ARTICLES

BARRIERS TO ADOPTING PREDICTIVE ALGORITHMS: A CRIMINAL JUSTICE FIELD EXPERIMENT

David S. Abrams, Brian M. Collopy, Aurelie Ouss, Megan Stevenson, and Colin D. Sullivan*

ABSTRACT

Artificial intelligence, machine learning, and algorithmic prediction tools have made huge advances in recent years. Some argue that we are on the precipice of a major revolution in our economy and society, in which eager adoption of these new technologies will transform how work is done.

This Article argues that change might come more slowly to the legal sphere than is commonly thought. We present the results of a criminal justice field experiment in which we provided novel sentence prediction software to public defenders. In some regards, the experiment was a failure. Usage of the prediction software was so low that we were unable to evaluate its impact on sentencing. This is despite strong a priori expressions of interest and tests showing that our algorithm is more accurate than the public defenders at predicting sentences.

However, this failure produced valuable insights about why predictive AI might face headwinds in the legal profession. Extensive interviews, a prediction "quiz," and our empirical results revealed the following takeaways. First, attorneys place a high bar on adopting new technology due both to workflow inertia and skepticism about benefits. Second, some attorneys were distrustful of an algorithm that did not have all of the information they had—even if the algorithm still provided more accurate information than their intuition. Third, algorithm design entails challenging ethical questions that can reduce trust and use among users. We discuss these issues in detail and suggest some possible paths forward.

---

* David S. Abrams: William B. and Mary Barb Johnson Professor of Law and Economics, Professor of Business Economics and Public Policy, University of Pennsylvania; Brian M. Collopy: Economist, Amazon (we note that this work was done prior to his joining Amazon); Aurelie Ouss: Assistant Professor of Criminology, University of Pennsylvania; Megan Stevenson: Professor of Law, University of Virginia; Colin Sullivan: Assistant Professor of Economics, University of Pittsburgh. Thanks to the Clark County Public Defender office and the wonderful attorneys and staff members who took time to make this project possible. Thanks to dedicated project managers Joseph Boskovski and Brittany Keesling. Thanks for amazing research assistance to Priyanka Goonetilleke, Connor Gregory, Jessie Levin, Danielle Moors, Alejandro Sanchez-Becerra, Ruth Woldemichael. Thanks also to JJ Prescott and Court Innovations. Additional thanks to the substantial excellent feedback at various stages of the project from numerous colleagues at Penn Carey Law, Wharton and throughout the legal and economics academy. Funding through a grant by Arnold Ventures is gratefully acknowledged. © 2026, David S. Abrams, Brian M. Collopy, Aurelie Ouss, Megan Stevenson & Colin D. Sullivan.

INTRODUCTION 50
I. ALGORITHMIC & AI PREDICTION FOR CRIMINAL JUSTICE EXPERTS . . 56
A. *Algorithmic Predictions in Law and Criminal Justice* 56
B. *The Potential of Predictive Algorithms in Public Defense* 59
II. EXPERIMENTAL DESIGN 62
A. *Criminal Justice in Clark County, Nevada* 62
B. *Algorithm Development* 64
C. *Interface Design and Testing* 67
D. *Experimental Setup* 70
III. EXPERIMENTAL RESULTS 71
A. *Overall Software Usage and Correlates* 72
B. *Algorithmic Predictions were Substantially Better than Attorneys* 75
C. *Attorney Feedback on the Software* 77
D. *Unique Circumstances Decreased Usage* 79
IV. IMPLICATIONS FOR AI IN CRIMINAL JUSTICE AND BEYOND 80
A. *The Bar is High for Algorithmic Performance* 80
B. *Barriers to Accuracy: Missing, Incomplete or Inaccessible Data* 82
C. *When "Fairness" Depends on Accuracy as Much as Bias* 84
CONCLUSION 85
APPENDIX TABLES 87

INTRODUCTION

A first-year public defender sits down at her desk with a stack of twenty open case files, all with plea offers from the prosecutor that she is expected to respond to in short order. Several of the cases are low-level drug offenses by offenders with minimal records—she has handled dozens of them before and they will almost certainly be resolved by plea with sentences of probation. But a few of the cases give her pause. One is an arson case, which she has never seen before. A couple others involve defendants with minor charges but lengthy criminal histories. Another is before a judge known to have particularly harsh views about low-level drug offenses.

In the normal practice of many public defender (PD) offices, she would have extremely limited time to devote to each case.[1] She could get advice from more senior attorneys, if they can be quickly located and have experience with the tricky cases. But experienced public defenders are in high demand and often burdened with the most complex cases.[2]

---

1. NICHOLAS M. PACE, MALIA N. BRINK, CYNTHIA G. LEE & STEPHEN F. HANLON, NATIONAL PUBLIC DEFENSE WORKLOAD STUDY 4 (2023).

2. Jule Pattison-Gordon, *Staffing and Funding Shortages Persist for Public Defenders*, GOVERNING (May 20, 2025), https://www.governing.com/workforce/staffing-and-funding-shortages-persist-for-public-defenders [https://perma.cc/DW3S-PFF3].

Imagine if this defender had access to algorithmic sentencing predictions that incorporate information from tens of thousands of cases and a huge range of case, defendant, and court characteristics. Using these predictions, she would be able to more effectively and efficiently handle not only the minor and familiar cases, but also the challenging and unfamiliar ones. When advising clients, she could recommend rejecting sentencing offers that are too high relative to what her experience and the algorithm suggest. This could improve outcomes for her clients, increase the amount of time she has to do work on other cases, and decrease idiosyncratic disparities in sentencing across attorneys.

This is not a fanciful scenario, but a reality made possible by a randomized field experiment that we performed in the Clark County (home of Las Vegas) Public Defender's office. Our five-person team spent several years developing a sentence prediction algorithm for public defenders to consult about the likely sentences for their cases. The algorithm incorporated eleven years of rich administrative data, including information on sentences as well as various factors likely to influence the sentence: the alleged offense, the defendant's criminal record, demographic and socioeconomic characteristics, as well as the courtroom the case was assigned to. We partnered with a software developer to build a user-friendly website that defenders could use to access the predictions. Throughout it all, we worked closely with the public defender's office to understand their needs and concerns. A member of our team flew to Las Vegas monthly to have face-to-face meetings, gather feedback, and improve our tool. We went through multiple iterations to optimize the user experience for defenders. Once the software was launched, a team member conducted one-on-one training sessions and frequent in-person check-ins to encourage adoption.

Our goal was to evaluate how sentencing predictions affected public defender behavior and case outcomes. We conducted a randomized, controlled trial in which half of the PDs were randomly assigned to receive access to the predictions (treatment group) while the other half (control group) went about their jobs as usual.[3] We hypothesized that our tool could help the public defender better serve her client by providing improved information about expected case outcomes. The predictions would help public defenders know what plea offer is reasonable according to court norms, and reject or negotiate ones that are unreasonable, thus lowering sentences and reducing idiosyncratic variation. We expected that the algorithm would be of

3. There is a large literature in economics about bargaining in which informational advantages improve outcomes. *See, e.g.*, Darryl A. Seale, Terry E. Daniel & Amnon Rapoport, *The Information Advantage in Two-Person Bargaining with Incomplete Information*, 44 J. ECON. BEHAV. & ORG. 177, 198 (2001) (finding that parties endowed with superior information receive significantly larger shares of the surplus in two-person bargaining). Information asymmetries can negatively impact settlement and plea negotiations in the legal context. Joyce Sadka, Enrique Seira & Christopher Woodruff, *Information and Bargaining through Agents: Experimental Evidence from Mexico's Labour Courts*, 91 REV. ECON. STUD. 3677, 3678–79 (2024) (finding that providing plaintiffs with more information increased settlement rates); Stephanos Bibas, *Plea Bargaining Outside the Shadow of Trial*, 117 HARV. L. REV. 2463, 2494–95 (2004) (explaining that information deficits in plea bargaining impede actual accuracy of bargains reached).

particular use to newer public defenders, for those less skilled at evaluating offers, and for PDs facing unusual case types.

Despite the enormous amount of time and resources spent on developing and optimizing the sentence prediction software, *public defenders rarely used the algorithmic predictions*. Overall, PDs consulted the predictions only about 10% of the time. As a result, there was insufficient data to evaluate most of the initial hypotheses. Access to the predictions may have affected sentences, but we lack the statistical power from the experiment to tell. Consistent with our hypothesis, usage of the algorithmic predictions was greatest among the least experienced attorneys. But even our most enthusiastic "power user" only consulted the predictions on less than half of their cases.

While the low usage of algorithmic predictions made it impossible to rigorously address the above hypotheses, it led naturally to the questions that are the subject of this article. Namely, why was there such low uptake of the algorithm, and what does this mean for the impact of algorithmic and AI predictions in the legal workplace?

Our first concern was that perhaps the algorithm did not provide meaningful accuracy improvements over the PDs' experience and intuition. With this in mind, we administered a quiz in which we provided PDs with case scenarios using data from real cases to see how attorney predictions about sentence length compared to those of our algorithm when both were provided with the same inputs. We then compared both the algorithmic and attorney predictions to the actual sentences handed down by judges. We found that the algorithm outperformed the attorneys substantially: the mean absolute error (i.e. the difference between the predictions and the actual sentences, expressed as a positive number) was 70% higher for the attorneys than for the algorithm.

In order to better understand the low usage of our algorithmic predictions, we interviewed the PDs to shed light on the factors hindering their adoption. Drawing from these structured follow-up surveys as well as extensive unstructured interviews, we extracted several lessons that we believe may have broad implications for the adoption of AI tools in criminal justice and other settings where experts must make rapid decisions in high-stakes settings.

First, attorneys set a high bar for adopting algorithmic and AI tools that overlap with their areas of expertise. Even if the tool provides improved information about likely sentences, public defenders were still reluctant to use it. We discuss a number of reasons for this reluctance, including inertia, difficulty establishing new work habits, insufficient role models and starter incentives as well as concerns about accuracy.[4]

Concerns about accuracy bring us to the second point: prediction models will not work as well when data is inaccurate and incomplete. Our sentencing tool was built using 11 years of data and dozens of inputs describing the criminal record, the

4. *See infra* Part III.C.

charges, and demographic information. Yet it did not include all of the data that public defenders thought was important. It did not include the prosecutor assigned to the case, for instance, or details about the charge that were not recorded in the PDs software system. Moreover, PDs expressed concern that our data sources—the case management system used by the public defender's office—had significant inaccuracies. For these reasons, they expressed skepticism about the accuracy of the predictions.

As the results of our quiz demonstrate, our tool was substantially more accurate than public defenders, even if it wasn't perfect. But defenders did not seem interested in consulting a prediction tool that wasn't perfect. In the training sessions, we explained that the tool provided baseline information about expected case outcomes for someone with a given set of characteristics, but that they should adjust up or down based on factors that were not included in the algorithm: details of the offense, their relationship with the prosecutor, etc. In our training sessions, they seemed to understand this and find it reasonable. But in our exit interviews, insufficient data was cited as one of the main reasons for disuse.

While AI performs well in settings where there are vast troves of rich data from which to learn, it will not be as successful when data is more limited. Criminal justice provides an example of this, but it is not the only one.[5] Data today is frequently siloed, collected by different firms or government agencies with different capacities for, and interest in, sharing.[6] Moreover, certain types of data require painstaking entry by low-level administrators who may not always be as careful as one might hope.[7]

Finally, our setting raises some fairness concerns that are different from those usually considered in the criminal justice setting. Comparing sentence predictions to risk assessments (algorithmic predictions of future arrest used to determine pretrial detention, sentencing and so forth) provides an illustration. Both risk assessments and sentence predictions will embed bias from the training data.[8] They are

---

5. Jens Ludwig & Sendhil Mullainathan, *Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System*, 35 J. ECON. PERSP. 71, 72 (2021) [hereinafter *Fragile Algorithms and Fallible Decision-Makers*] ("Decades of empirical work by economists shows that in almost every data application, the data is incomplete, not fully representing either the objectives or the information that decision-makers possess. For example, judges rely on much more information than is available to algorithms, and judges' goals are often not well-represented by the outcomes provided to algorithms.").

6. GILLIAN DIEBOLD, CTR. FOR DATA INNOVATION, OVERCOMING BARRIERS TO DATA SHARING IN THE UNITED STATES 1 (2023), https://www2.datainnovation.org/2023-data-sharing-barriers.pdf [https://perma.cc/5S7P-YX96] (observing that data held by separate agencies and firms often remains siloed due to differing capacities and incentives to share).

7. *See, e.g.*, Aileen Rothbard, *Quality Issues in the Use of Administrative Data Records*, *in* ACTIONABLE INTEL. 77, 86 (John Fantuzzo & Dennis P. Culhane eds., 2015) ("The most common errors generally found are coding and input errors") (citing John M. Abowd & Lars Vilhuber, *The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers*, 23 J. BUS. & ECON. STAT. 133 (2005)); Jorge Martinez, *67 Data Entry Statistics for 2025*, DOCUCLIPPER (Mar. 5, 2025), https://www.docuclipper.com/blog/data-entry-statistics/ [https://perma.cc/G4GD-F3UJ] (noting that manual data-entry "has an error rate as high as 4% . . . humans would commit between 100 and 400 errors [per 10,000 entries]").

8. Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2263 (2018).

trained on data that is the outcome of a racially disparate world, and the predictions will embed those racial disparities. Criminal record data, for instance, is an essential input to both, yet the criminal record is the product of racially disparate policing and prosecution practices.[9]

However, the real-world implications of having racially biased predictions differ between our sentence predictions and standard risk assessment tools. When it comes to the use of risk assessments to determine incarceration decisions, historical biases embedded in the algorithm disadvantage defendants from the biased-against group.[10] Accordingly, the use of race as an input in risk assessments is broadly considered unconstitutional and the use of racial "proxies" is highly controversial.[11] Given these concerns, scholars have focused on designing tools that reduce or eliminate historical biases in algorithms.[12]

Our setting, however, raises a different set of concerns than those considered in the risk assessment context. When the algorithm is trying to predict the behavior of the judge who is deciding sentencing, excluding race or race proxies from the algorithm could disadvantage the defendant, since it gives their attorney less accurate information. If the algorithm falsely leads the public defender to believe that a Black defendant can expect a similar sentence to a white defendant, they may give their clients bad advice. From this perspective, sentence predictions should include race as long as race is a factor in sentencing. On the other hand, including race in the algorithm could serve to re-entrench bias by influencing norms and expectations. Thus, the fairness concerns that arise with including race in sentence predictions are more complex than those in risk assessments.

Based on a number of conversations with stakeholders and other academics, we ultimately did not include race in the algorithm. However, we did use variables that correlated with race, such as zip code. Interviews with the public defenders revealed mixed opinions on the question, and some expressed skepticism in the

---

9. *See, e.g.*, Emma Pierson et al., *A Large-Scale Analysis of Racial Disparities in Police Stops Across the United States*, 4 NATURE HUM. BEHAV. 736, 736 (2020) (finding consistent racial bias in police stop and search rates using a nationwide "veil of darkness" design); M. Marit Rehavi & Sonja B. Starr, *Racial Disparity in Federal Criminal Sentences*, 122 J. POL. ECON. 1320 (2014) (demonstrating that racial disparities in federal sentencing largely arise from prosecutors' charging decisions); Carlos Berdejó, *Criminalizing Race: Racial Disparities in Plea-Bargaining*, 59 B.C. L. REV. 1187, 1214 (2018) (showing significant racial gaps in plea outcomes, with Black defendants less likely to receive charge reductions).

10. Mayson, *supra* note 8, at 2229–30; Megan T. Stevenson & Jennifer L. Doleac, *Algorithmic Risk Assessment in the Hands of Humans* 1 (Inst. of Lab. Econ., Working Paper No. 12853, 2019); Peter K. Yu, *Artificial Intelligence, the Law-Machine Interface, and Fair Use Automation*, 72 ALA. L. REV. 187, 202–03 (2020); *Fragile Algorithms and Fallible Decision-Makers*, *supra* note 5, at 81–82; Leah Wisser, Note, *Pandora's Algorithmic Black Box: The Challenges of Using Algorithmic Risk Assessments in Sentencing*, 56 AM. CRIM. L. REV. 1811, 1819 (2019).

11. MODEL PENAL CODE: SENT'G § 6B.06(2)(a), cmt. c (AM. L. INST., Proposed Final Draft 2017) (disapproving the consideration of race or ethnicity in risk tools and noting constitutional concerns); *see, e.g.*, Bernard E. Harcourt, *Risk as a Proxy for Race: The Dangers of Risk Assessment*, 27 FED. SENT'G REP. 237, 238 (2015) (citing concerns with using race proxies in risk scores).

12. *See generally* Bo Cowgill & Megan T. Stevenson, *Algorithmic Social Engineering*, 110 AEA PAPERS & PROCS. 96 (2020) (describing attempts to reduce bias through tool design).

algorithm for failing to include race. Designing algorithms that will be embraced by users requires navigating these challenging questions.

The three lessons revealed by our study—hesitance of experts, incomplete/inaccurate data, and fairness concerns—show that there are substantial barriers to uptake when it comes to adopting predictive algorithms. We think these lessons may be applicable across a variety of domains. The features of our setting are not unique. In fact, they may be especially common in areas where gains from AI could be most profound: resource-constrained settings in which experts need to make high-stakes evaluations rapidly. Experts in many domains are likely to be hesitant about adopting new methods. Experts tend to be busy and set a high bar for what counts as helpful enough to warrant a change in their workflow. They also may be reluctant to hand over the reins due to pride or career interests. Deferring to a predictive tool could challenge their domain of expertise.

As for data, criminal justice is not the only high-stakes, resource-constrained domain in which data is limited or inaccurate. By virtue of being a high-stakes setting, the data used may have privacy implications that warrant legal and institutional protections. Beyond criminal justice, medicine and healthcare provides another example of a high-stakes setting where data can be hard to access.[13] Without access to detailed, reliable data, AI and prediction tools will be less accurate. Resource-constrained settings also frequently suffer from data inaccuracies, in part because data can be time intensive and expensive to input and maintain. Data on homeless populations, for instance, requires shelter employees or caseworkers to take time away from direct services to input data.[14] Inaccuracies are expected.

Finally, we expect concerns about fairness and ethics to influence adoption in many settings. Fairness may mean different things in different contexts and to different people. But as long as fairness remains important to end users, they will be hesitant to adopt a tool that does not fully address their concerns.

Learning to extract knowledge from the AI revolution will require learning how to take advantage of it in domains in which it excels, while relying on alternative

---

13. *See, e.g.*, Mohammad Ali Saberi, Hamid Mcheick & Mehdi Adda, *From Data Silos to Health Records Without Borders: A Systematic Survey on Patient-Centered Data Interoperability*, 16 INFO. 106, 106 (2025), https://doi.org/10.3390/info16020106 [https://perma.cc/BB9B-AZ95] (describing "fragmented health data silos across multiple healthcare providers and countries" that slow access to complete patient records); Nathan Denton, Monique Molloy, Samantha Charleston, Craig Lipset, Jonathan Hirsch, Andrew E. Mulberg, Paul Howard & Eric D. Marsh, *Data Silos Are Undermining Drug Development and Failing Rare Disease Patients*, 16 ORPHANET J. RARE DISEASES 161, 161 (2021), https://doi.org/10.1186/s13023-021-01806-4 [https://perma.cc/TXN2-SB76] (noting that balkanized biomedical databases "will inevitably fragment the expanding knowledge base and undermine . . . efforts to develop much-needed treatments").

14. Teale W. Masrani, Helen Ai He & Geoffrey Messier, *The Human Behind the Data: Reflections from an Ongoing Co-Design and Deployment of a Data-Navigation Interface for Front-Line Emergency Housing Shelter Staff*, *in* CHI EA '23: EXTENDED ABSTRACTS OF THE 2023 CHI CONF. ON HUM. FACTORS IN COMPUTING SYS. (Albrecht Schmidt et al. eds., 2023), https://doi.org/10.1145/3544549.3585694 [https://perma.cc/A96T-TGDH] (observing that shelter staff manage cognitively and emotionally taxing work and find data-navigation tools burdensome).

sources of information where it does not. The "failure" of our experiment offers lessons which we hope will be valuable as the world learns to take advantage of the AI revolution as best it can.

The rest of the article proceeds as follows. Part I provides background on predictive algorithms in criminal justice and their potential for public defenders more specifically. Part II describes the setting and design of the experiment as well as the development of the algorithm and user interface. Part III reports the results of the experiment, the quiz comparing attorney and algorithm predictive ability, and our interviews. Part IV discusses the implications of the empirical results and the lessons they hold for the use of algorithms and AI by experts, and the final part concludes.

I. ALGORITHMIC & AI PREDICTION FOR CRIMINAL JUSTICE EXPERTS

This part provides a brief overview of how algorithmic predictions are used in law and criminal justice, then discusses their potential for use in public defense.

A. Algorithmic Predictions in Law and Criminal Justice

The ability to make accurate predictions about potential legal outcomes is a critical lawyering skill.[15] To advise a client effectively, a lawyer must have a reasoned sense of the potential risks and benefits of various strategies, which requires predictions of the outcome of these strategies. Lawyers draw on their knowledge of the law, prior experiences, and advice from others to choose between a variety of case strategies based on their respective risks. Prediction algorithms have the potential to improve information about likely case outcomes, thereby reducing risk and uncertainty. Lawyers in many domains have already adopted such tools.[16]

In the criminal justice space, however, predictive algorithms have primarily been used to help judges, police, and administrators make decisions—not to help lawyers advise clients. Predictive tools are used by law enforcement to forecast crime patterns and inform their decisions about operations, staffing, and response priorities.[17] Law enforcement also uses sophisticated technologies to surveil

15. Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 102 (2014).

16. Ashley Hallene & Jeffrey M. Allen, *Using AI for Predictive Analytics in Litigation*, A.B.A. (Oct. 16, 2024), https://www.americanbar.org/groups/senior_lawyers/resources/voice-of-experience/2024-october/using-ai-for-predictive-analytics-in-litigation/ [https://perma.cc/769T-EYCK].

17. Lyria Bennett Moses & Janet Chan, *Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability*, 28 POLICING & SOC'Y 806, 807–08 (2018); Zach Friend, *Predictive Policing: Using Technology to Reduce Crime*, FBI (Apr. 9, 2013), https://leb.fbi.gov/articles/featured-articles/predictive-policing-using-technology-to-reduce-crime [https://perma.cc/7234-HMAS]; JOEL HUNT, NAT'L INST. OF JUST., FROM CRIME MAPPING TO CRIME FORECASTING: THE EVOLUTION OF PLACE-BASED POLICING 4 (2019). For example, Place-Based Algorithmic Patrol Management (PAPM), also referred to as "place-based predictive policing," enables police to use previous crime data to statistically predict "geospatial areas with an increased probability of criminal activity" to aid their allocation of resources for "intervention and prevention strategies and tactics." DUNCAN PURVES & RYAN JENKINS, A MACHINE LEARNING EVALUATION FRAMEWORK FOR PLACE-BASED ALGORITHMIC PATROL MANAGEMENT 1–2 (2023). *See also* Chaz Arnett, *Black Lives Monitored*, 69 UCLA L.

suspects.[18] Many departments, for example, have access to facial recognition software like Clearview AI, a searchable tool that draws from "a database of more than 3 billion photos collected from social media networks and other public websites" to make identifications.[19] Some tools even predict the risk levels of persons at certain addresses—categorizing them as "green, yellow, or red"—which police may use to prepare for a situation upon arrival.[20]

Algorithmic risk assessment tools are used in almost every state for decisions regarding pretrial detention, sentencing, prison management, or parole decisions.[21] These risk assessment tools predict the likelihood of new criminal justice contact (e.g. arrest) or other types of misconduct on the basis of demographics, criminal record, offense characteristics, socioeconomic factors, and sometimes even attitudes and personality measures.[22] Judges then use these predictions to determine whether or not the person will be detained pretrial, receive a carceral sentence, placed in a high security prison or released on parole. The stakes of these decisions are high, oftentimes months or years of incarceration.

There has been a great deal of discussion in the academic literature in recent years about the implications of the rise of AI and algorithms in criminal justice, ranging from generally optimistic to highly critical.

On the positive side, researchers have praised AI tools in law enforcement as a way to (1) improve the real-time decision-making of police officers and (2) provide data to audit past officer actions.[23] Because many of these programs are automated, they may allow police departments to better direct their limited resources.[24]

Rev. 1384, 1395–1400 (2023) (describing advancements in technology and law enforcement partnership with private companies to enable surveillance).

18. Ric Simmons, *Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System*, 52 U.C. DAVIS L. REV. 1067, 1069 (2018); Arnett, *supra* note 17, at 1395–1400; SARAH BRAYNE, PREDICT AND SURVEIL: DATA, DISCRETION, AND THE FUTURE OF POLICING 57–58 (Oxford Univ. Press 2021).

19. Kashmir Hill, *The Facial-Recognition App Clearview Sees a Spike in Use After Capitol Attack*, N.Y. TIMES (Jan. 9, 2021), https://www.nytimes.com/2021/01/09/technology/facial-recognition-clearview-capitol.html [https://perma.cc/X6MH-B5L2] (describing the use of Clearview to quickly identify rioters on January 6th). *See also* Arnett, *supra* note 17, 1397–98 (describing use of facial recognition software to identify protesters).

20. Justin Jouvenal, *The New Way Police are Surveilling You: Calculating Your Threat 'Score'*, WASH. POST (Jan. 10, 2016), https://www.washingtonpost.com/local/public-safety/the-new-way-police-are-surveilling-you-calculating-your-threat-score/2016/01/10/e42bccac-8e15-11e5-baf4-bdf37355da0c_story.html [https://perma.cc/N698-M75M].

21. *AI in the Criminal Justice System*, ELEC. PRIV. INFO. CTR, https://epic.org/algorithmic-transparency/crim-justice/ [https://perma.cc/796M-BF9M] (last visited Oct. 24, 2025). *See* Simmons, *supra* note 18, at 1072; Robin A. Smith, *Opening the Lid on Criminal Sentencing Software*, DUKE TODAY (July 19, 2017), https://today.duke.edu/2017/07/opening-lid-criminal-sentencing-software [https://perma.cc/64PU-G6GT] (regarding courts' use of COMPAS).

22. Shaila Dewan, *Judges Replacing Conjecture With Formula for Bail*, N.Y. TIMES (June 26, 2015), https://www.nytimes.com/2015/06/27/us/turning-the-granting-of-bail-into-a-science.html [https://perma.cc/5MU7-NTA7].

23. Sharad Goel, Justin M. Rao & Ravi Shroff, *Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy*, 10 ANNALS OF APPLIED STAT. 365, 367, 383, 388 (2016) (noting the flaws in police actual stop-and-frisk when compared with the data, but recommending statistical analysis as a means to identify situations where stop-and-frisk remains appropriate).

24. Simmons, *supra* note 18, at 1069; Meghan J. Ryan, *Understanding Criminal Justice Innovations*, 5 J. L. & INNOVATION 38, 46 (2023).

Results from a field study in Los Angeles indicated that predictive policing tools were better than skilled police analysts at predicting patterns of crime, which led to a slight decrease in crime volume.[25] Facial recognition algorithms, trained on datasets of millions of photos, have the potential to reduce the harms and biases associated with fallible eyewitness testimony, although benefits cannot be ascertained with certainty when the algorithms are not transparent.[26]

Policy simulations have suggested the potential for large gains in adopting pretrial risk assessment. A study found that an algorithm created by economists and other scientists would have performed significantly better than judges at detecting defendant risk when making pretrial release decisions: the algorithm would have "reduce[d] crimes committed by released defendants by up to 25% without having to jail any additional people," and, "without increasing the crime rate at all" would have "jail[ed] up to 42% fewer people."[27]

Algorithms have been criticized primarily for their risk of perpetuating bias due to biased data inputs or general shortcomings with the technology.[28] The use of predictive policing is criticized for entrenching racial bias and creating self-perpetuating cycles of predictions begetting over-policing.[29] Likewise, the widely-used tool for predicting recidivism risk on probation or parole, COMPAS, is only accurate sixty to seventy percent of the time and was more likely to incorrectly flag Black defendants as high risk than White defendants.[30] A study found that "[B]lack defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified [by COMPAS] as higher risk compared to their White counterparts."[31] Reliance on past events—a necessary aspect of predictive algorithms in criminal justice—risks entrenching any biases that tainted prior arrests, adjudications, risk assessments, and sentences.[32]

There are also additional concerns over privacy, oversight, and errors. Observers argue for more public oversight over privately developed, "black box" algorithms.[33]

---

25. Moses & Chan, *supra* note 17, at 816.

26. Brandon L. Garrett & Cynthia Rudin, *The Right to a Glass Box: Rethinking the Use of Artificial Intelligence in Criminal Justice*, 109 CORN. L. REV. 561, 581–83 (2024).

27. Jon Kleinberg, Jens Ludwig & Sendhil Mullainathan, *A Guide to Solving Social Problems with Machine Learning*, HARV. BUS. REV. (Dec. 8, 2016), https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning [https://perma.cc/4TLK-2BU5]; Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig & Sendhil Mullainathan, *Human Decisions and Machine Predictions*, 133 Q.J. ECON. 237, 237 (2018) [hereinafter *Human Decisions and Machine Predictions*].

28. Yu, *supra* note 10, at 202–03; *Fragile Algorithms and Fallible Decision-Makers*, *supra* note 5, at 81–82; Mayson, *supra* note 8, at 2263–67; Wisser, *supra* note 10, at 1819.

29. Moses & Chan, *supra* note 17, at 810.

30. Jeff Larson, Surya Mattu, Lauren Kirchner & Julia Angwin, *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm [https://perma.cc/U8WV-TLXG]; Smith, *supra* note 21.

31. Larson et al., *supra* note 30.

32. Mayson, *supra* note 8, at 2218, 2263–67; Wisser, *supra* note 10, at 1819.

33. Andrea Nishi, Note, *Privatizing Sentencing: A Delegation Framework for Recidivism Risk Assessment*, 119 COLUM. L. REV. 1671, 1685, 1705–09 (2019) (urging for more government oversight of risk assessment

Algorithmic surveillance and identification techniques raise Fourth[34] and Sixth Amendments concerns.[35] Concerns about the reliability of these technologies also abound, both in terms of evidentiary admissibility as well as basic questions of justice.[36]

B. The Potential of Predictive Algorithms in Public Defense

To date, algorithmic prediction tools in criminal justice have primarily been designed for use by judges, parole boards, prisons, and police.[37] Ours is, to our knowledge, the first instance of an algorithmic prediction tool designed for public defenders.

We focus on public defense for several reasons. First, public defenders in the U.S. are often incredibly resource constrained.[38] Tasked with providing legal defense to indigent defendants, individual PDs often handle hundreds of legal cases per year.[39] Given this caseload, they may have only minutes to devote to each case.[40] Many public defender offices struggle to retain experienced attorneys

---

tools and noting the current "overly broad and undefined delegation of power to private actors" creating risk assessment tools used for sentencing).

34. *See* Arnett, *supra* note 17, 1405–07 (discussing the barriers faced by BLM protestors to raise Fourth Amendment challenges against aerial surveillance and facial recognition, that Fourth Amendment jurisprudence deems expectations of privacy to be reduced when in public places, and that private companies developing surveillance technologies may not be captured as state actors); Maneka Sinha, *The Automated Fourth Amendment*, 73 EMORY L.J. 589, 620, 630 (2024) (discussing the "doctrinal gap" between Fourth Amendment jurisprudence and the "pace [of] technology-driven policing," namely that courts are not equipped under current doctrine to consistently evaluate the reliability of police technology as a basis for reasonable suspicion or probable cause).

35. Andrea Roth, *Machine Testimony*, 126 YALE L.J. 1972, 2039–2051 (2017) (discussing a criminal defendant's right to question the credibility of machines used against him under the Sixth Amendment's Confrontation Clause and concluding that, at least sometimes, a defendant's right of confrontation is triggered).

36. Valena Beety, *Considering "Machine Testimony": The Impact of Facial Recognition Software on Eyewitness Identifications*, 60 DUQ. L. REV. 271, 275, 277–79 (2022) (explaining that facial recognition software "has fundamental accuracy problems, undermining its reliability" and its use could negatively impact witness identifications which are already prone to inaccuracy).

37. Doaa Abu Elyounes, *Bail or Jail? Judicial Versus Algorithmic Decision-Making in the Pretrial System*, 21 COLUM. SCI. & TECH. L. REV. 376, 408–11, 416 (2020); Sharad Goel, Justin M. Rao & Ravi Shroff, *Personalized Risk Assessments in the Criminal Justice System*, 106 AM. ECON. REV. 119, 119 (2016) (parole boards); *AI in the Criminal Justice System*, ELEC. PRIV. INFO. CTR, https://epic.org/algorithmic-transparency/crim-justice/ [https://perma.cc/WPY7-URVC] (last visited Oct. 24, 2025) (prison management); Andrew Guthrie Ferguson, *Policing Predictive Policing*, 94 WASH. U. L. REV. 1109, 1117 (2017) (police).

38. *See, e.g.*, PAUL HEATON, GIDEON'S PROMISE VERSUS GIDEON'S REALITY: RESOURCE SHORTFALLS IN PENNSYLVANIA PUBLIC DEFENSE 13–18 (2024) (surveying resource shortfalls across public defender offices in Pennsylvania); Richard A. Oppel Jr. & Jugal K. Patel, *One Lawyer, 194 Felony Cases, and No Time*, N.Y. TIMES (Jan. 31, 2019), https://www.nytimes.com/interactive/2019/01/31/us/public-defender-case-loads.html [https://perma.cc/H6WU-V4FY] (describing how public defenders in many states face excessive caseloads).

39. *See, e.g.*, Oppel & Patel, *supra* note 38.

40. Stephen B. Bright & Sia M. Sanneh, *Fifty Years of Defiance and Resistance After* Gideon v. Wainwright, 122 YALE L.J. 2150, 2164 (2013) ("Many poor people do not see a lawyer until moments before the court proceeding in which their cases are resolved. They have a few minutes of conversation with harried lawyers. . . .").

due to the difficult working conditions and low pay[41] (although this is less true in Clark County).[42] To a certain extent, all of these factors are true of courts, prisons, and other criminal justice bodies. But the problems tend to be more severe in public defense, which remains at the bottom of the totem pole in terms of pay and other benefits.[43]

Introducing new technology has the potential to lessen some of these burdens by increasing the efficiency of an otherwise constrained defense system. When a busy defense attorney may only have five minutes to consult with her client before a critical hearing,[44] a predictive tool that can reduce her workload by even minutes has the potential to meaningfully impact the quality of her representation.[45]

Second, public defenders are unlike judges or criminal justice administrators in that they must favor the interests of their client and their client only. At least in theory, judges, prosecutors and other criminal justice actors are tasked with "seeking justice," which does not necessarily mean seeking long sentences. In contrast, public defenders are charged with zealously representing their clients, which usually means seeking the shortest sentence possible.[46] This makes it much more straightforward to evaluate whether access to algorithmic predictions improves workplace performance. If it results in shorter sentences, then public defenders are being more effective at accomplishing their goals.

In contrast, it would be much harder to evaluate "success" of a sentence prediction tool placed in the hands of prosecutors or judges. Justice is not clearly defined and therefore cannot be easily measured. The fact that there aren't multiple competing concerns also means that issues of algorithmic fairness are substantially different and arguably less severe in this context, which we discuss in Part IV.C.

A third reason why we chose a public defender's office as the site of our study is that public defense is enormously impactful in this country. Public defenders represent millions of criminal defendants annually.[47] Nationally, about 80% of people

41. *See* Jonah E. Bromwich, *Hundreds Have Left N.Y. Public Defender Offices Over Low Pay*, N.Y. TIMES (June 9, 2022), https://www.nytimes.com/2022/06/09/nyregion/nyc-public-defenders-pay.html [https://perma.cc/N93C-YUKL].

42. David S. Abrams & Albert H. Yoon, *The Luck of the Draw: Using Random Case Assignment to Investigate Attorney Ability*, 74 U. CHI. L. REV. 1145, 1145 (2007).

43. Samantha Jaffe, Note, *"It's Not You, It's Your Caseload": Using Cronic to Solve Indigent Defense Underfunding*, 116 MICH. L. REV. 1465, 1465–66 (2018); Ellen B. Primus, *Defense Counsel and Public Defense*, *in* 3 REFORMING CRIM. JUST. 121 (Erik Luna ed., 2017) (describing chronic underfunding of public defense systems, including low salaries, high caseloads, and limited investigative resources relative to prosecutors).

44. Oppel & Patel, *supra* note 38.

45. *See, e.g.*, Aviv Caspi, Overworking Public Defenders 2–4, 8 (Mar. 27, 2023,) (unpublished working paper); Amy Mahler, The Impact of Working Conditions on Productivity: Evidence from the US Public Defense System 25–26 (Jun. 5, 2024) (unpublished working paper).

46. *Performance Guidelines for Criminal Defense Representation (Black Letter)*, NAT'L LEGAL AID & DEF. ASS'N. § 1.1(a), https://www.nlada.org/defender-standards/performance-guidelines/black-letter [https://perma.cc/6RT7-ZYEL] (last visited Nov. 17, 2025) ("The paramount obligation of criminal defense counsel is to provide zealous and quality representation to their clients at all stages of the criminal process.").

47. DONALD J. FAROLE JR. & LYNN LANGTON, BUREAU OF JUST. STATS., COUNTY-BASED AND LOCAL PUBLIC DEFENDER OFFICES, 2007 1 (2010).

charged with a felony in state court were represented by a public defender.[48] Because the vast majority of criminal cases are resolved through plea bargains,[49] giving public defenders access to better information could have a substantial impact on their ability to negotiate in the best interests of their clients. If our tool proved effective and was adopted nationwide, it could have large benefits.

In a 1978 survey, a defense attorney described that learning "what [the] case is worth" guards defense attorneys from being pressured by prosecutors into unfavorable pleas.[50] The author described the ability of experienced defenders to develop a "feel for a case" as essentially "an impressionistic multiple regression model carried in the attorney's head."[51] More experienced defenders may have decades of experience in developing this "feel," but new attorneys have to "learn the hard way" by "get[ting] battered around."[52]

As an author of this paper has shown in a previous study, it is not just the inexperienced defense attorney who takes the battering, but her clients too, in the form of worse case outcomes.[53] Disparities in sentencing outcomes between public defenders may be due to differences in skill, motivation, or personal connections with the judge or prosecutor,[54] but part of these differences in skills could be linked to attorneys' ability to predict the likely outcome of a given case, which could help them negotiate more favorable plea bargains and better allocate their time to cases where they can help the most.[55]

We anticipated that aiding PDs' decision-making with a sentence prediction tool could help attorneys develop a sense of sentencing norms for unfamiliar cases—without needing many years of experience. This could help in bargaining with prosecutors. Judges typically defer to the plea deals reached between prosecutors and defendants when handing down sentencing decisions, so the ability of defense attorneys to negotiate a favorable agreement will directly influence their client's ultimate sentence.[56]

---

48. CAROLINE W. HARLOW, BUREAU OF JUST. STAT., DEFENSE COUNSEL IN CRIMINAL CASES 1 (2000).

49. Approximately 98% of federal offenses and 95% of state offenses are resolved through plea bargaining. Lucian E. Dervan, *Fourteen Principles and a Path Forward for Plea Bargaining Reform*, A.B.A. (Jan. 22, 2024), https://www.americanbar.org/groups/criminal_justice/resources/magazine/2024-winter/fourteen-principles-path-forward-plea-bargaining-reform/ [https://perma.cc/FH4U-YLXJ].

50. MILTON HEUMANN, PLEA BARGAINING: THE EXPERIENCES OF PROSECUTORS, JUDGES, AND DEFENSE ATTORNEYS 76 (Univ. Chi. Press 1978).

51. *Id.*

52. *Id.* at 77.

53. Abrams & Yoon, *supra* note 42, at 1173–76.

54. *Id.* at 1173.

55. Stephanos Bibas, *Incompetent Plea Bargaining and Extrajudicial Reforms*, 126 HARV. L. REV. 150, 164–65 (2012).

56. *See, e.g.*, Anne Metz, John Monahan, Luke Siebert & Brandon Garrett, *Valid or Voodoo: A Qualitative Study of Attorney Attitudes Toward Risk Assessment in Sentencing and Plea Bargaining* 3 (Univ. Va. Pub. L. & Legal Theory, Abstract No. 3552018, 2020).

The law can be an opaque domain,[57] where experience with understanding what *actually* happens is more important than a textbook understanding of what *should* happen.[58] Wealthier defendants can hire attorneys who are not only more effective at a given task, but, given their knowledge of the system, can also make better predictions and thus better decisions.[59] Access to algorithms offers the prospect that one would not be penalized for one's lawyer's inexperience or inaccurate expectations.

II. EXPERIMENTAL DESIGN

To determine whether access to sentencing predictions would impact public defender behavior and the sentences of their clients, we developed a novel machine learning algorithm to predict sentence lengths in felony cases. We tested it through a randomized field experiment in the Clark County Public Defender's Office in 2019-2020. Our goal was to isolate the impact of the predictions on sentences. The expectation was that insights gained from access to the algorithmic sentencing predictions would help public defenders quickly decide what the best response to a plea offer would be, thereby freeing up time and resources for other work. In this Part, we detail the implementation, development, and testing of our novel algorithm aimed at predicting sentencing outcomes in felony cases. We first provide an overview of the experiment, then discuss the development of the algorithm, and then explain our software design choices, which were made to maximize usage.

A. Criminal Justice in Clark County, Nevada

Our study took place in collaboration with the Public Defender's Office of Clark County, Nevada. Prior to moving forward with Clark County, we reached out to over twenty PD offices in large cities and had preliminary discussions with offices in Chicago, Dallas, and San Diego. We chose to proceed with Clark County due to a superior data infrastructure and greater support from the office. One of the authors also had worked with the office previously on a separate project several years prior.[60]

Clark County is the most populous county in Nevada with about 2.3 million residents and is the eleventh most populous county in the United States.[61] It is home to Las Vegas, Nevada's most populous city.[62] The county is similar in size to such counties as Kings County, New York (home to Brooklyn), Dallas County, Texas, and Miami-Dade County, Florida.[63] According to 2024 demographic estimates,

---

57. Stephanos Bibas, *Transparency and Participation in Criminal Procedure*, 81 N.Y.U. L. REV. 911, 924 (2006) (describing how, due to the opacity of the criminal justice system, outsiders rely on unrepresentative and sensationalistic media accounts).

58. HEUMANN, *supra* note 50, at 75.

59. Abrams & Yoon, *supra* note 42, at 1173–74.

60. *Id*. at 1145.

61. US COUNTY POPULATIONS 2025, WORLD POPULATION REV., https://worldpopulationreview.com/us-counties [https://perma.cc/N3DG-N3C3] (last visited Feb. 3, 2025) (showing estimates for 2025).

62. KRISTEN CARNEY, NEV. DEMOGRAPHICS, NEVADA CITIES BY POPULATION, https://www.nevada-demographics.com/cities_by_population [https://perma.cc/R6HJ-2A5K] (last updated Aug. 7, 2025).

63. WORLD POPULATION REV., *supra* note 61.

Clark County is about 37.5% white (non-Hispanic), 32.8% Hispanic or Latino, 14% Black, and 12.2% Asian.[64] Clark County reported 12,100 violent crimes in 2019,[65] putting its crime rate slightly higher than the national average.[66]

As required by the Sixth Amendment,[67] Clark County, Nevada provides indigent defendants with legal counsel at all phases of serious criminal proceedings.[68] Defendants are deemed indigent when they cannot, without substantial hardship, obtain competent, qualified counsel on their own.[69] The Clark County Public Defender's Office (CCPD) is the primary provider of indigent defense in the county[70] and the subject of this study. In 2019, CCPD defenders were appointed to over 26,500 criminal cases—18,921 felony cases, 2,091 gross misdemeanors, and 4,754 misdemeanors, not including juvenile cases.[71] Almost all of the CCPD's cases come from Las Vegas.[72]

The CCPD offers vertical representation, meaning the attorney assigned to a case is responsible for that case throughout its duration.[73] This is in contrast to some other PD offices, which only offer horizontal services where different attorneys handle different stages of a case.[74] In the CCPD, defenders are randomly assigned cases before the judge, prosecutor, or CCPD team chief are aware of the defendant's identity or charges.[75] Newly hired PDs in the office spend six to twelve months exclusively on misdemeanor cases before they are assigned to any felony cases.[76]

Particularly relevant to this study is the process of criminal sentencing in Nevada. Nevada follows an indeterminate sentencing regime in which judges set

---

64. *Quick Facts: Clark County, Nevada*, U.S. CENSUS BUREAU, https://www.census.gov/quickfacts/fact/table/clarkcountynevada/PST045224 [https://perma.cc/YF5X-EJW4] (last visited Feb. 3, 2025) (compared to the national average, there is a higher representation of Hispanic or Latino residents and a lower representation of white non-Hispanic residents in Clark County); *Quick Facts: United States*, U.S. CENSUS BUREAU, https://www.census.gov/quickfacts/ [https://perma.cc/7DM5-YDK9] (last visited Feb. 3, 2025).

65. NEV. DEP'T OF PUB. SAFETY, UNIFORM CRIME REPORTING 2019 REPORT 51 (2025), https://www.rccd.nv.gov/siteassets/content/about/ucr/2019.pdf [https://perma.cc/T8NQ-AVQ5] (reporting 12,100 total violent crimes and a population of 2,270,103 in Clark County in 2019 for a violent crime rate of 533 per 100,000 inhabitants).

66. FED. BUREAU OF INVESTIGATION, CRIME IN THE UNITED STATES, 2019: VIOLENT CRIME (2020), https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/violent-crime.pdf [https://perma.cc/E2KQ-BYLD] (an estimated 1,203,808 violent crimes in the U.S. in 2019, or 366.7 violent crimes per 100,000 inhabitants).

67. Gideon v. Wainwright, 372 U.S. 335, 339–40 (1963).

68. NEV. JUDICIARY, MODEL PLAN FOR THE PROVISION OF APPOINTED COUNSEL FOR URBAN COURTS IN NEVADA, 1–2 (May 22, 2008), https://nvcourts.gov/__data/assets/pdf_file/0028/25984/model_plan_for_the_provision_of_appointed_counsel_for_urban_courts_in_nevada.pdf [https://perma.cc/MJG2-TB7M].

69. *See In re Review of Issues Concerning Presentation of Indigent Defendants in Criminal and Juvenile Delinquency Cases*, ADKT No. 411 (2008) (motion for an order approving Clark County Administrative Plan for Appointment of Counsel); *see* NEV. JUDICIARY, *supra* note 68, at 3.

70. STATE OF NEV. DEP'T. OF INDIGENT DEF. SERV., ANNUAL REPORT FISCAL YEAR 2020 18 (2020).

71. *See id.*

72. Abrams & Yoon, *supra* note 42, at 1149.

73. *Id.* at 1164.

74. *See* David S. Abrams & Priyanka Goonetilleke, *Assembly-Line Public Defense*, 100 NYU L. REV. 2025 (forthcoming) (manuscript at 4–5) (providing an empirical investigation of the costs and benefits of horizontal and vertical systems); Abrams & Yoon, *supra* note 42, at 1164.

75. Abrams & Yoon, *supra* note 42, at 1164.

76. *Id.* at 1161.

both a minimum and maximum term within the guidelines set by the legislature.[77] The prisoner must serve at least the minimum sentence, but the actual time served by prisoners is determined by parole boards according to the rehabilitative record of the prisoner.[78] According to our interviews, most people are released not long after they've served the minimum sentence.

B. Algorithm Development

In order to develop and train our prediction algorithm, we obtained eleven years of historical data including all felony cases handled by the CCPD from 2007-2017. We compiled data on cases, clients and their criminal histories, disposition and sentencing, as well as administrative information.[79] This data came from the PD case management system. A major consideration as we developed the algorithm was what variables to include, and, in particular, whether to include the defendant's race in the algorithm training procedure. We defer discussion of this aspect of the algorithm development to Part IV.C.

We primarily considered two leading machine learning approaches: random forests and stochastic gradient boosting. Random forest models are assemblies of multiple decision trees, where each tree splits the data based on ranges of values the variables can take on.[80] For example, a tree based on criminal history may predict a sentence five years longer for those with criminal histories higher than average. By assembling many trees, a random forest model makes predictions based on past cases that look most similar to the current case.[81] In developing a random forest model, it is necessary to choose various "tuning parameters," including the

77. ALISON LAWRENCE, NAT'L CONF. OF STATE LEGISLATURES, MAKING SENSE OF SENTENCING: STATE SYSTEMS AND POLICIES 4 (2015), https://documents.ncsl.org/wwwncsl/Criminal-Justice/sentencing.pdf [https://perma.cc/W7G8-6MK3].

78. *Id.* at 9; PAULA M. DITTON & DORIS JAMES WILSON, BUREAU OF JUST. STATS., TRUTH IN SENTENCING IN STATE PRISONS (1999), https://bjs.ojp.gov/redirect-legacy/content/pub/pdf/tssp.pdf [https://perma.cc/Z2W8-RU2C] (noting truth in sentencing laws emerged in response to calls for "tough on crime" policies in many states operating under indeterminate sentencing regimes in the 1980s).

79. The following variables were used to train the algorithm and make sentencing predictions: Defense ID number; arraignment date, lead offense, lead statute code, second offense, second statute code, judge, number of charges, max severity, max class, prior cases, open cases, ongoing suspended, sentence, in custody, number of prior convictions, number of prior violent cases, number of prior felony cases, number of open felony cases, number of prior felony convictions, number of days spent incarcerated for prior cases, number of prior suspended sentences, number of prior probation revocations, all charges, all statute codes, the severity level of the case (felony, gross misdemeanor, misdemeanor), prior bench warrants, intoxicants, modifiers, and enhancers (which were summarized into special condition #1 and #2 on the website), Uniform Crime Report category of crime, month the case was received, Justice Court Judge, year the case was opened, age, zip code. Note: max severity is the maximum severity classification for an offense while max class is the maximum type of felony classification. Ongoing suspended refers to whether the defendant had a current suspended sentence; sentence is the length of the sentence in the training data. Justice Court Judge is the identity of the judge that handled the preliminary hearing.

80. Sendhil Mullainathan & Jann Spiess, *Machine Learning: An Applied Econometric Approach*, 31 J. ECON. PERSP. 87, 93–94 (2017) (explaining distinctions between OLS and machine learning and advantages of machine learning).

81. *Id.* at 102 (explaining distinctions between OLS and machine learning and advantages of machine learning).

minimum number of observations in the terminal nodes (the ends of branches); the number of trees; and the number of candidate variables for splitting using the trees.[82]

Gradient boosting is similar to random forests in that it also takes a random subset of the variables and a random subset of the data initially.[83] While predictions from the initial iteration of a model may be poor, model parameters are updated such that more weight is given to those models that have the greatest gain in the direction of the correct prediction.[84] The tuning parameters for gradient boosting include the depth of each tree, the number of trees, and the weight given to subsequent trees (the learning rate, or shrinkage).[85]

Our model was trained to predict the sentence minimum among those who were incarcerated. We focus on the sentence minimum because the public defenders we spoke to while developing the algorithm reported that the amount of time actually served is generally close to the minimum of the sentence range. Both classes of model are trained using historical cases where all of the case information as well as the actual sentence length is known. To determine which algorithm performs best, the models were trained using 80% of the historical data (training set). The remaining 20% (hold-out set) of the data was then used to evaluate the predictions.

The models were evaluated by using several measures of the distance of the predictions from the real sentences, including mean absolute error, mean squared error, and the fraction of predictions that were within a specified time interval of the actual sentence.[86] The "error" is the predicted sentence minus the actual sentence, and therefore includes both positive and negative numbers. The absolute error is the error expressed as a positive number (e.g. if the error was -5 months, the absolute error would be 5 months). The mean absolute error is therefore the average of all of the errors expressed as positive numbers. The mean squared error is the average of all the squared errors. The random forest algorithm outperformed gradient boosting, although the differences there were fairly small. We chose the random forest algorithm to generate sentencing predictions to be given to the public defenders.

In recent years, there has been some debate about whether the more modern machine learning methods such as random forests and gradient boosting provide a substantial improvement in accuracy over more traditional prediction methods, such as ordinary least squares (OLS).[87] OLS produces the multidimensional line

82. Jaehyun Yoon, *Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach*, 57 COMPUTATIONAL ECON. 247, 253 (2021).

83. *Id.* at 249–51.

84. *Id.*

85. *Id.* at 253.

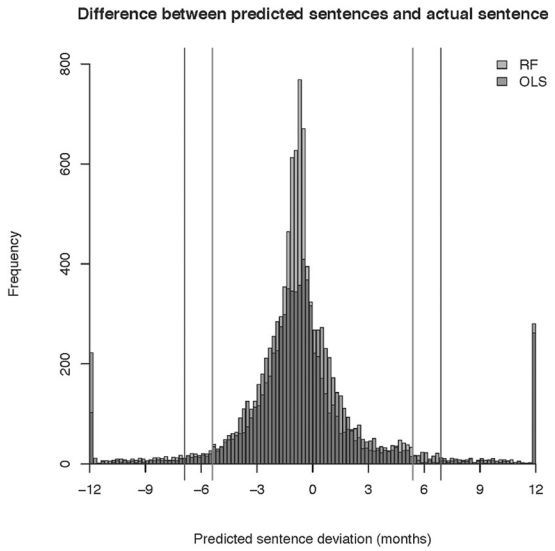
86. Mean absolute error is computed by calculating the absolute value of the difference between the actual and predicted sentences for each case and then taking the average of all the absolute values; mean square error is similar, but rather than computing the absolute value, it computes the square of the difference between the actual and predicted sentence for each case and then takes the average; the last measure computes what fraction of predictions were within 2 months of the actual sentence, for example.

87. Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel & Daniel G. Goldstein, *Simple Rules to Guide Expert Classifications*, 183 J. ROYAL STAT. SOC'Y SERIES A 771 (2020).

that most closely fits the data and predictions all lie along that line.[88] OLS regressions have long been the workhorse econometric approach.[89] However, OLS imposes the assumption that sentences respond linearly to the inputs, which can be restrictive.

In Figure 1 we provide a comparison of the performance of our selected model, random forests (light grey) against OLS (dark grey). For each model we compute the prediction error, defined as the deviation between the predicted and actual sentence for each case in the test set, and then plot all of these errors (measured in months) as a histogram. The random forest errors form a tighter distribution, clustered substantially closer to zero than the errors from OLS. This can be seen visually with the higher peak in the light grey bars around zero and greater prevalence of the dark grey bars in the tails. We also plot the 5th and 95th percentiles of each distribution as solid vertical lines. The narrower spread for random forests suggests less frequent large errors compared to OLS.[90]

Figure 1. Distribution of Predicted and Actual Sentence Deviations for Random Forest and OLS



88. WILLIAM H. GREENE, ECONOMETRIC ANALYSIS 396–97 (Rod Banister et al. eds., 5th ed. 2002).

89. *Id.*

90. Ours is not the first use of random forests in a criminal justice application. *See, e.g.*, Melanie-Angela Neuilly, Kristen M. Zgoba, George E. Tita & Stephen S. Lee, *Predicting Recidivism in Homicide Offenders Using Classification Tree Analysis*, 15 HOMICIDE STUD. 154, 156, 164 (2011) (predicting homicide offender recidivism); Richard A. Berk, Brian Kriegler & Jong-Ho Baek, *Forecasting Dangerous Inmate Misconduct: An Application of Ensemble Statistical Procedures*, 22 J. QUANTITATIVE CRIMINOLOGY 131, 138 (2006) (predicting misconduct among incarcerated prisoners); Richard Berk, Lawrence Sherman, Geoffrey Barnes, Ellen Kurtz & Lindsay Ahlman, *Forecasting Murder Within a Population of Probationers and Parolees: A High Stakes Application of Statistical Learning*, 172 J. ROYAL STAT. SOC'Y 191, 191–92 (2009) (predicting murders by criminals on probation or parole); Richard A. Berk, Susan B. Sorenson & Geoffrey Barnes, *Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions*, 13 J. EMPIRICAL LEGAL STUD. 94, 95 (2016) (predicting domestic violence).

C. Interface Design and Testing

For a technological tool to be useful, it must be designed in a way that is legible to users.[91] Therefore, a great deal of effort was expended on the design to convey sentencing predictions, including contracting with the software company Court Innovations[92] to develop an interface that was simple to learn and versatile. Initial discussion with PDs resulted in software that could be accessed through a browser on a computer or mobile device. Sentencing predictions and other case information could also be easily printed in a useful format for presenting to clients. The software evolved over time until it took the form described here.

When attorneys accessed the software, they were initially presented with a list of their open cases as displayed in Figure 2. In addition to the case number, this screen includes features that allow for quick identification: the lead charge statute and description, defendant age and sex, and date of case assignment.

Choosing a case led to a subsequent web page that included two parts: the sentencing prediction at the top (Figure 3) and more detailed information at the bottom (Figure 4). Sentence predictions were displayed as a minimum and maximum range because this is how sentences are given in Clark County. A person who is sentenced to fourteen to forty-eight months will be required to serve a minimum of fourteen months before they are eligible for parole. Their sentence is capped at forty-eight months.

In addition to the predicted sentence, we also provided a measure of the precision of the prediction, conveyed through the "precision meter" which used a color-coded dial to indicate one of four levels: low, moderate-low, moderate, or good.[93] When precision was too low, roughly 5% of the time, the page displayed "No estimates generated for this case" and no sentencing range was provided.

In addition to providing the sentence prediction and confidence, the top of the screen allowed for feedback from attorneys. Through the "Submit Feedback" button, they were asked to enter anything that might be incorrect about the case details, which may have been due to human error in input or to changes not entered.[94] This occurred only rarely.

At the bottom of the page, we provided detailed case characteristics which were updated daily based on data from the public defender. This information, shown in

---

91. The Technology Acceptance Model (TAM) is a theory aimed at understanding what motivates users to accept or reject technology. Fred D. Davis, *Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology*, 13 MIS Q. 319, 320 (1989). Perceived usefulness and perceived ease of use of a technology are fundamental predictors of user uptake. *Id.* at 333. A study applying the TAM model to evaluate police officers' uptake of computer technology found that officers considered ease of use to be most important to them, followed by usefulness of the technology for their duties. Caran A. Colvin & Angeline Goh, *Validation of the Technology Acceptance Model for Police*, 33 J. CRIM. JUST. 89, 91 (2005).

92. Court Innovations is a software company whose founders came from the University of Michigan and provides software solutions to increase access to courts. *See* MATTERHORN BY COURT INNOVATIONS, http://www.courtinnovations.com/ [https://perma.cc/MLF7-VP6E] (last visited Oct. 24, 2025).

93. To produce prediction intervals, we used the empirical distribution of past prediction errors to estimate future prediction errors. This process, called "conformal prediction," is akin to bootstrapping and can be used with any prediction algorithm to generate prediction intervals with any level of certainty. *See* Glenn Shafer & Vladimir Vovk, *A Tutorial on Conformal Prediction*, 9 J. MACH. LEARNING RSCH. 371, 371, 373 (2008).

Figure 2. Case Selection Interface

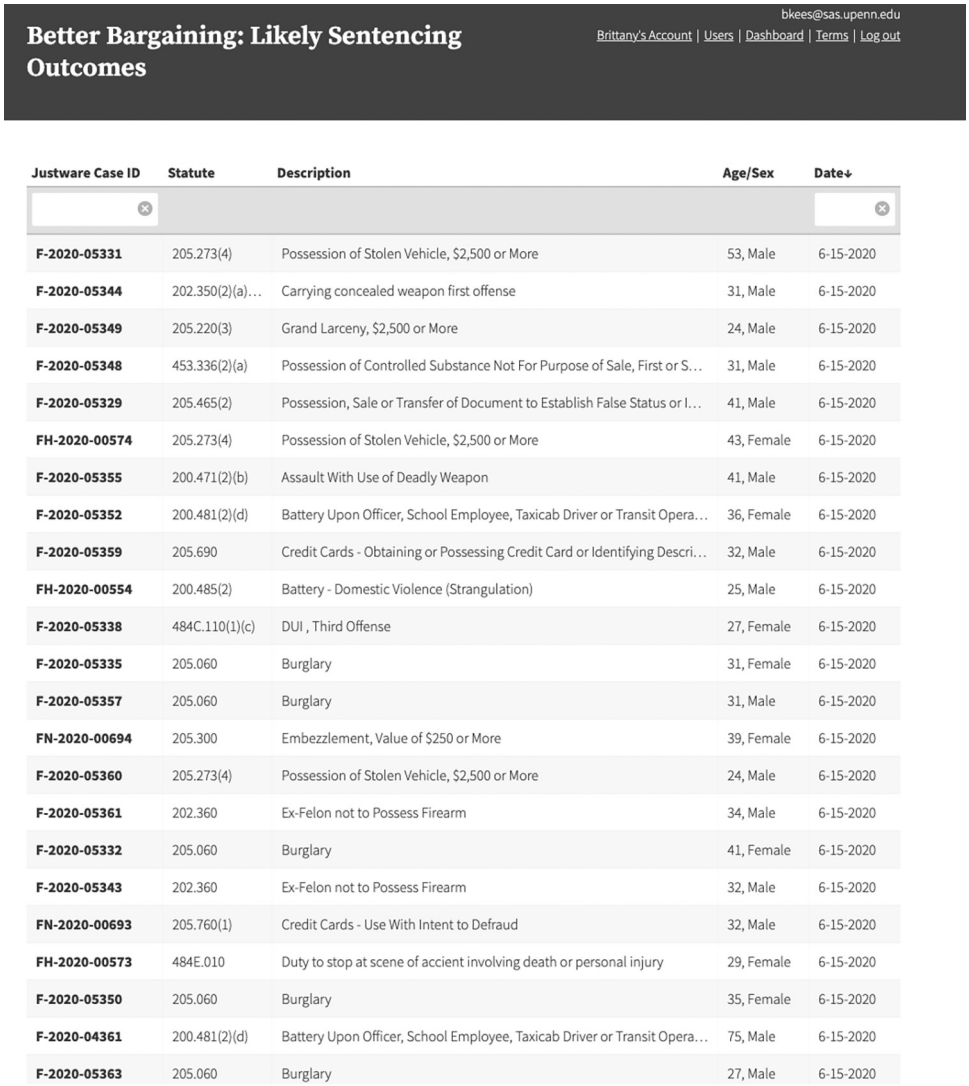
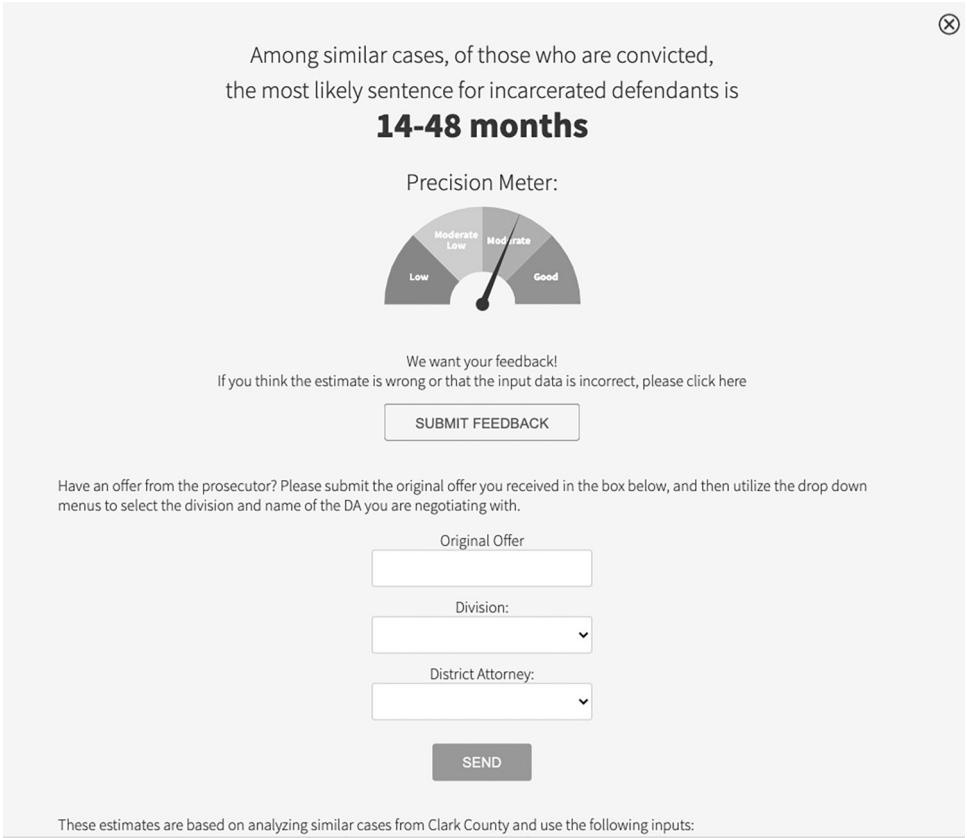


Figure 4, serves several purposes. First, it showed the attorneys what information was used to develop the prediction for that case. This allowed them to observe which information was taken into account and which wasn't, as well as to verify its accuracy. The second reason we provided this information is because PDs

94. PDs were also instructed in initial trainings and frequently reminded not to rely on predictions if any of the data presented was incorrect. We also initially asked for information about plea offers from prosecutors on this screen, although we later discontinued this data collection.

Figure 3. Sentence Prediction



expressed interest in it; there was previously no other platform or document in which they could find all of the variables about a case in one place. Thus, in addition to the sentence predictions, we hoped that this simple display of case information would be useful.

We went through multiple iterations in the process of developing this final product. To maximize usage of the software, we held multiple feedback sessions with small groups of PDs and a multi-week pilot test of the software. This led to a number of revisions to the software design. For instance, we revised the language used on the prediction page because we found it was imprecise and led to misinterpretation.

Most notably, we chose to refer to our generated numbers as *estimates* rather than *predictions*. This language change was made to help underscore the fact that PDs have access to information about each case that the algorithm does not, such

Figure 4. Case Characteristics for Attorney Verification

Case Inputs

Justware Case ID:	F-2020-05348
Arraignment Date:	Jun 15, 2020
Lead Offense:	Possession of Controlled Substance Not For Purpose of Sale, First or Second Offense, Schedule I, II, III, or IV Substance
Lead Statute Code:	453.336(2)(a)
Second Offense:	
Second Statute Code:	
Number of Charges:	1
Max Severity:	Felony
Max Class:	E
Prior Cases:	0
Open Cases:	0
Ongoing Suspended Sentence:	0
In Custody:	No
Special Condition #1:	Methamphetamine intoxicant present
Special Condition #2:	N/A
Number of Prior Convictions:	
Number of Prior Violent Cases:	0
Number of Prior Felony Cases:	0
Number of Open Felony Cases:	0
Number of Prior Felony Convictions:	0
Number of Days Spent Incarcerated for Prior Cases:	0
Number of Prior Suspended Sentences:	0
Number of Prior Probation Revocations:	0

*__Note:__ Criminal history includes only occasions where defendant was represented by a Clark County public defender in the last 10 years.

as the strength of the evidence. In trainings, we emphasized that the algorithm provides an estimate for an average case with similar characteristics, but that the defender must incorporate their own knowledge of factors that are not included in the algorithm and adjust up or down as needed.

D. Experimental Setup

While the CCPD office has over seventy attorneys, quite a few were ineligible to participate in the study because they were part of special units, such as those that focus solely on homicide, sexual assault or immigration cases. Other attorneys were ineligible if they were primarily in a supervisory role or if they had long expected periods of absence from the office, such as maternity leave. Of the forty-four PDs eligible to participate, we used a random number generator to choose twenty-two to have access to our predictions while the other twenty-two operated

as usual. The main analysis includes thirty-nine attorneys, as five had to be dropped after leaving the office or being switched to a special unit which was outside of the domain of the study.

Several steps were taken to maximize attorney usage of the software. Prior to the main experiment, in early 2018, we shared access to the software with a handful of the PDs in order to refine the design and implementation. We then ran a ten-week pilot study between March and May 2018 in order to further refine the study setup and ensure that the data infrastructure was able to produce new sentencing predictions daily. The full-scale experiment began on July 15, 2019 and was intended to run for one year, but was forced to end in March 2020 due to the impact of the COVID-19 pandemic on the courts.

All attorneys in the treatment group were trained on how to use the software by our project manager who returned to the office periodically. Attorneys were shown a PowerPoint presentation that explained how the software worked and then met one-on-one with a team member to ensure that they were able to use it. During the training, we emphasized the fact that the software was available to assist the attorneys in the decision-making process, but that it was ultimately up to them how or whether to make use of it.

In order to maximize take-up, a project manager made many trips to the Clark County Public Defender's Office, when we were designing the prediction tool, during the pilot, and during the full study. In total, the project manager came to Vegas six times during the design and pilot period (March 2018 to March 2019), twice to train PDs who were part of the full study (April and May of 2019), and six times during the run of the experiment (May 2019 to November 2019). In total, she visited thirteen times over eighteen months and was in frequent contact over the phone in between visits.

Our project manager was hired, in part, due to her people skills. We hoped that the public defenders would enjoy chatting with her, which would help facilitate communication and increase uptake. Our impression was that she forged many friendly relationships with the public defenders and that usage would have been substantially lower without this personal touch. We discuss this more below.

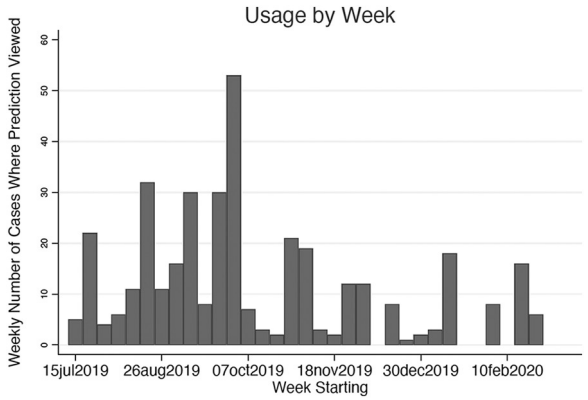
III. EXPERIMENTAL RESULTS

The biggest finding from the field experiment was that, despite extensive effort, usage of the software was low. To better understand this, we conducted a survey, a prediction "quiz," and a number of follow-up interviews. This Part reports what we learned, both from the experiment and these supplementary sources. We first report usage rates overall and how they varied by case and attorney characteristics. Second, we investigate whether perhaps our predictions were actually no better than those of the attorneys themselves, and therefore did not add value. Third, we report results of structured interviews with some of the public defenders in the treatment group on why they did not make more use of the predictions. Finally, we discuss circumstances specific to the experiment that likely impacted usage.

A. Overall Software Usage and Correlates

To quantify software usage, we observed the number of cases where the public defender viewed the sentencing prediction, using metadata generated by the software (Figure 5). After the full experiment began on July 15, 2019, usage of the software rose through the end of the summer and into the fall, when it began to decline, with intermittent spikes. By the time the experiment ended on March 14, 2020, usage had declined a fair amount. Overall, only 10.3% of predicted sentences were viewed by public defenders. Even at peak usage in October 2019, PDs looked at predictions only for a minority of cases.

Figure 5. Weekly Usage of Sentencing Predictions

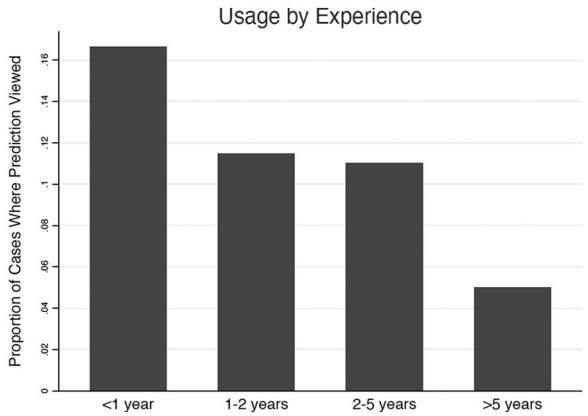


While overall usage was low, there was substantial heterogeneity across attorneys. About 30% of attorneys viewed sentencing predictions for less than 1% of their cases, which includes those viewed while being trained on how to use the software. But there was another group of attorneys (also constituting about 30%) who viewed predictions for about one in four cases. The biggest power user viewed predictions for 44% of cases–still less than half of their caseload.

Usage rates differed significantly by attorney experience, as was expected. Public defenders with less than a year of experience accessed the predictions almost 17% of the time—over three times the rate of colleagues with over five years of experience, who used the predictions only 5% of the time (Figure 6).

The difference in usage rates across the categories of attorney experience is strongly statistically significant. The F-test is a statistical test for whether a variable is equal across different categories. In our setting, the F-test (F-statistic $= 15.63$) shows that there is less than a one in one thousand chance that usage rates would differ so much across the categories if usage was completely random.

Figure 6. Software Usage by Attorney Experience



There is also suggestive evidence that the attorneys were more likely to access the predictions for serious cases. Table 1 shows usage rates by the predicted incarceration probability. This information was generated by an alternative random forest model trained to predict the likelihood of being incarcerated versus receiving a sentence of probation or fines. This is in contrast to our primary prediction level, which predicts the sentence minimum among those who are incarcerated. The model predicting the likelihood of incarceration was not shared with attorneys in the experiment because in pilot testing we determined probabilities were hard for public defenders to interpret. Also, PDs told us that the intensive margin of sentence length was of more interest than the extensive margin of whether the person would be incarcerated at all. Table 1 shows that the rate of accessing the sentencing predictions increases with the likelihood of incarceration, ranging from 8.9% usage for lower than 25% likelihood of incarceration to 13.2% usage rate when the likelihood of incarceration ranges from 75 to 100%. The correlation between probability of incarceration and likelihood of usage was not statistically significant, but it provides suggestive evidence of a relationship.[95]

Another piece of evidence also suggests that usage was higher for more serious cases. Appendix Table 1 shows prediction usage rates broken down by the felony class of the top charge. Class A and B are the most serious felony charges and represent a higher share of cases where predictions were consulted. The predictions were less likely to be consulted for Class E (the least serious charges) and Class NA (charges that were likely misdemeanors). As with the prior results, the correlation was not statistically significant, but this may be due to low statistical power.

95. This can be seen by the fact that the F-statistic in Table 1 has an associated p-value of close to 0.13. This means that there is a 13% probability that usage really did not vary across likelihood of incarceration categories, but we observed this level of difference or greater.

Table 1: Comparison of Usage Rates by Probability of Incarceration

Incarceration Probability	Used Prediction	
	mean	sd
<25	0.089	(0.285)
25-50	0.089	(0.284)
50-75	0.110	(0.314)
75-100	0.132	(0.339)
Total	0.103	(0.303)

F-stat: 1.89 p-value = 0.1298

We also observed that usage was substantially higher on days when our project manager visited the office to meet with public defenders in person. These visits were meant to answer questions and encourage usage of the software. Days with visits had over twice the number of logins and over three times as many predictions viewed, compared with days without visits (Table 2). The personal touch was thus important in terms of jumpstarting use. Unfortunately, this momentum did not last. We discuss implications of this further in Part IV.

Table 2: Impact of Team Member Visit on Software Usage

	Day of Visit					
	(1) Yes		(2) No		(3) Difference	
	mean	sd	mean	sd	b	se
Daily Logins	1.533	1.302	0.718	1.094	0.815*	(0.347)
Daily Predictions	6.467	9.738	1.750	4.221	4.717	(2.537)
Observations	15		156		171	

*p<0.05; **p<0.01; ***p<0.001

Quite a few other characteristics were examined to see if they impacted usage, but no others had much predictive value. This includes the defendant's gender and age, type of crime, whether the defendant was currently incarcerated, and information about the defendant's criminal history. It is possible that some of these results would be statistically significant if there were a larger sample.

It should be noted that while we thought that the predictions would help some PDs, we were also aware prior to the study that the predictions would not be universally useful. Well before beginning the experiment, in a pre-analysis reports we wrote,

> "The information provided by the software may be helpful to attorneys as they consider how to advise clients in the plea bargaining process. For a veteran public defender with a client facing typical charges in front of a well-known judge, the information we provide will likely not be very valuable. In these cases, we wouldn't expect much effect. But in a variety of situations, the information could be extremely valuable: we expect it to be most useful with new attorneys, new judges, unusual clients, or unusual charges."

We discuss this further in Part IV.

B. Algorithmic Predictions were Substantially Better than Attorneys

If PDs outperform our algorithm in predictive ability, this could explain the low uptake of the technology. In that case, public defenders are rationally choosing not to use a tool that has no value to them. In order to test this hypothesis, we created a quiz using data from actual cases and administered it to the public defenders in fall 2020. We provided the public defenders with all of the inputs used in the algorithm and asked them to tell us what they thought the sentence would be. This differs from a real-world setting in which public defenders have access to much more information than the algorithm, but allows us to put both humans and algorithms on a level playing field. As we communicated to the defenders, our algorithm was designed to provide a benchmark of average sentences for a case with given characteristics, and they should adjust their expectations based on private information. The test thus evaluated whether our algorithm provided improved information about the benchmark, which should theoretically lead to improved prediction of sentences. This would support our hypothesis that defenders would better be able to evaluate a case with use of the algorithm then they would without.

Because the quiz was based on real cases, we knew the actual sentences and were able to calculate how far off both the algorithmic and attorney predictions were for each case. We then compared the prediction errors of the PDs to those made of the algorithm.

We found that the algorithmic predictions were substantially more accurate than those made by the public defenders (Table 3). The primary measure used was the mean absolute error. To compute this, we first estimated the prediction error for each case for both algorithm and the attorneys, calculated as the difference between the prediction and the actual sentence. We then took the absolute value (so all differences were made to be positive) and averaged the absolute values for all of the quiz responses.

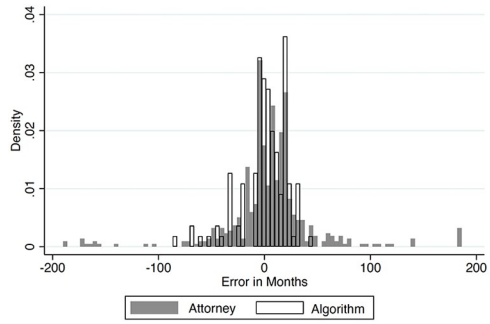
The mean absolute error for attorneys was over 70% higher than that for the algorithm. The disparity was even greater for errors that overestimated the sentence length, which were the vast majority.

Table 3: Attorney vs Algorithm Sentencing Prediction Errors (sentences in months)

	Algorithm			Attorney			Comparison	
	Mean	SD	# obs	Mean	SD	# obs	t-statistic	p-value
Absolute Error	15.28	15.8	545	26.4	31.24	545	-8.47	<0.001
Error|Error>0	13.26	9.71	313	27.66	32.1	482		
Error|Error<0	-19.9	21.36	210	-18.89	22.01	57		

Figure 7 shows the distribution of the prediction errors for both attorneys and the algorithm. While the algorithm errors cluster around relatively small positive and negative values, the attorney errors were frequently quite large. Indeed, all of the negative errors made over 100 months were made by the attorneys, as were all positive errors over 50 months.

Figure 7. Histogram of Prediction Errors for Attorneys and Algorithm



Two additional analyses shed some light on the source of the attorney prediction errors. First, in Table 4, we compare sentencing prediction errors by attorney experience. Attorneys with less than a year of experience as a public defender make absolute prediction errors that are an average of 30% larger than those with more than a year of experience. This result is not statistically significant at the 5% level but is at the 10% level. It would not be surprising if new public defenders were less accurate in their predictions; this could also explain the higher usage rates among less experienced public defenders, as shown in Figure 6.

Table 4: Prediction Performance by Attorney Experience

	Experience <=1 year			Experience >1 year			Comparison	
	Mean	SD	# obs	Mean	SD	# obs	t-stat	p-value
Absolute Error	32.5	42.0	106	25.0	27.9	439	2.2475	0.0802
Error|Error>0	34.6	43.5	95	26.0	28.5	387		
Error|Error<0	-18.2	20.9	9	-19.0	22.4	48		

One might also hypothesize that attorneys with lower performance on the prediction quiz would be more likely to use the algorithmic predictions. We investigate this hypothesis in Table 5, which shows the relationship between quiz prediction performance and use of the sentencing predictions. It reports a comparison of the average absolute error on the prediction quiz for the group of treatment attorneys that were the lightest users of the predictions and those that were the heaviest.[96] As expected, the heavy users have errors that are over 35% higher than those that rarely use the predictions. These results are near the threshold of statistical significance at the 5% level.

96. For this analysis, we limit to people who completed at least 15 questions on the quiz. Among these, we compared the four participants who viewed the least predictions against the four participants who viewed the most.

Table 5: Prediction Error by Algorithm Usage

	Prediction Tool Usage Level					
	(1) Light		(2) Heavy		(3) Difference	
	mean	sd	mean	sd	b	se
Absolute Error	21.773	27.199	29.797	29.985	-8.024	(4.133)
Observations	86		106		192	

*p<0.05; **p<0.01; ***p<0.001

Light users are the 4 users who viewed the least predictions during the experiment (0-6 views) and also answered at least 15 questions of the quiz. Heavy users are the 4 users who viewed the most predictions (32-75 views) during the experiment and also answered at least 15 questions of the quiz.

C. Attorney Feedback on the Software

In order to gain further insight into why usage of the predictions was not very high, we interviewed attorneys in the treatment group in the spring of 2021. We observed a few major sets of concerns in the attorneys' responses, the most frequent of which was about the accuracy of the predictions. We had received this feedback at other junctures as we were developing the algorithm and interface, and we took measures to attempt to address it. But these concerns persisted.

The general worry about accuracy was captured by one attorney who stated, "If I was a little more confident in the numbers and ranges that it's spitting out, then I would have more of an inclination to utilize it." Other attorneys were more explicit that they thought their predictions were more accurate than the algorithm, saying, "Um, well, you tried your best. But—I just don't think that it's smarter than me. I have more information than it does. Granted my information is all over the place, but I still have more of it." More than one attorney was concerned that the case management system used by the office to track data, Justware, was inaccurate and that would necessarily make the predictions so as well: "I think Justware is unreliable. At least it was in the past. It's probably gotten better now—but if Justware is unreliable, then your predictions probably are too."

Attorneys' concerns regarding the information used as the basis for prediction came up in multiple ways. Some attorneys worried about whether the judge was factored into predictions, since judges have a huge impact on cases: "Also, my track has two very different judges . . . so my cases will play out differently with each judge." Although we included judge fixed effects, the judge was generally not yet assigned by the time our predictions were generated, and thus our predictions included the average effect of the two judges that could have been assigned to a case on that track. This was a lower level of specificity than the attorneys wanted.

Others were concerned with how the algorithm made use of information about client race. As explained in Part II.A., we did not explicitly include defendant race

as one of the predictors in our algorithm, though we did include the defendant's home zip code and several other covariates that are highly correlated with race.[97] After much testing, we confirmed that not explicitly including race did not result in differential prediction errors by race or decrease the accuracy of the predictions. Nevertheless, one attorney stated:

> I don't really get why race isn't on here. Everyone knows that plays a role. I'm Black. I know it. Everyone here knows it. Maybe the zip code will work, but I don't know enough about it. And, if you can have zip code on there, and that's supposed to determine race—then why not just have race. Doesn't really make sense to me. But hey—I'm not a scientist. I don't know how this works.

Several PDs, including new ones, thought the sentencing predictions were more useful for newer attorneys: "I just never used it that much because it wasn't helpful for me. Maybe it was for the newer people on staff?" But the newer attorneys had their own issues. One observed, "Honestly, I am [so] overwhelmed by my work [being a new PD] that I just forget to use it." Another said, "It seems like it will be helpful to me as a new defender. But, I don't know enough to know if the predictions are off." A couple of attorneys saw accessing the software as extra work they didn't have time for, were uninterested in additional technological assistance, or just didn't want to change their routines.

Not being able to receive predictions for all cases decreased attorneys' usage. As explained in Part II.C, we made the choice not to provide predictions for cases with high uncertainty—which only happened about five percent of the time—to make sure that we were fairly confident that information was potentially useful and relevant. However, this had a downside: attorneys could look up a case and not see a prediction. This was discouraging to them. Similarly, some commented that the cases they viewed as difficult to predict—the ones for which the algorithm might be most useful—also seemed to have less reliable predictions, as captured by our precision meter. As one attorney said, "It seemed confused on the same cases that I was confused about. So, we have something in common."

In the first month of our experiment we ran into a technical problem that prevented predictions from being shown. Even though this problem was fixed and the software worked for the rest of the experiment, some users were put off: "I didn't like that a lot of my cases were missing in the beginning. That was annoying." In addition, attorneys could temporarily shift to a different unit, which meant they handled cases for which we didn't provide predictions as part of the experimental study. Even if these assignments lasted for only a couple of weeks, it would break the habit of using the software.

On the positive side, several attorneys noted that they could use the predictions to attempt to persuade uncertain clients of their belief in the best course of action,

---

97. *See supra* note 79 (describing variables used to train the algorithm and make sentencing predictions).

although it was unclear the predictions would actually influence attorney beliefs: "If the program told me something that I liked, and the client needed that reinforcement . . . then I might use the software to help explain to the client." However, another was concerned that it could have the opposite effect when presented to clients: "Clients are hard to deal with. Sometimes they will latch on to certain information in the beginning . . . and it's hard to move on from that in the long run. Sometimes it clouds their judgment in the long run, that can hurt their decision-making process."

Another attorney described how the software was being used, but that shifting one's workflow could be difficult: "Whenever I get a file, I run a conflict check, I call my client, and then I run the calculator . . . to help me understand the value . . . and then work from there. I need to get it into my normal practice, and I feel better moving forward. At this point it's really useful to young PDs . . . just a matter of using it." These discussions with the attorneys who used the software were extremely valuable in our assessment of broader lessons from this experiment.

D. Unique Circumstances Decreased Usage

Based on concerns raised during the pilot stage—similar to those discussed above—we designed the experiment with these issues in mind to encourage greater usage. These steps included having the cooperation and endorsement of the Chief Public Defender, having a clear group presentation to introduce the software, as well as one-on-one trainings for all PDs in the treatment group, and having frequent follow-ups with attorneys, both in-person and remotely.

Two unanticipated events had substantial impacts not only on our experiment but on the office more broadly. First, the Chief Public Defender resigned abruptly shortly before we were to begin our experiment. This delayed our start by over half a year. More importantly, it meant that we were working with a new head of the office during our experiment. While he allowed the experiment to continue, he was not nearly as involved as the prior Chief was in sharing his support for the project with the PDs. We were also not allowed to use some incentives that we had planned to try in the hope of increasing usage, such as gift certificates for coffees or meals when attorneys met participation goals.

Second, the worldwide COVID-19 pandemic began as we were about two-thirds of the way through our anticipated one-year experiment. The pandemic impacted the Clark County courts and public defender's office as it did most other aspects of society. The courts slowed down immensely, delayed many cases, and shifted many to be held remotely. Our team could no longer travel to help facilitate communication and increase uptake in person. As a result of this massive change to the nature of the work and the huge decrease in workload, we made the decision to end the experiment earlier than planned. This meant we had fewer months of data and therefore a smaller sample to evaluate usage patterns.

IV. IMPLICATIONS FOR AI IN CRIMINAL JUSTICE AND BEYOND

While we were unable to accomplish our original goal, the field experiment described above yielded insight about the adoption of predictive tools in the legal workplace. These insights are likely to be relevant to other settings, both in the legal space and beyond. Public defenders are legal experts making rapid, high-stakes decisions in order to process a large workload. Many legal workplaces have similar characteristics as other professional spaces: engineering, medicine, etc. We hope that the insights learned here are relevant to other settings in which people try to harness the potential of algorithms and AI to improve workflows. Below, we present three main takeaways of our research.

A. The Bar is High for Algorithmic Performance

In order for experts to adopt a new technology that purports to assist decision-making in the realm of their expertise, they must be convinced that it can deal with nuance.[98] This is true whether it is software to assist with medical diagnoses, financial decision-making, or legal practice.[99]

The biggest concern and reason for lack of regular use of our software was that the attorneys were not convinced that it was accurate *enough*.[100] A frequent piece of feedback was that the public defenders did not think the predictions were precise. Why not? Attorneys mentioned concerns with Justware, the system the PD's office uses to track cases, which we used to feed the algorithm. But they also mentioned concerns that there were factors that wouldn't be captured by our software. For example, one attorney said, "I think your relationship with the judges and the DAs play a big role in how a case goes."[101] Again, we told PDs that the two-judge track was incorporated as well as that the software provided average sentences, and they could adjust accordingly. However, because the algorithm did not capture all relevant dimensions of sentencing, many PDs rejected its value altogether.

In other words, the real world is complicated, and there is currently skepticism that software can capture the subtlety that humans can.[102] Anyone who has used an

98. *See, e.g.*, Metz et al., *supra* note 56; Jennifer M. Logg, Julia A. Minson & Don A. Moore, *Algorithm Appreciation: People Prefer Algorithmic to Human Judgment*, 151 ORG. BEHAV. & HUM. DECISION PROCESSES 90, 99 (2019) (noting that algorithmic advice tends to fall on deaf ears when it is pitted against the judgement of experts).

99. Nikhil Agarwhal, Alex Moehring, Pranav Rajpurkar & Tobias Salz, *Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology* 27–33 (Nat'l Bureau of Econ. Rsch., Working Paper No. 31422, 2023) (medicine); Jennifer M. Logg, *Theory of Machine: When Do People Rely on Algorithms?* 5 (Harv. Bus. Sch. Working Paper, No. 17–086, 2017) (financial decisions); Metz et. al, *supra* note 56 (legal practice).

100. There have of course also been instances of the opposite problem—attorneys trusting software too much—which has ended up at times with citations to fictitious court cases.

101. Prosecutors also report that relationships with judges and defense attorneys influence case outcomes. *See* HEUMANN, *supra* note 50, at 120.

102. *See Human Decisions and Machine Predictions*, *supra* note 27, at 278–79 (discussing factors that may influence judge-made decisions about bail not captured by the study's algorithmic prediction model).

AI system for a task and was frustrated because it did not quite get the subtlety will understand this.[103] In the Clark County Public Defender office, as everywhere, data is imperfect, and data will never capture the entirety of any setting. But what we saw in Part III.B was that the data available and the algorithm constructed were good enough to substantially outperform attorney predictions. Even so, a small number of inaccurate predictions frustrated users and was one of the factors that decreased usage, even among those who were initially interested.

Beyond accuracy, our interviews also revealed inertia and inconsistency as barriers to adoption. Some saw the value in the prediction tool but just never managed to build it into their workflow. Although we sought to make the user interface as easy as possible, it still required logging on to a separate website instead of being incorporated into the case management software they were already using. Additionally, there were occasional disruptions to use. It turns out that the Public Defenders in this office take temporary rotations on different "tracks" fairly frequently—something we hadn't fully understood in advance.[104] When this happened, it made generating sentencing predictions more complicated and in many cases impossible.[105] Although these track changes were usually temporary, typically lasting a couple weeks, that was long enough for the attorneys to get out of the habit of using our software.

These experiences suggest several key takeaways. First, the algorithmic system must be developed with input from users to capture the nuance of the setting. Despite our lengthy process of design and iteration, there were still nuances of the defenders' workflow that we did not understand, such as the frequency of track rotation. If we had designed the system to be more flexible from the beginning, perhaps we could have avoided the interruptions that arose from track rotations.

Second, users must be trained to recognize the abilities and limitations of the system. We did not conduct the quiz demonstrating how much more accurate our algorithm was than public defenders until after the conclusion of the experiment. If we had collected and disseminated this evidence in advance, users might be more convinced of its value. They would know that, even despite the valid concerns about incomplete data, that the tool provides valuable information that they could learn from.

In addition to improving trust in the predictions and integrating the software more seamlessly into existing routines, successful technology rollout can benefit from the support of senior leadership, high take-up among role models, and temporary incentives to induce initial adoption. Rather than having strong, supportive messaging coming from the Head Public Defender, all of the messaging and

---

103. *See* Cade Metz, *Chatbots May 'Hallucinate' More Often Than Many Realize*, N.Y. TIMES (Nov. 6, 2023), https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates.html [https://perma.cc/HWV7-U83T].

104. A track is simply a group of roughly four to six public defenders and two judges to whom their cases are assigned.

105. This was due to the fact that we did not generate predictions for specialized tracks, such as domestic violence or homicide.

training came solely from our research team. The lack of vigorous support from up top may have made PDs more hesitant to adopt a new workflow.

Other role models may also be important in adoption of new technologies. We noted that PDs with over five years of experience used the software at less than a third of the rate than those in their first year (Figure 6). Additionally, many in leadership roles weren't even eligible to participate in our experiment. This choice was made because they tended to handle either a much smaller or no caseload directly and therefore would not have generated much data for the experiment. But this lack of senior level role models may also have hurt software usage and consistent use. Incorporating PDs in leadership roles into the experiment as well as the trainings could have helped encourage more junior employees to adopt.

Finally, we had hoped to provide incentives in the form of coffees, meals or gift cards to employees who met certain targets in software usage, whether in absolute levels or relative to peers. Even when new technologies are ultimately very useful to a user, learning something new or developing new habits can be difficult. Evidence from behavioral economics studies shows that temporary incentives can provide initial "activation energy" to get over inertia and develop new work processes.[106] Unfortunately, we did not get approval from the head public defender for this.

The challenges discussed here will be faced in many employment settings where AI software is introduced as a tool to help workers. Leadership from management and office role models may provide that activation energy, as may incentives linked to usage. Ultimately, though, if the new technology is to be successful and useful in the long run, it must prove to workers that it helps them do their jobs, even after short-run incentives and exhortations are long over.

B. Barriers to Accuracy: Missing, Incomplete or Inaccessible Data

In reviewing the data, one fact that was immediately clear is that usage waned over time, even among attorneys who gave the software a serious try in the initial phases. Why? Our survey and usage data suggest that prediction accuracy remained the biggest concern. As discussed above, PDs seem to want the algorithm to be perfect, and quickly lost trust when it proved to be fallible—even though humans can be fallible, too. Here we discuss the biggest barriers to improving the accuracy of our predictions: incomplete, inaccurate and inaccessible data.

Digitized data relevant to sentencing predictions is not all in one place. The public defender, police, prosecutor, courts, and both state and national criminal record systems all maintain separate data systems and sometimes won't share with each

106. *See, e.g.*, Yaman Roumani, Joseph K. Nwankpa & Yazan F. Roumani, *The Impact of Incentives on the Intention to Try a New Technology*, 27 TECH. ANALYSIS & STRATEGIC MGMT. 126, 136–39 (2015); Emma Riley, Abu Shonchoy & Robert Darko Osei, *Incentives and Endorsement for Technology Adoption: Evidence from Mobile Banking in Ghana*, 1–58 (Ctr. for Effective Glob. Action, Working Paper No. 240, 2024), https://escholarship.org/uc/item/8335r6n6 [https://perma.cc/SZQ4-TF7C].

other, let alone with outside researchers. We sought to get data from a variety of these sources and were shut down for various reasons. The courts might have been amenable to it, but they did not have an IT person capable of bulk downloading court records. State and national criminal record systems are protected by law and not available to external researchers. The prosecutor's office did not feel comfortable providing data to us. For these reasons, our algorithm lacked a number of key pieces of information: additional details about the offense as compiled in the police report, the prosecutor assigned to the case, and more detailed criminal record information beyond cases handled by the Clark County Public Defender. These limitations are fairly standard with criminal justice data, but they are meaningful limitations nonetheless.[107]

Furthermore, the case management data we had access to was inputted manually by a team of paralegals. Again, this is common with criminal justice data, which is entered somewhat painstakingly and tends to be error-prone.[108] If we had access to court, prosecutor and police records, we might have been able to crosscheck the data and improve its accuracy. But legal, cultural, and resource restrictions prevented that.

Thus, we conducted our experiment in a setting where there are serious data limitations. Criminal justice data is notoriously messy and inaccurate, as well as siloed across different sources. Without sufficient data, even the best algorithms cannot make useful predictions.

We chose not to present predictions in cases where the predictions were particularly uncertain: about 5% of cases. These omitted predictions disproportionately involved uncommon offenses and atypical cases, which, by definition, were underrepresented in the training data. But this proved a frustration to public defenders, since these are often the cases where they wanted predictions most. Deciding whether to supply algorithmic advice, even when uncertainty is great, requires navigating tough decisions with no clear answers.

In our modern lives, almost everything is digitized to a certain extent. But the quality, completeness, and accessibility of data can vary widely across domains. In some domains, the normal daily process of human life generates vast amounts of data: we send emails, use our phones, post on social media. Many people make their living generating information for the Internet—journalists, bloggers, advertisers, content creators, etc.—thus generating rich data sources that can be mined by AI and predictive algorithms. AI tools are likely to be most successful in domains with rich and accurate data.[109]

In contrast, predictive tools face more challenges in domains with lower quality data. Criminal justice is a setting in which data is limited, inaccurate, and difficult

---

107. Wayne A. Logan & Andrew Guthrie Ferguson, *Policing Criminal Justice Data,* 101 MINN. L. REV. 541, 548–49 (2016).

108. Sarah Esther Lageson, *Criminally Bad Data: Inaccurate Criminal Records, Data Brokers, and Algorithmic Injustice*, 2023 U. ILL. L. REV. 1771, 1778–86 (2023).

109. Hal R. Varian, *Big Data: New Tricks for Econometrics,* 28 J. ECON. PERSPS. 3, 3–5 (2014).

to access, but it's not the only one. Healthcare data, for instance, is governed by restrictive laws protecting privacy.[110] Private firms collecting data may be reluctant to share both due to reputational concerns or business interests. Legal protections such as copyright provide further barriers, although the extent of such limitations is still being litigated.[111] Beyond the barriers to accessing already-digitized data, there are also the limits on what *types* of data get digitized. Conversations in hallways between attorneys, lunchtime chats with judges, courtroom attitudes and behaviors—even in our digital world, there remain important parts of our life that don't leave a digital footprint. When digital data is sparse, error-prone, or unavailable, there are limits to what even the most advanced algorithms can predict.

C. When "Fairness" Depends on Accuracy as Much as Bias

There is a large literature on algorithmic fairness. The major concern is that algorithms will replicate historic bias when it is trained on data where bias is present.[112] Consider a sentence prediction algorithm that included race as an input. If the algorithm is trained on historical data where there were racial disparities in sentencing, the algorithm's recommendations will have disparities as well. A Black defendant whose case and criminal history is identical to a white defendant's would be predicted to get a longer sentence. A public defender equipped with these predictions might recommend that the Black defendant accept a plea offer while, if the same plea offer was provided to a white defendant, the PD would recommend fighting for something better.

On the one hand, this is concerning. Baking historical disparities into a sentence prediction has the potential to reinforce them. Unless biases are challenged, they could cycle through in perpetuity.

On the other hand, if the system is biased against Black defendants, the public defender *should* take this into account when making recommendations to her clients. Telling Black clients to fight for the same sentence that might be offered to a white defendant might backfire. A Black defendant who rejects a plea deal that is seen as reasonable by court actors in order to take the case to trial could wind up with a lengthier sentence due to the trial penalty.[113]

This context—providing sentencing predictions to public defenders—raises very different algorithmic fairness issues than those typically considered in criminal justice.

---

110. Nicolas P. Terry, *Regulatory Disruption and Arbitrage in Health-Care Data Protection*, 17 YALE J. HEALTH POL'Y L. & ETHICS 143, 148–59 (2017); HIPAA Privacy Rule, 45 C.F.R. pts. 160, 164 (2025).

111. N.Y. Times Co. v. Microsoft Corp., 777 F. Supp. 3d 283, 300–01 (S.D.N.Y. 2025).

112. *See* Harcourt, *supra* note 11, at 237–38; (describing how parole risk assessment tools use race as a proxy for recidivism risk); Sahar Takshi, *Algorithmic (Un)Fairness: Artificial Intelligence-Based Discrimination Against Individuals with Disabilities*, 57 CRIM. L. BULL. 505, 505–06 (2021); Mayson, *supra* note 8, at 2259–63.

113. The trial penalty is sometimes defined as the difference between a sentence after a defendant is found guilty at trial relative to a plea offer. *But see* David S. Abrams, *Is Pleading Really a Bargain?*, 8 J. EMPIRICAL LEGAL STUD. 200, 220 (2011); David S. Abrams, Viet Nguyen, Aurelie Ouss & Julia Reinhold, *A (Plea) Offer You Can Refuse* (Aug. 2024) (unpublished working paper) (on file with author) (discussing why other definitions may be more relevant in defendant decision-making).

Most predictive algorithms in criminal justice are risk assessments: predictions of the likelihood that the person will be charged with a new crime or some sort of misconduct in the future. With risk assessments, a prediction that is built on historical biases disadvantages the biased-against group, since it makes people overestimate their relative potential for crime. With sentence predictions, the incorporation of historical biases has a more ambiguous effect on the well-being of the defendant. On the one hand, it could perpetuate these biases by making them part of sentencing expectations. On the other hand, if such biases remain part of sentencing, this can't be ignored. A public defender advocating zealously for their client would need to advise Black and white clients differently, since the sentencing process works differently.

We spent a great deal of time discussing this issue both with PDs, other attorneys, and academics who have researched the topic. We discovered that people had differing intuitions even in our context—some very strongly held. Given the degree of controversy over including race in the algorithm, we were concerned that it might lead some PDs to decline to participate in the study or to be less likely to make use of the predictions. To better understand the issue, we investigated whether there was any decrease in predictive accuracy when defendant race was excluded from the predictions. We found no loss of accuracy overall or by race. This is because race is highly correlated with other variables that we included in the dataset, including zip code.[114] We decided to exclude race as a prediction variable in our algorithm, assured that doing so would not affect prediction accuracy. Of course, simply excluding race does not mean that the algorithm is free of historical biases, which could be captured both by ZIP Code and the criminal record. We communicated this to the defenders in our training sessions.

Even though concerns about algorithmic bias are arguably less severe in this context (because embedding historical biases into the prediction is arguably to the defendant's advantage) we still learned that they had an impact on usage. Fairness concerns may be magnified in other criminal justice contexts, where biased algorithms are more clearly problematic for the discriminated-against group. Thus, it is incumbent on developers of software algorithms to take fairness concerns seriously and also to take a great amount of care in clearly explaining the choices made to users.

CONCLUSION

Now is a moment of tremendous promise and potential change to the job of the lawyer due to revolutionary new technologies. In particular, AI and algorithmic prediction hold potential to make attorneys more productive and get better results for their clients. At the same time, some have expressed concern about how software may change our lives and jobs for the worse. But this promise or peril will only be realized if legal experts—and experts more broadly—are willing to make use of the new technologies.

---

114. Mayson, *supra* note 8, at 2263.

In this article we have presented results from a field experiment that made a novel predictive technology available to public defenders. But this article is not about the impact that software had on criminal sentences. Despite a great amount of effort spent on the software and experiment, that analysis is not possible—because the public defenders made little use of the algorithmic predictions.

Nevertheless, there is a great deal to learn from the experiment as well as the additional survey and quiz data we collected from the public defenders. We identify three primary lessons from our study. First, legal experts have a very high bar for adopting new technology in their field of expertise. This is due to the fact that their extensive knowledge can make them skeptical of the utility of the new technology and also because of inertia in one's work. We suggest a few ways to potentially address these issues, including endorsement from leadership, role models making use of the new software, and incentives for initial use.

Second, algorithmic and AI predictions will not be effective in some important areas due to a lack of data availability. This can come from legal restrictions and data siloing, but also a lack of digitized data or unusual cases making accurate predictions impossible. Without consistently accurate predictions, expert legal users are unlikely to regularly use the new algorithms.

Finally, ethical and fairness concerns may have a substantial impact on adoption. Users are well aware of issues such as algorithmic bias and may reject algorithms they see as unfair or problematic. Algorithmic developers should work closely with various stakeholders as well as users of the algorithm to ensure that design decisions are made conscientiously and communicated clearly.

We will undoubtedly continue to see an increase in the use of AI and predictive algorithms in criminal justice and the law more broadly. But there remain substantial barriers to uptake. Learning to extract benefits from the AI revolution will require learning how to get past these barriers; to take advantage of it in domains in which it excels, while relying on alternative sources of information where it does not.

APPENDIX TABLES

Table A1: Sentencing Prediction Usage by Felony Class

	Used Prediction		
	Mean	SD	N
A	0.114	(0.320)	70
B	0.105	(0.307)	1233
C	0.116	(0.321)	352
D	0.103	(0.305)	117
E	0.090	(0.287)	256
NA	0.086	(0.280)	269
Total	0.103	(0.304)	2297

F-statistic: 0.45, p-value: 0.82

Note: Class A is the most serious felony, Class E is the least serious, and NA indicates the charge was likely a misdemeanor.