie, and Crystal S Yang, "Algorithmic Recommendations When the Stakes Are High: Evidence from Judicial Elections," in "AEA Papers and Proceedings," Vol. 114 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2024, pp. 633–637.

_ , Will S Dobbie, and Crystal Yang, "Algorithmic recommendations and human discretion," Technical Report, National Bureau of Economic Research 2023.

Arnold, David, Will Dobbie, and Peter Hull, "Building Nondiscriminatory Algorithms in Selected Data," *American Economic Review: Insights*, 2025, *7* (2), 231–249.

Battaglini, Marco, Luigi Guiso, Chiara Lacava, Douglas L Miller, and Eleonora Patacchini, "Refining public policies with machine learning: The case of tax auditing," *Journal of Econometrics*, 2024, p. 105847.

Becker, Luc, Bastian Wurm, and Thomas Hess, "Will Algorithms Replace Managers? A Systematic Literature Review on Algorithmic Management," 2023.

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

Berk, Richard A, Susan B Sorenson, and Geoffrey Barnes, "Forecasting domestic violence: A machine learning approach to help inform arraignment decisions," *Journal of empirical legal studies*, 2016, *13* (1), 94–115.

Bhuller, Manudeep, Gordon B Dahl, Katrine V Løken, and Magne Mogstad, "Incarceration, recidivism, and employment," *Journal of political economy*, 2020, *128* (4), 1269–1324.

Boyd, Christina L, Lee Epstein, and Andrew D Martin, "Untangling the causal effects of sex on judging," *American journal of political science*, 2010, *54* (2), 389–411.

Callaway, Brantly and Pedro HC Sant'Anna, "Difference-in-differences with multiple time periods," *Journal of econometrics*, 2021, *225* (2), 200–230.

Chaisemartin, Clément De and Xavier d' Haultfoeuille, "Two-way fixed effects estimators with heterogeneous treatment effects," *American economic review*, 2020, *110* (9), 2964–2996.

Chilton, Adam S and Marin K Levy, “Challenging the randomness of panel assignment in the federal courts of appeals,” *Cornell L. Rev.*, 2015, *101*, 1.

Chyn, Eric, Brigham Frandsen, and Emily C Leslie, “Examiner and Judge Designs in Economics: A Practitioner’s Guide,” Technical Report, National Bureau of Economic Research 2024.

Cohen, Alma and Crystal S Yang, “Judicial politics and sentencing decisions,” *American Economic Journal: Economic Policy*, 2019, *11* (1), 160–191.

Cram, W Alec, Martin Wiener, Monideepa Tarafdar, Alexander Benlian et al., “Algorithmic Controls and their Implications for Gig Worker Well-being and Behavior.,” in “ICIS,” Vol. 2020 2020, pp. 1–17.

Eren, Ozkan and Naci Mocan, “Emotional judges and unlucky juveniles,” *American Economic Journal: Applied Economics*, 2018, *10* (3), 171–205.

— **and** — , “Juvenile punishment, high school graduation, and adult crime: Evidence from idiosyncratic judge harshness,” *Review of Economics and Statistics*, 2021, *103* (1), 34–47.

Fabri, Marco and Philip M Langbroek, “Is There a Right Judge for Each Case-A Comparative Study of Case Assignment in Six European Countries,” *Eur. J. Legal Stud.*, 2007, *1*, 292.

Fortin, Nicole, Thomas Lemieux, and Sergio Firpo, “Decomposition methods in economics,” in “Handbook of labor economics,” Vol. 4, Elsevier, 2011, pp. 1–102.

Frandsen, Brigham, Lars Lefgren, and Emily Leslie, “Judging judge fixed effects,” *American Economic Review*, 2023, *113* (1), 253–277.

Glynn, Adam N and Maya Sen, “Identifying judicial empathy: Does having daughters cause judges to rule for women’s issues?,” *American Journal of Political Science*, 2015, *59* (1), 37–54.

Goodman-Bacon, Andrew, “Difference-in-differences with variation in treatment timing,” *Journal of econometrics*, 2021, *225* (2), 254–277.

Gramckow, Heike P., Omniah Ebeid, Erica Bosio, and Jorge Luis Silva Mendez, “Good practices for courts report: Helpful elements for good court performance and the World Bank’ s quality of judicial process indicators - Key elements, lessons learned, and good practice examples,” 2016.

Harris, Allison P, “Can racial diversity among judges affect sentencing outcomes?,” *American Political Science Review*, 2024, *118* (2), 940–955.

— **and Maya Sen**, “Bias and judging,” *Annual Review of Political Science*, 2019, *22* (1), 241–259.

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li, “Discretion in hiring,” *The Quarterly Journal of Economics*, 2018, *133* (2), 765–800.

Hübert, Ryan, “Biased Judgments without Biased Judges: How Legal Institutions Cause Errors,” *The Journal of Politics*, 2021, *83* (2), 753–766.

Jarrahi, Mohammad Hossein, Mareike Möhlmann, and Min Kyung Lee, “Algorithmic management: The role of AI in managing workforces,” *MIT Sloan Management Review*, 2023.

Jin, Changwei, "Analysis on the Mechanism of AI Case Division," *Journal of CUPL*, 2020, *2020* (02). China University of Political Science and Law.

Kahn, Matthew E and Pei Li, "The effect of pollution and heat on high skill public sector worker productivity in China," Technical Report, National Bureau of Economic Research 2019.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human decisions and machine predictions," *The quarterly journal of economics*, 2018, *133* (1), 237–293.

_ , Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer, "Prediction policy problems," *American Economic Review*, 2015, *105* (5), 491–495.

Kling, Jeffrey R, "Incarceration length, employment, and earnings," *American Economic Review*, 2006, *96* (3), 863–876.

Lee, Min Kyung, Daniel Kusbit, Evan Metsky, and Laura Dabbish, "Working with machines: The impact of algorithmic and data-driven management on human workers," in "Proceedings of the 33rd annual ACM conference on human factors in computing systems" 2015, pp. 1603–1612.

Liu, Ernest, Yi Lu, Wenwei Peng, and Shaoda Wang, "Judicial independence, local protectionism, and economic integration: Evidence from China," Technical Report, National Bureau of Economic Research 2022.

Liu, Zhuang, "Does reason writing reduce decision bias? Experimental evidence from judges in China," *The Journal of Legal Studies*, 2018, *47* (1), 83–118.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan, "The unreasonable effectiveness of algorithms," in "AEA Papers and Proceedings," Vol. 114 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2024, pp. 623–627.

Macfarlane, Katherine A, "Constitutional Case Assignment," *NCL Rev.*, 2023, *102*, 977.

Rittenhouse, Katherine, Emily Putnam-Hornstein, and Rhema Vaithianathan, "Algorithms, Humans and Racial Disparities in Child Protection Systems: Evidence from the Allegheny Family Screening Tool," *Unpublished Working Paper*, 2023.

Rosenblat, Alex and Luke Stark, "Algorithmic labor and information asymmetries: A case study of Uber' s drivers," *International journal of communication*, 2016, *10*, 27.

Shayo, Moses and Asaf Zussman, "Judicial ingroup bias in the shadow of terrorism," *The Quarterly journal of economics*, 2011, *126* (3), 1447–1484.

Stevenson, Megan T and Jennifer L Doleac, "Algorithmic risk assessment in the hands of humans," *American Economic Journal: Economic Policy*, 2024, *16* (4), 382–414.

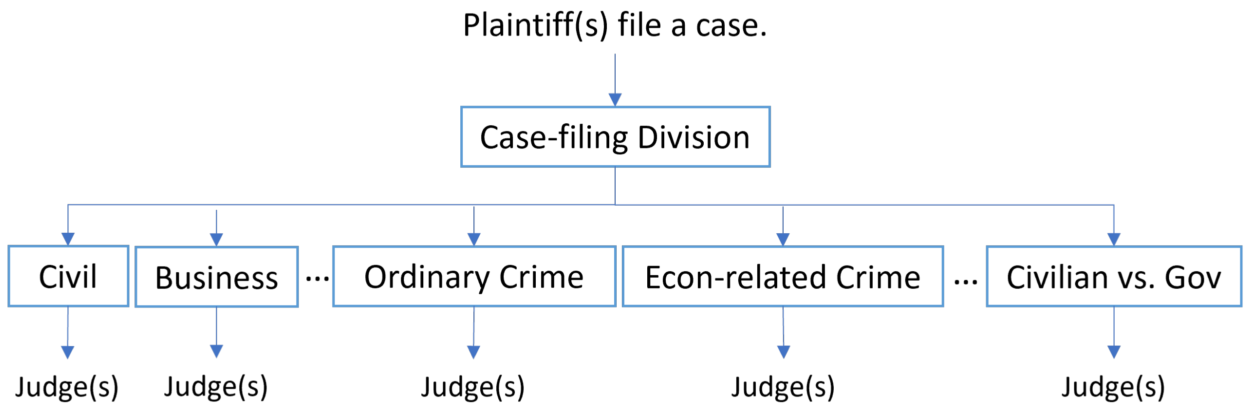
Stoddard, Greg, Dylan J Fitzpatrick, and Jens Ludwig, "Predicting police misconduct," Technical Report, National Bureau of Economic Research 2024.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of econometrics*, 2021, *225* (2), 175–199.

A Tables and Figures

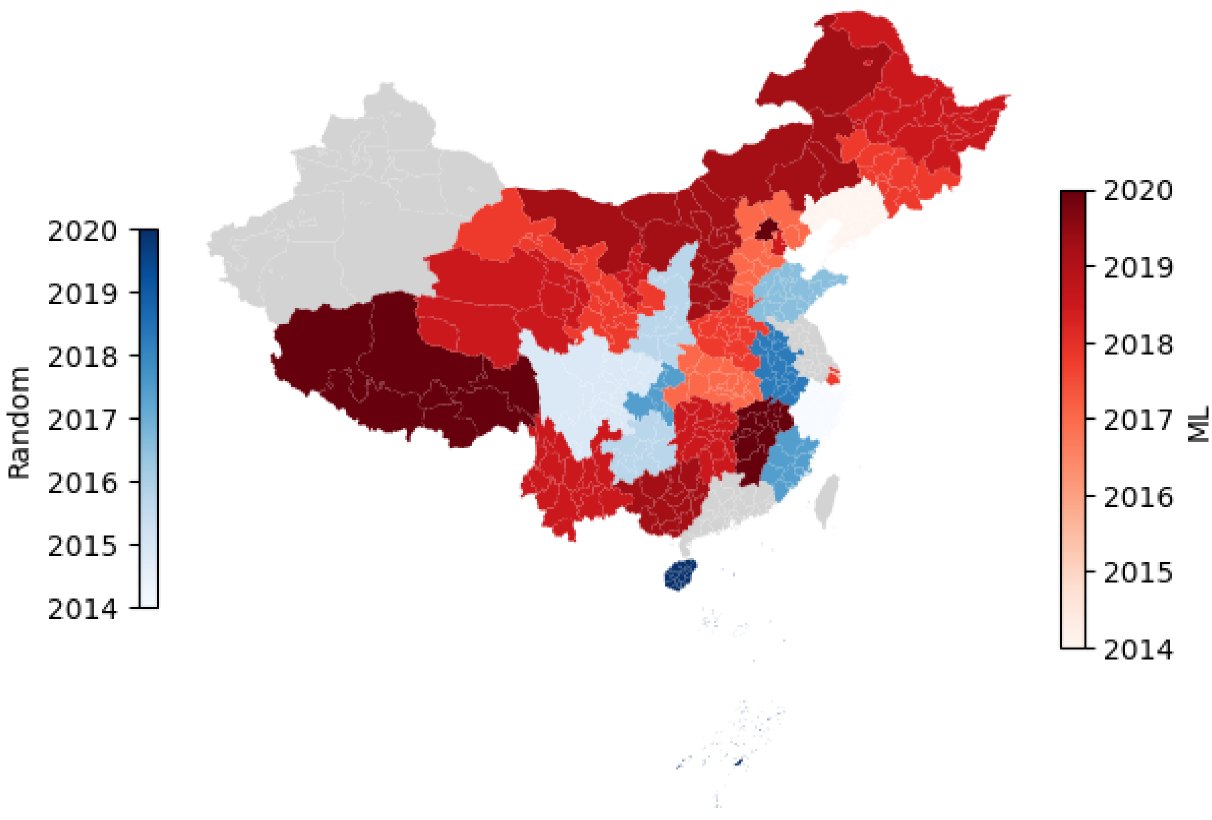
A.1 Figures

Figure A1. Flow Chart of Judicial Process



Notes: This figure illustrates the case assignment process in China. Plaintiffs file the case at the case-filing division. Then, the case-filing division distributes to each division according to the case type. Then, each division assigns the case to the judge(s) by either a human decision maker or an algorithm. Judge(s) only hear cases within their division. All divisions within a court, and all courts within a province, use the same type of case assignment approach.

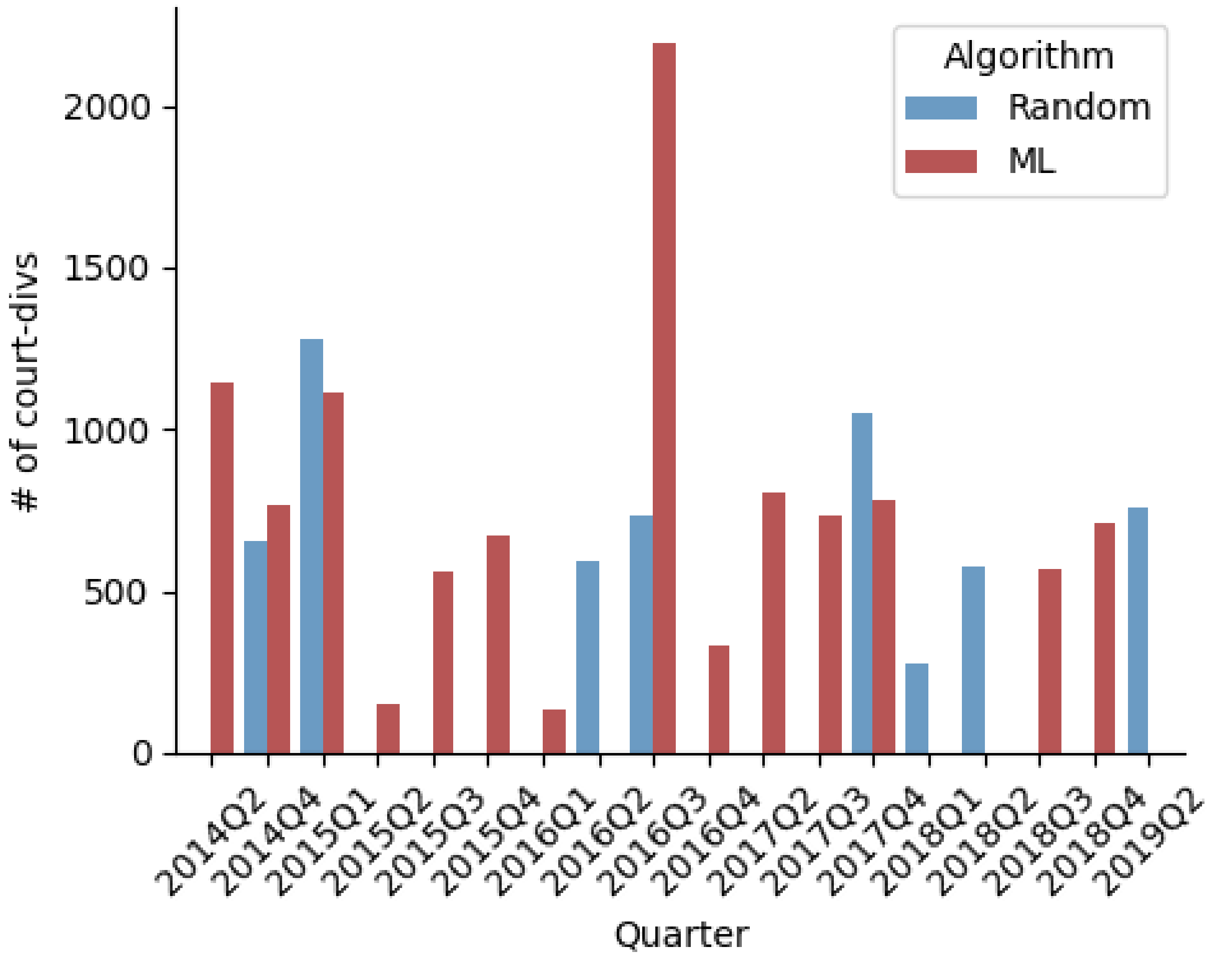
Figure A2. Algorithmic Case Assignment System Adoption by Prefecture



Notes: This figure illustrates the timing and type of case assignment systems in China' s prefectures. The blue (red) gradient scale indicates prefectures using random (ML-based) assignment systems, with darker shades showing later adoption. Gray areas represent prefectures with unknown system details, which are excluded from analysis. Thin white lines denote prefecture borders. According to the data source, prefectures in the same province generally adopt the same system simultaneously. While prefectural variations likely exist, additional validation is needed. This analysis uses uniform timing per province. The future versions will incorporate prefectural variation.

Source: China Government Procurement Online.

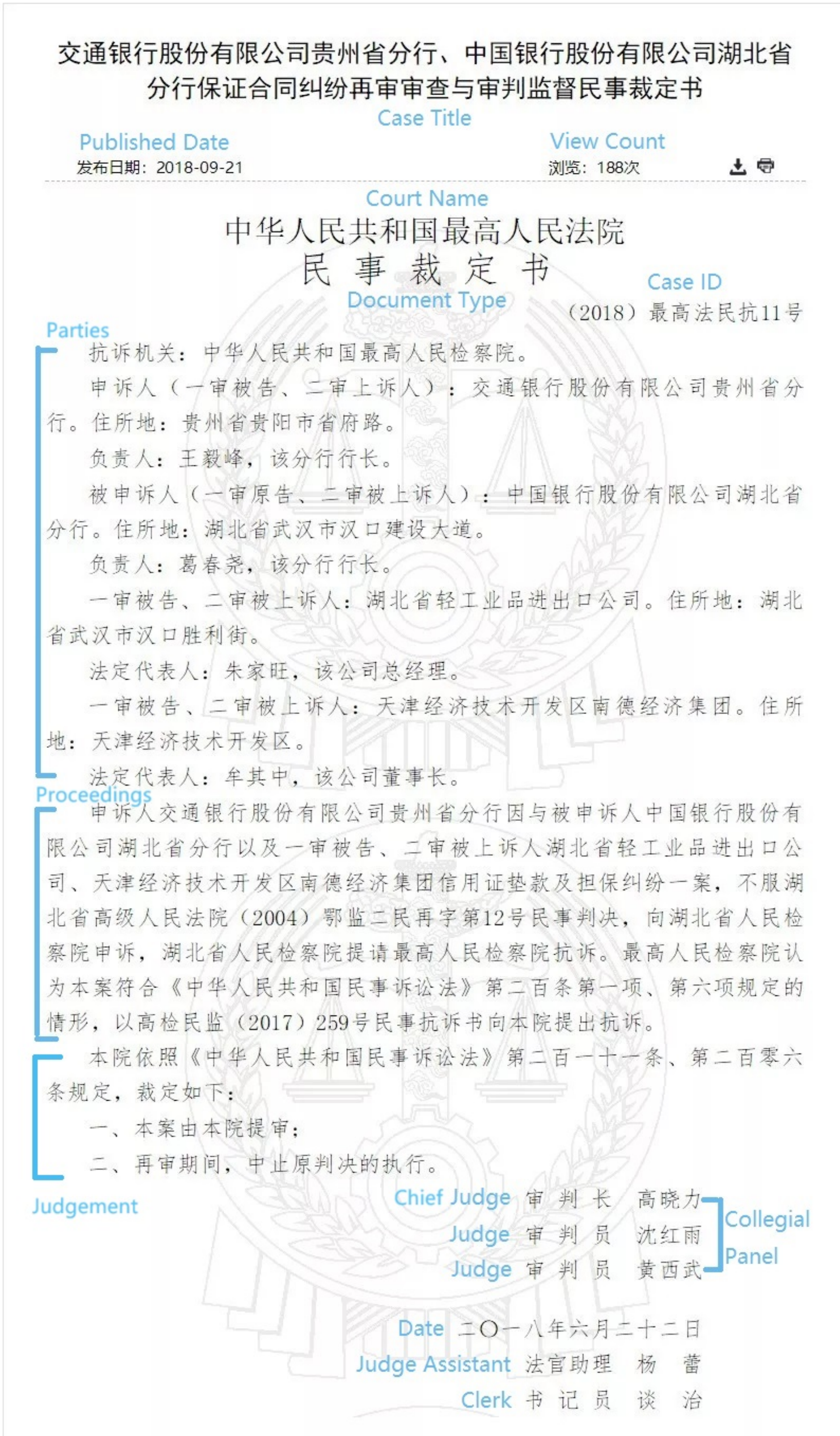
Figure A3. Algorithmic Case Assignment System Adoption by Courts



Notes: This figure illustrates the timing and type of case assignment systems by court. The blue (red) bar indicates courts using random (ML-based) assignment systems. From this figure, we can see the rollout of the algorithmic case assignment from 2014 to 2020. According to the data source, courts in the same province generally adopt the same system simultaneously. While prefectural variations likely exist, additional validation is needed. This analysis uses uniform timing per province. The future versions will incorporate prefectural variation.

Source: China Government Procurement Online.

Figure A4. Case Document Example



交通银行股份有限公司贵州省分行、中国银行股份有限公司湖北省分行保证合同纠纷再审审查与审判监督民事裁定书

发布日期：2018-09-21　　　　浏览：188次

中华人民共和国最高人民法院

民 事 裁 定 书

（2018）最高法民抗11号

抗诉机关：中华人民共和国最高人民检察院。

申诉人（一审被告、二审上诉人）：交通银行股份有限公司贵州省分行。住所地：贵州省贵阳市省府路。

负责人：王毅峰，该分行行长。

被申诉人（一审原告、二审被上诉人）：中国银行股份有限公司湖北省分行。住所地：湖北省武汉市汉口建设大道。

负责人：葛春尧，该分行行长。

一审被告、二审被上诉人：湖北省轻工业品进出口公司。住所地：湖北省武汉市汉口胜利街。

法定代表人：朱家旺，该公司总经理。

一审被告、二审被上诉人：天津经济技术开发区南德经济集团。住所地：天津经济技术开发区。

法定代表人：牟其中，该公司董事长。

申诉人交通银行股份有限公司贵州省分行因与被申诉人中国银行股份有限公司湖北省分行以及一审被告、二审被上诉人湖北省轻工业品进出口公司、天津经济技术开发区南德经济集团信用证垫款及担保纠纷一案，不服湖北省高级人民法院（2004）鄂监二民再字第12号民事判决，向湖北省人民检察院申诉，湖北省人民检察院提请最高人民检察院抗诉。最高人民检察院认为本案符合《中华人民共和国民事诉讼法》第二百条第一项、第六项规定的情形，以高检民监（2017）259号民事抗诉书向本院提出抗诉。

本院依照《中华人民共和国民事诉讼法》第二百一十一条、第二百零六条规定，裁定如下：

一、本案由本院提审；

二、再审期间，中止原判决的执行。

审 判 长　高晓力

审 判 员　沈红雨

审 判 员　黄西武

二〇一八年六月二十二日

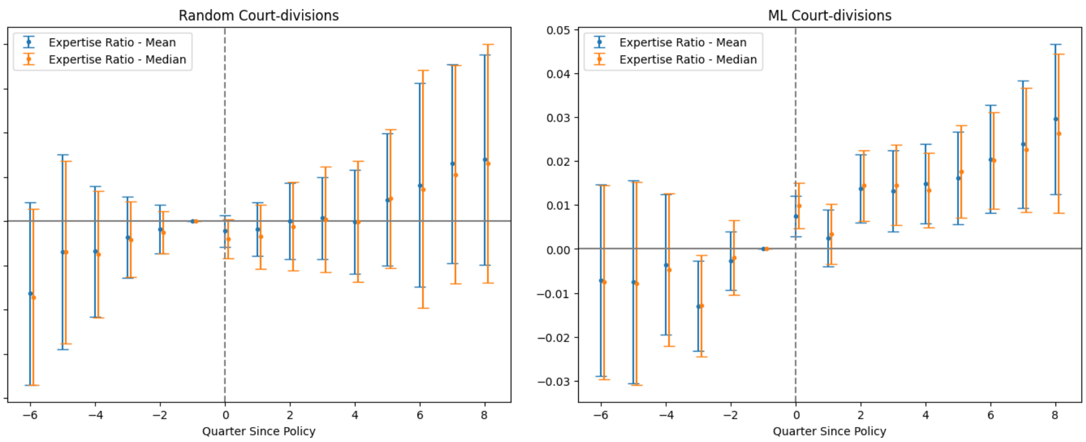
法官助理　杨　蕾

书 记 员　谈　治

Notes: This figure shows an example of the case document. A case document has of four parts: title, fact, verdict and closing. The title contains case title, case type, court name, document type and case ID. The fact contains litigant(s) name, their information, facts, evidence, previous proceeding and related case IDs. The verdict contains the judgment and the reasoning. The closing contains the judge(s) and clerk(s) names and the date of judgment.

Source: China Judgment Online.

Figure A5. Event Study of Algorithmic Assignment on Judge Expertise Ratio

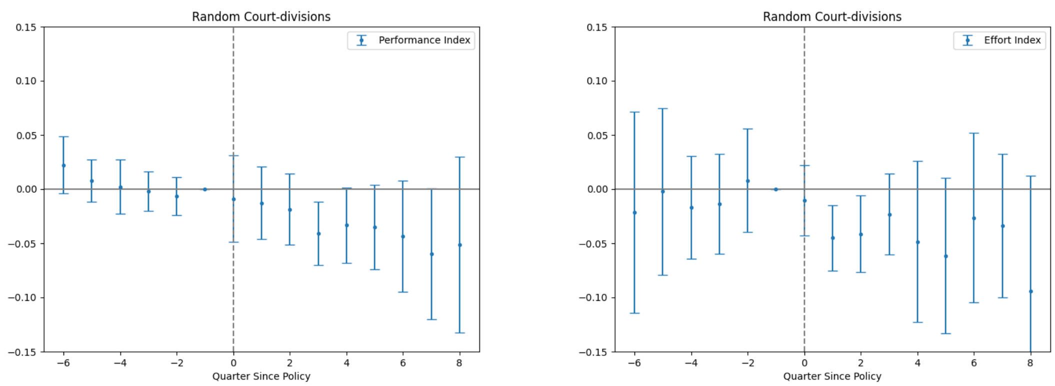


(a) Random Assignment

(b) ML Assignment

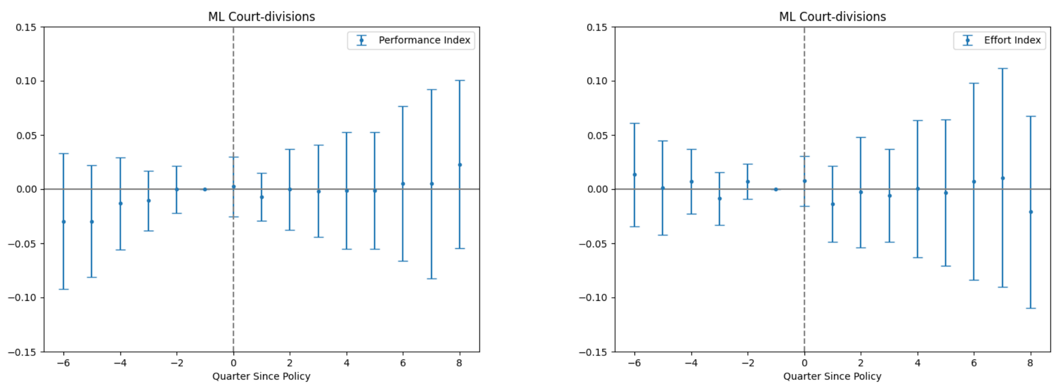
Notes: The figure presents event-study estimates of the effects of random/ML case assignment system on expertise ratio of judges. The outcome variable, judge expertise ratio, is measured as the share of cases within a judge's specialization out of the total cases assigned to the judge, I take the mean (blue) and median (orange) across all judges within a court-division in each quarter. The left panel contains the random court-divisions. The right panel contains the ML court-divisions. Only the right panel display post-treatment and anticipatory effects, with 95% confidence intervals, derived from an event-study model corresponding to Equation 1. The unit of analysis is court-division and the unit of time is quarter. The treatment group includes court-divisions in provinces already implementing random/ML assignment, while the control group comprises not-yet-implemented random/ML assignment court-divisions. Control variables include caseload inflow, and standard errors are clustered at the provincial level.

Figure A6. Event Study of Random Assignment on Performance Index and Efforts Index



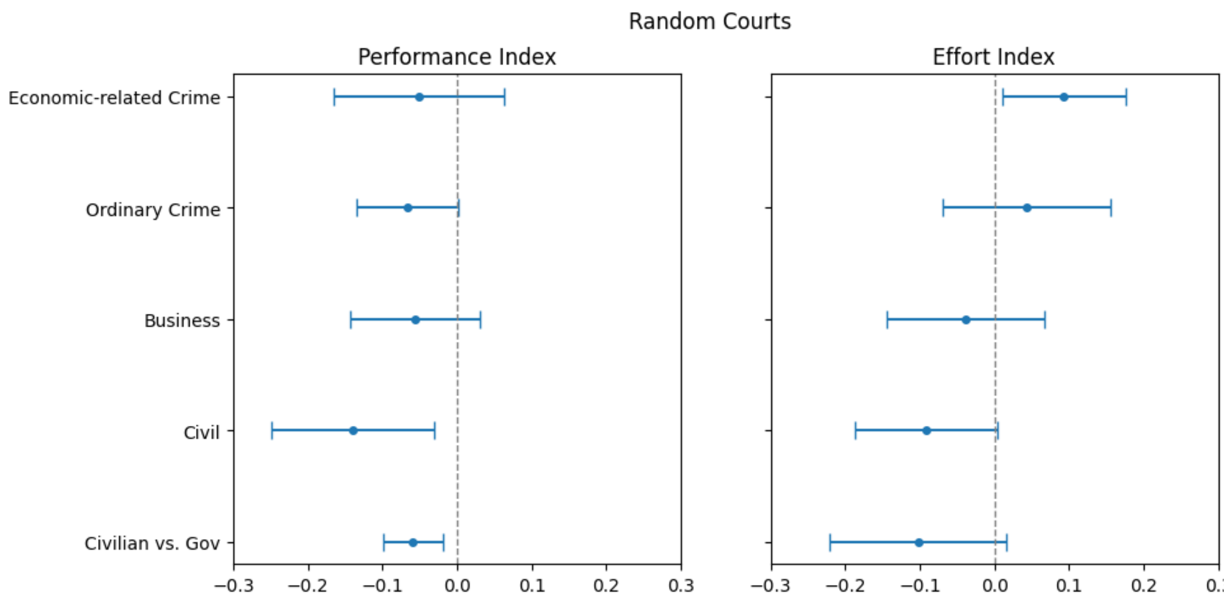
Notes: The figure presents event-study estimates of the effects of a random case assignment system on performance (output) and effort (input) indices. The left panel uses the performance index as the outcome variable, a weighted average of normalized # of case processed, not appeal rate and not reversal rate according to Anderson (2008). The right panel uses the effort index, a weighted average of normalized verdict word count, # of provisions cited and # of provisions written according to Anderson (2008). Both panels display post-treatment and anticipatory effects, with 95% confidence intervals, derived from an event-study model corresponding to Equation 1. The unit of analysis is court-division and the unit of time is quarter. The treatment group includes court-divisions in provinces already implementing random assignment, while the control group comprises not-yet-implemented random assignment court-divisions. Control variables include caseload inflow, and standard errors are clustered at the provincial level.

Figure A7. Event Study of ML-Based Assignment on Performance Index and Efforts Index



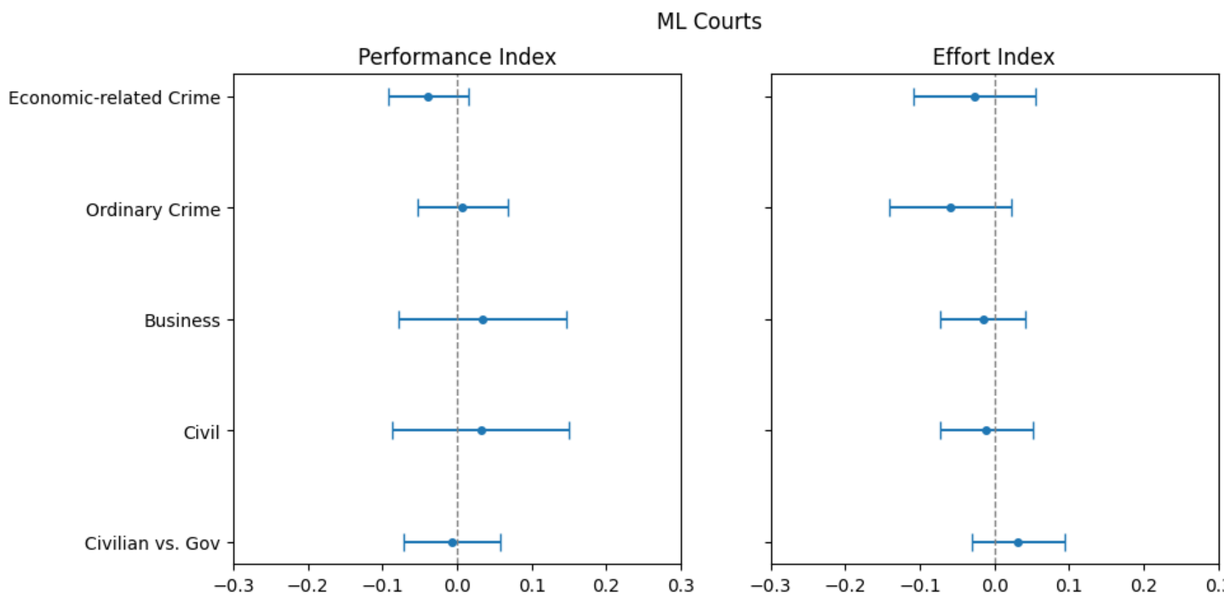
Notes: The figure presents event-study estimates of the effects of a ML-based case assignment system on performance (output) and effort (input) indices. The left panel uses the performance index as the outcome variable, a weighted average of normalized # of case processed, not appeal rate and not reversal rate according to Anderson (2008). The right panel uses the effort index, a weighted average of normalized verdict word count, # of provisions cited and # of provisions written according to Anderson (2008). Both panels display post-treatment and anticipatory effects, with 95% confidence intervals, derived from an event-study model corresponding to Equation 1. The unit of analysis is court-division and the unit of time is quarter. The treatment group includes court-divisions in provinces already implementing ML-based assignment, while the control group comprises not-yet-implemented ML-based assignment court-divisions. Control variables include caseload inflow, and standard errors are clustered at the provincial level.

Figure A8. ATT of Random Assignment on Performance Index and Efforts Index across Divisions



Notes: The figure shows the average treatment effects of an random case assignment system on performance (output) and effort (input) indices across five court divisions: civil vs. government, civil, business, ordinary crime, and economic-related crime. The left panel uses the performance index as the outcome variable, a weighted average of normalized # of case processed, not appeal rate and not reversal rate according to Anderson (2008). The right panel uses the effort index, a weighted average of normalized verdict word count, # of provisions cited and # of provisions written according to Anderson (2008). Both panels display post-treatment effects with 95% confidence intervals, derived from the CSDiD corresponding to Equation **??**. Each coefficient represents a separate regression. The unit of analysis is court-division and the unit of time is quarter. The treatment group includes court-divisions in provinces already implementing random assignment, while the control group comprises not-yet-implemented random assignment court-divisions. Control variables include caseload inflow, and standard errors are clustered at the provincial level.

Figure A9. ATT of ML-Based Assignment on Performance Index and Efforts Index across Divisions



Notes: The figure shows the average treatment effects of an ML-based case assignment system on performance (output) and effort (input) indices across five court divisions: civil vs. government, civil, business, ordinary crime, and economic-related crime. The left panel uses the performance index as the outcome variable, a weighted average of normalized # of case processed, not appeal rate and not reversal rate according to Anderson (2008). The right panel uses the effort index, a weighted average of normalized verdict word count, # of provisions cited and # of provisions written according to Anderson (2008). Both panels display post-treatment effects with 95% confidence intervals, derived from the CSDiD corresponding to Equation **??**. Each coefficient represents a separate regression. The unit of analysis is court-division and the unit of time is quarter. The treatment group includes court-divisions in provinces already implementing ML-based assignment, while the control group comprises not-yet-implemented ML-based assignment court-divisions. Control variables include caseload inflow, and standard errors are clustered at the provincial level.

A.2 Tables

Table A1. Summary Statistics of the Random and ML-Based Court-Divisions Before Treatment

Variable	Random	ML	Difference
Panel A: Case Characteristics Before Algo			
# of Cause of Actions (COA)	0.99	1.00	-0.01***
	(0.17)	(0.17)	(0.00)
$ involved (Yuan)	4572965.36	4889624.26	-316658.90
	(30535612.29)	(34873296.30)	(90646.96)
# of litigants	2.43	2.40	0.03**
	(1.22)	(1.40)	(0.00)
Ratio of corporate litigants	0.30	0.27	0.03
	(0.30)	(0.30)	(0.00)
Ratio of female litigants	0.08	0.09	-0.01
	(0.11)	(0.13)	(0.00)
Panel B: Judge Characteristics Before Algo			
Predicted gender	0.29	0.29	-0.002***
	(0.24)	(0.25)	(0.00)
Experience (quarters)	9.84	7.05	2.79***
	(7.18)	(5.71)	(0.02)
Historical # cases processed/quarter	49.43	33.57	15.87
	(57.46)	(55.61)	(0.16)
Historical NOT appeal rate	0.79	0.76	0.03
	(0.16)	(0.17)	(0.00)
Historical NOT reversal rate	0.99	0.98	0.01
	(0.03)	(0.05)	(0.00)
Historical verdict word count	535.67	575.16	-39.48
	(227.00)	(251.71)	(0.66)
Historical # of provisions cited	1.73	1.82	-0.10
	(0.70)	(0.74)	(0.00)
Historical # of provisions written	0.86	0.88	-0.02
	(0.57)	(0.62)	(0.00)

Notes: This table shows the case and judge characteristics of the random and ML-based court-divisions before the algorithm adoption. Column (1) shows the pre-treatment average of random court-divisions, column (2) shows the pre-treatment average of the ML-based divisions, and column (3) shows the difference between two groups. Panel A presents the case characteristics, in which I use the first four to calculate case complexity. Panel B presents the judge characteristics. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A2. Summary Statistics of the Random and ML-Based Court-Divisions Before Treatment

Variable	Random	ML	Difference
Panel A: Outcomes before Algo - Performance			
Performance Index	0.06	-0.05	0.12***
	(0.58)	(0.74)	(0.00)
# of cases processed/quarter	66.81	44.91	21.89***
	(161.42)	(167.59)	(0.36)
NOT appeal rate	0.79	0.77	0.02***
	(0.26)	(0.28)	(0.00)
NOT reversal rate	0.98	0.97	0.01***
	(0.08)	(0.10)	(0.00)
Panel B: Outcomes before Algo - Effort			
Effort Index	0.05	-0.04	0.09***
	(0.75)	(0.80)	(0.00)
Verdict word count	502.87	484.31	18.57***
	(403.94)	(416.85)	(0.89)
# of provisions cited	1.62	1.54	0.08***
	(1.05)	(1.09)	(0.00)
# of provisions written	0.78	0.71	0.07***
	(0.97)	(0.94)	(0.00)

Notes: This table shows the performance and effort of the random and ML-based court-divisions before the algorithm adoption. Column (1) shows the pre-treatment average of random court-divisions, column (2) shows the pre-treatment average of the ML-based divisions, and column (3) shows the difference between two groups. Panel A and panel B presents the performance index and compositing variables and effort index and compsiting variables, respectively. Panel B presents the judge characteristics. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A3. ATT of Random Assignment on Performance Index and Efforts Index

	Panel A: Outcome Variables - Performance					
Index	**Normalized Variables**			**Raw Variables**		
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Performance	# Cases Processed	Not Appealed Rate	Not Reversal Rate	# Cases Processed	Appeals	Reversals
-0.054***	0.051	-0.110*	-0.060	8.056	6.066**	0.209
(0.02)	(0.04)	(0.06)	(0.04)	(5.00)	(2.44)	(0.17)
	Panel B: Outcome Variables - Effort Index					
Index	**Normalized Variables**			**Raw Variables**		
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Effort	Verdict Word Count	Provisions Cited	Provisions Written	Verdict Word Count	# Provisions Cited	# Provisions Written
-0.114***	-0.097***	-0.097***	0.012	-46.533***	-0.096***	-0.012
(0.02)	(0.04)	(0.03)	(0.03)	(12.57)	(0.02)	(0.02)
Unit	Court-division					
Cluster	Province					
Obs	350,552					

Notes: This table presents the ATT of random assignment system on performance index, effort index, the normalized and raw elements composing the index. In Panel A, Column (1) shows the ATT on the performance index. Columns (2)-(4) display normalized variables (# cases processed, not appealed rate, and not reversal rate) that compose the index, while Columns (5)-(7) show their raw counterparts (# cases processed, appeals, and reversals). In Panel B, Column (1) shows the ATT on the effort index. Columns (2)-(4) include normalized variables (verdict word count, provisions cited, and provisions written), and Columns (5)-(7) show the raw counterparts for these components. The treatment group includes court-divisions in provinces already implementing random assignment, while the control group comprises not-yet-implemented random assignment court-divisions. Control variables include caseload inflow, and standard errors are clustered at the provincial level. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A4. ATT of ML Assignment on Performance Index and Efforts Index

			Panel A: Outcome Variables - Performance			
Index		**Normalized Variables**			**Raw Variables**	
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Performance	# Cases Processed	Not Appealed Rate	Not Reversal Rate	# Cases Processed	Appeals	Reversals
0.025	0.001	0.042	0.023	-3.289	-0.889	-0.155
(0.03)	(0.04)	(0.06)	(0.03)	(5.25)	(1.46)	(0.12)
			Panel B: Outcome Variables - Effort Index			
Index		**Normalized Variables**			**Raw Variables**	
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Effort	Verdict Word Count	Provisions Cited	Provisions Written	Verdict Word Count	# Provisions Cited	# Provisions Written
-0.010	0.021	0.002	-0.031	13.150	0.018	-0.009
(0.03)	(0.04)	(0.04)	(0.03)	(7.21)	(0.03)	(0.01)
Unit				Court-division		
Cluster				Province		
Obs				534,601		

Notes: This table presents the ATT of ML-based assignment system on performance index, effort index, the normalized and raw elements composing the index. In Panel A, Column (1) shows the ATT on the performance index. Columns (2)-(4) display normalized variables (# cases processed, not appealed rate, and not reversal rate) that compose the index, while Columns (5)-(7) show their raw counterparts (# cases processed, appeals, and reversals). In Panel B, Column (1) shows the ATT on the effort index. Columns (2)-(4) include normalized variables (verdict word count, provisions cited, and provisions written), and Columns (5)-(7) show the raw counterparts for these components. The treatment group includes court-divisions in provinces already implementing ML-based assignment, while the control group comprises not-yet-implemented ML-based assignment court-divisions. Control variables include caseload inflow, and standard errors are clustered at the provincial level. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A5. DFL Decomposition in Mean Sentencing Outcome, Random vs. ML Courts

	Panel A: Random Courts				
(1) Effect	(2) Total Custody Term	(3) Immediate Custody	(4) If Probation	(5) Probation Length	(6) Fine (Chinese Yuan)
Observed difference	-0.133	-0.028	0.015	-0.143	-596.982
	(0.011)***	(0.013)**	(0.005)***	(0.020)***	(47.265)***
Structure effect	-0.190	0.012	-0.014	-0.290	-202.754
	(0.042)***	(0.060)	(0.022)	(0.110)***	(173.915)
Obs			39,504		
	Panel B: ML Courts				
(1) Effect	(2) Total Custody Term	(3) Immediate Custody	(4) If Probation	(5) Probation Length	(6) Fine (Chinese Yuan)
Observed difference	-0.039	-0.167	0.105	0.177	641.422
	(0.015)***	(0.018)***	(0.007)***	(0.029)***	(56.352)***
Structure effect	-0.121	-0.101	0.036	-0.104	298.027
	(0.057)**	(0.063)	(0.026)	(0.115)	(211.008)
Obs			46,128		
Unit			Case		
Year & Province FE			Yes		

Notes: This table presents results from DFL decomposition analysis of sentencing outcomes for 10% stratified sample of dangerous driving cases. The decomposition separates observed mean differences between pre- and post-algorithm periods into two components: the composition effect (explained by changes in observable case characteristics) and the structure effect (unexplained differences). Each outcome displays the observed difference and structure effect followed by standard errors in parentheses, calculated using 100 bootstrap replications. Panel A examines random assignment courts; Panel B examines machine learning assignment courts. Column (2)-(6) show five sentencing outcomes: total custody sentence, consecutive custody sentence, whether probation, probation length, and monetary fine. Control variables include blood alcohol content, traffic accident severity, flight from scene, expression of remorse, prior criminal record, valid license status, and time and province fixed effects. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table A6. DFL Decomposition in Sentencing Outcome Variance, Random vs. ML Courts

			Panel A: Random Courts		
(1) Effect	(2) Total Custody Term	(3) Immediate Custody	(4) If Probation	(5) Probation Length	(6) Fine (Chinese Yuan)^2
Observed difference	-0.149	-0.057	-0.001	-0.934	-5506384.852
	(0.017)***	(0.028)**	(0.000)***	(0.065)***	(755971.549)***
Structure effect	0.830	1.402	0.247	2.932	11696652.228
	(0.011)***	(0.015)***	(0.000)***	(0.026)***	(347606.210)***
Obs			39,504		
			Panel B: ML Courts		
(1) Effect	(2) Total Custody Term	(3) Immediate Custody	(4) If Probation	(5) Probation Length	(6) Fine (Chinese Yuan)^2
Observed difference	-0.089	0.016	0.005	-0.896	666474.325
	(0.025)***	(0.042)	(0.001)***	(0.103)***	(1019598.728)
Structure effect	1.118	1.548	0.249	4.013	20038265.851
	(0.010)***	(0.015)***	(0.000)***	(0.035)***	(451326.766)***
Obs			46,128		
Unit			Case		
Year & Province FE			Yes		

Notes: This table presents results from DFL decomposition analysis of sentencing outcome variance for dangerous driving cases. The decomposition separates observed variance differences between pre- and post-algorithm periods into two components: the composition effect (explained by changes in observable case characteristics) and the structural effect (unexplained differences). Each outcome displays the observed difference followed by standard errors in parentheses, calculated using 100 bootstrap replications. Panel A examines random assignment courts; Panel B examines machine learning assignment courts. Columns show five sentencing outcomes: total custody sentence, consecutive custody sentence, probation assignment, probation length, and monetary fine. Control variables include blood alcohol content, traffic accident severity, flight from scene, expression of remorse, prior criminal record, valid license status, and time and province fixed effects. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

B Case Assignment: Judge-Case Correlation

While expertise matching examines whether the right judge handles each case, it is equally important to understand how the burden of difficult cases is distributed across judges. In this section, I examine how judge characteristics relate to case complexity before and after the introduction of random and ML-based assignment systems.

These systems are expected to influence judge-case correlations in different ways. A random assignment system is designed to eliminate such correlations by assigning cases without considering judge characteristics. This breaks existing links between characteristics like experience, past performance, and effort, and the types of cases judges receive. In contrast, an ML-based system uses historical data to optimize assignments. It may replicate prior patterns by learning from past assignments or introduce small adjustments to improve match quality through predictive modeling.

To estimate these changes, I estimate a regression of case complexity on judge characteristics, including an interaction with the policy indicator. The analysis includes all judges who served in the same court both before and after the reform:

$$\text{case\_complexity}_i = \beta_0 + \beta_1 \text{Post}_t + \beta_2 \text{judge\_char}_{jt} + \beta_3 (\text{judge\_char} \times \text{Post})_{jt} + \epsilon_{ijt} \qquad (7)$$

In this model, case_complexity_i represents the complexity index for case i, calculated from pre-assignment factors such as the cause of action, monetary stakes, and litigant composition.[30] The binary variable Post_t indicates whether quarter t occurs after the policy's implementation, distinguishing pre- and post-policy periods. The vector judge_char_{jt} includes characteristics of judge j at time t, such as historical case complexity, historical performance, historical effort, gender, and experience (normalized)[31]. The interaction term $(\text{judge\_char} \times \text{Post})_{jt}$ captures any shifts in the relationship between judge characteristics and case complexity due to the policy. Standard error is clustered at the provincial level, as in the main regression.

The key coefficients are β_2 and β_3. The coefficient β_2 represents the baseline correlation between judge characteristics and case complexity under the previous system. In contrast, β_3 reflects changes in this correlation following the implementation of random or ML-based assignments. By analyzing these coefficients, we can determine the extent to which each system changes, preserves, or disrupts existing assignment patterns[32].

[30]The complexity index reduces the dimensionality of multiple indicators into a single measure, based on Anderson (2008). All variables are determined before assignment to ensure they are exogenous to the judge's influence. Alternative dimensionality reduction methods like PCA yielded principal components with less than 30% explanatory power.

[31]Historical complexity is calculated by the judge's average complexity index before the policy. Historical performance and effort indices are calculated as averages over prior quarters, following the method in Anderson (2008).

[32]It is worth noting potential limitations of this difference-in-differences (DiD) approach. As explained in section 2, staggered policy implementation across provinces may introduce bias in traditional DiD estimates. Although generalized DiD methods address staggered adoption, they focus on β_3 and do not provide information on β_2, which is essential for understanding baseline correlations. Furthermore, generalized DiD methods often require binary variables, limiting the precision of continuous judge characteristics. While CSDiD methods may offer additional insights, they require substantial data preprocessing, which risks perceptions of data manipulation. For this analysis, the combination of binary (gender) and normalized continuous variables (indices and experience) is expected to

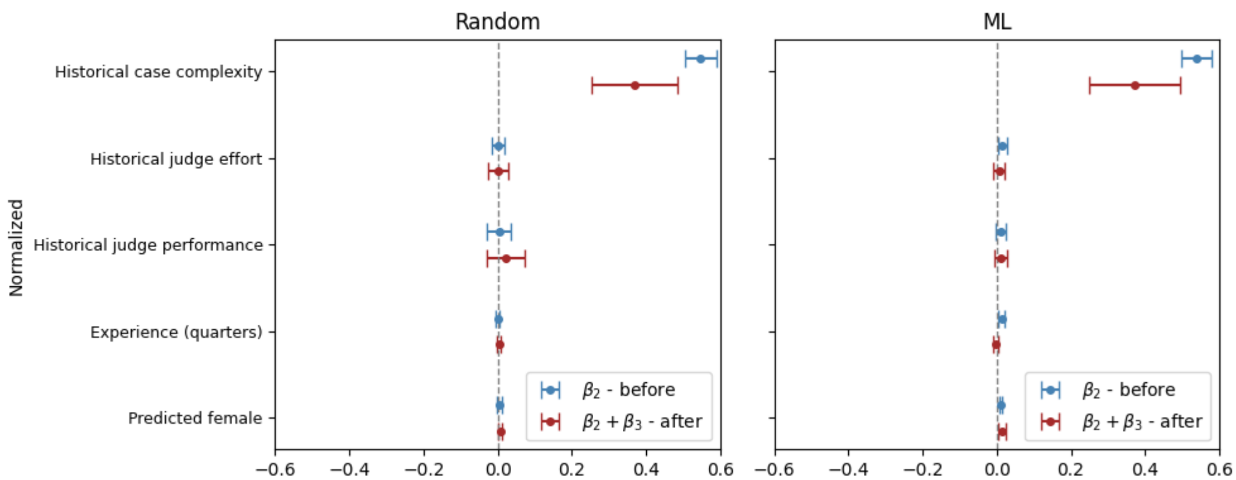
Pre-policy Judge-Case Correlation The blue ink in Figure A10 presents the baseline correlations between judge characteristics and case complexity before the implementation of random and ML-based assignment, captured by β_2. In both random and ML court divisions, judges' current case complexity is strongly correlated with their historical complexity. A one standard deviation increase in historical complexity is associated with a 0.54 (random) and 0.56 (ML) standard deviation increase in current complexity. In random court divisions, other characteristics, such as historical performance, effort, experience, and gender, show no significant correlation with case complexity. In ML court divisions, historical effort and experience are positively correlated with case complexity, but the effect sizes are limited.

Post-Policy Judge-Case Correlation The red ink in Figure A10 shows the post-policy correlations, $\beta_2 + \beta_3$, for both systems. While historical complexity remains positively correlated with current case complexity in both random and ML court divisions, the strength of this relationship is noticeably reduced compared to the pre-policy period. In random divisions, a one standard deviation increase in historical complexity now corresponds to a 0.42 standard deviation increase in current complexity; in ML divisions, the figure is slightly higher at 0.43. This difference is significant at the 1% level. In ML courts, the correlation between experience and case complexity becomes statistically insignificant, suggesting that experience no longer affects case assignments.

Overall, both systems reduce reliance on traditional assignment patterns, particularly those based on past case complexity. By the post-policy period, neither system shows significant correlations between other judge characteristics and case complexity.

capture judge-case correlation changes effectively.

Figure A10. Case Assignment Pattern Relative to Case Complexity Index Change Due to Random/ML



Notes: This figure illustrates how judge characteristics correlate with case complexity before and after the transition to algorithmic assignment systems. The analysis estimates the following equation:

$$\text{case\_complexity}_i = \beta_0 + \beta_1 \text{Post}_t + \beta_2 \text{judge\_char}_{jt} + \beta_3 (\text{judge\_char} \times \text{Post})_{jt} + \epsilon_{ijt} \quad (8)$$

The figure displays β_2 (pre-reform correlation) in blue and $\beta_2 + \beta_3$ (post-reform correlation) in red. The left panel shows courts that transitioned to random assignment, while the right panel shows those that transitioned to ML-based assignment. The analysis includes all cases involving judges observed in both pre- and post-reform periods. Historical case complexity, judge effort, and judge performance indices are standardized weighted averages normalized to pre-treatment observations. Experience is similarly normalized. Predicted female is a binary variable indicating female judges based on name prediction. The difference between β_2 and $\beta_2 + \beta_3$ for historical case complexity is statistically significant at the 1 percent level in both random and ML court divisions. The difference in experience is also statistically significant at the 1 percent level in ML court divisions. The other differences are not significant.