Predictive Algorithms in the Frontlines of the Administrative State: Empirical Evidence from Child Protection Screening

Amit Haim*[1], Rhema Vaithianathan[2]

[1]Tel Aviv University, Israel
[2]Centre for Social Data Analytics Auckland University of Technology, New Zealand

January 25, 2026

Abstract

Administrative agencies increasingly deploy predictive algorithms to support frontline decision-making, yet evidence on how such tools affect real-world practice remains limited. This study examines the introduction of a predictive risk model (PRM) used by Child Protection Services (CPS) in a large U.S. county to assist screening decisions for allegations of abuse or neglect. A technical constraint in the county's case management system generated quasi-experimental variation in access to the PRM: the score could be displayed in morning screening sessions but not in afternoon sessions. Using approximately 130,000 client-referral observations from 2020–2023 and a difference-in-differences design, we estimate the intention-to-treat effect of PRM availability on screening and investigation-track assignments.

We find modest but systematic impacts. Access to the PRM reduced screening-in rates for low-risk referrals and increased the likelihood that screened-in cases were assigned to the less intrusive assessment track. Effects did not extend to higher-risk referrals and attenuated over time, suggesting that organizational routines and worker discretion mediate algorithmic influence. The PRM's presence improved alignment between predicted and assigned investigation tracks, though effects were small in magnitude.

*We thank Sandy Handan-Nader, Jacob Goldin, Haggai Porat, Megan Stevenson, Alex Albright, David Freeman Engstrom, Rob Maccoun, and Katherine Strandburg for valuable comments. We thank participants at the 2024 meeting of the American Law and Economics Association and the 2024 Conference of Empirical Legal Studies for their input. We thank Larissa Lorimer for help in producing data. All mistakes are ours. Partial funding for this study was provided by Stanford Law School.

These findings highlight the importance of evaluating predictive tools within the conditions of field deployment rather than in isolation. In hybrid human-algorithm systems, the effects of predictive models depend not only on their statistical performance but also on the institutional, temporal, and behavioral contexts in which they are embedded.

Contents

1 Introduction

Algorithmic tools, including machine learning and predictive techniques, are becoming increasingly common in public administration (Engstrom and Haim 2023; Engstrom, D. E. Ho, et al. 2020). Their adoption reflects longstanding concerns about decisional accuracy, consistency, and fairness in the administrative state (Davis 1969; Ho and Sherman 2017; Lipsky 2010). Across many domains, officials make high-volume, high-stakes decisions that often exhibit unwarranted variation (Kahneman, Sibony, and Sunstein 2020; Ramji-Nogales, Schoenholtz, and Schrag 2007; Ho, Marcus, and Ray 2021).

A growing literature argues that algorithmic tools may serve as internal governance mechanisms that help standardize decision processes, reduce noise, and improve the allocation of interventions (Glaze et al. 2022; Kleinberg et al. 2018; Grove and Meehl 1996). Yet most empirical research evaluates algorithms in isolation—implicitly assuming that they replace human decision-makers. In practice, however, algorithms typically *supplement* rather than supplant human discretion, especially in high-stakes legal and social services settings.

Recent work across public agencies emphasizes the need to study algorithmic tools as socio-technical systems, where human decision-makers retain meaningful authority (Gillis, McLaughlin, and Spiess 2021; Engstrom and Haim 2023). A small but growing empirical literature examines such hybrid systems in real-world contexts, particularly in criminal justice (M. T. Stevenson and Doleac 2024; M. Stevenson 2018; Albright 2024; Angelova, Dobbie, and C. S. Yang 2023; Goel et al. 2021). Comparatively little is known about algorithmic implementation in local social services, despite their scale and policy importance. Within child protection, recent studies show mixed effects, ranging from improved accuracy and reduced disparities (Grimon and Mills 2025; Rittenhouse, Putnam-Hornstein, and Vaithianathan 2023) to heightened surveillance of low-income families (Parker et al. 2022; Eiermann 2024). Other work finds that frontline staff may strategically mitigate algorithmic recommendations (Cheng et al. 2022).

This study examines how child protection workers use and respond to a predictive risk model (PRM) deployed in a large county in the Western United States. The PRM produces a 1–20 risk score for each referral, intended to support—but not dictate—screening decisions. A technical limitation in the agency's case management system created quasi-random variation in tool availability: the score could be displayed only in morning screening sessions, yielding a natural treatment (morning) and comparison (afternoon) group. This design provides an opportunity to estimate an intention-to-treat (ITT) effect of tool availability while holding constant the institutional context and workforce.

Our analysis uses approximately 130,000 client-referral observations between 2020 and 2023. We study whether access to the PRM affected (i) the probability that a referral was screened in; (ii) the assignment of screened-in referrals to investigation tracks; and (iii) alignment between assignment decisions and model-predicted risk. Importantly, because the PRM does not incorporate facts from the current referral narrative, its predictions do not fully correspond to

the legal standard for screening—an issue that bears directly on evaluation and interpretation.

Previewing our results, we find modest but systematic effects. Access to the PRM reduced screening-in rates for low-risk referrals and shifted some screened-in cases toward the less intrusive assessment track. Effects are concentrated in the early months of deployment and attenuate over time. Consistent with prior work on hybrid systems, worker discretion and organizational routines appear to moderate the influence of algorithmic information.

This study contributes to scholarship on administrative algorithms by providing field evidence on how frontline workers integrate predictive data into time-constrained, team-based decision-making. Rather than evaluating algorithmic accuracy in isolation, we assess how the presence of a predictive tool changes human decisions, with implications for caseloads, triage, and internal governance.

2 Background

2.1 Accuracy and Consistency in Administrative Decision-making

Poor quality and inconsistent administrative decision-making is found across various domains, including immigration (Ramji-Nogales, Schoenholtz, and Schrag 2007), social security benefits (Hausman 2021), food inspections (Ho 2017); and in all levels of government. Unwarranted variation in decisions is present between officials and even within the same officials over time (Ho and Sherman 2017).

Inconsistent and poor-quality decisions by administrative agencies are detrimental to the rule of law, and to the fair and efficient delivery of government services. When an outcome depends on the proclivity of an assigned official or on irrelevant extraneous factors, a basic notion of legality and legitimacy is undermined. For instance, systemic biases where decisions are based on irrelevant characteristic, such as race or socio-economic conditions, are antithetical to notions of fairness and may undermine trust in the government.

Poor quality of decisions in mass adjudication programs is widespread, despite various efforts to improve them (Ames et al. 2020). These problems have long preoccupied scholars and policymakers: Davis warned more than fifty years ago that administrative discretion and its arbitrariness is one of the most fundamental flaws in modern democracy Davis 1969). In the domains where policy is implemented by workers with considerable discretion, so called Street-Level Bureaucracies, inconsistencies are often endemic and can be devastating to the public legitimacy of these agencies (Lipsky 2010).

In street-level bureaucracies, the delivery of policy and services is administered through an array of frontline workers. A fundamental aspect of street-level bureaucracies is the discretion workers employ. Discretion is often touted as an advantage because it allows decisions to be tailored to the specific, unforeseen

circumstances of the person. Yet worker discretion may also lead to idiosyncratic (and unjustifiable) variation. Another pervasive problem of street-level bureaucracies is the difficulty to improve quality of services rather than quantitative expansion, as additional capacity is absorbed in reproducing the same level of quality at a higher volume (Lipsky 2010, pg. 38).

One approach to tackle the quality problem has focused on internal administrative processes. As Ho and Sherman suggest, there is a wide range of mechanisms at agencies' disposal, including quality assurance, peer review, monitoring, training, and those tend to work better than ex post methods (Ho and Sherman 2017). More recently, administrative agencies have begun to adopt various kinds of machine learning and artificial intelligence algorithms in their decision-making processes (Engstrom, D. E. Ho, et al. 2020). While the motivation for the adoption of these algorithmic instruments vary, one objective is to improve decision quality and consistency. As Glaze et al (2022) show in the context of Social Security adjudications, AI tools have the potential to significantly improve how decision-makers consider cases. Other theoretical and empirical work has shown that algorithms can more accurately allocate interventions to recipients (Kleinberg et al. 2018), even with simpler statistical models (Grove and Meehl 1996).

One issue to consider is that most research to date has focused on evaluating algorithmic tools in and of themselves, essentially assuming that they are supplanting human decision-makers. However, in the real world of administrative decision-making, algorithmic tools are more likely to serve as aids to humans. That is for various legal, reputational, and political reasons; at least for the foreseeable future and in high-stakes scenarios.

This approach has begun to change somewhat, with some works highlighting that algorithms should be studied notwithstanding human decision-makers retaining discretion (Gillis, McLaughlin, and Spiess 2021), and recognize the importance of hybrid decision-making and focusing on implementation (Engstrom and Haim 2023). But in the empirical scholarship, only a handful of recent studies have taken this approach and focused on real-world scenarios of implementation of algorithmic tools (Ludwig, Mullainathan, and Rambachan 2024). Those studies are mostly concerned with the criminal justice domain and judicial decision-making (M. T. Stevenson and Doleac 2024; M. Stevenson 2018; Albright 2024; Angelova, Dobbie, and C. S. Yang 2023; Cordella and Gualdi 2024; Starr 2014; Goel et al. 2021). Other administrative settings, and particularly in the local social services domain, have gained less attention in this regard – despite having substantial impact.

2.2 Predicting Risk in Child Protective Services

Child Protective Services (CPS) are administrative agencies with a precarious mandate. They are in charge of choosing whether to investigate allegations of child abuse and neglect. Following investigation, CPS may make a finding of substantiated abuse or neglect. They may also open a case, and continue to engage with the family, offering resources to mitigate risks and safety concerns.

In some case, CPS may seek to remove the child from their parents if it is deemed that the child is in danger and risks cannot be sufficiently mitigated.

A false negative assessment on the part of CPS (i.e. a determination that a child is not at risk when they are) could end in dire and even fatal results, making such mistakes worrisome and highly salient for agencies. Agencies could thus be expected to "over-investigate" , especially as failure to prevent calamities will likely result in negative publicity and political pressure.

On the other hand, needless CPS investigation is burdensome and often stigmatizing. Moreover, the majority of initial referrals that are received by child abuse hotlines come from community-based reporters, such as teachers or healthcare providers, who may be over-sensitized to report any and every hint of maltreatment because of their own liability. This is especially troubling when over-investigation is conflated with low socio-economic status or racial attributes, propagating the perception of child protection as part of a surveillance state which disproportionately targets the disadvantaged (Redden, Dencik, and Warne 2020).

Thus, CPS agencies find themselves in a double-bind, between family preservation and protection (Scherz 2011). This leads to a needle in a haystack type of problem – out of the numerous reports of potential maltreatment, CPS need to sift out those high-risk cases which truly necessitate an investigation. Overall, around half of all referrals of abuse and neglect to hotlines in the United States are investigated. The formal legal basis for CPS involvement and initiating legal action is the statutory definitions of conduct involving harm or creating substantial risk of imminent harm to a child.[1] Even if the agency can substantiate an allegation and ascertain what had happened, it must also consider if an event or similar behavior is likely to reoccur. Besides this determination, which involves assessing both former conduct and future risk, it is mostly also needed to establish whether CPS involvement will mitigate future harm, and relatedly what will be the outcomes of different interventions (such as supervision, counseling, foster care placement, or others). Thus, the decision to intervene regularly involves some prediction of future parental behavior and levels of risk.

Because state law adopts a low threshold for hotline referrals, CPS systems admit a large volume of cases at the “front door,” even though only a fraction proceed to formal investigation or court involvement. Therefore, it is sensible for agencies to concentrate on improving case triage, as there is reason to believe too many cases are entering the process. The limited resources that agencies have are better invested in cases with real potential of danger.

CPS agencies have used risk assessment methods for decades in multiple stages of the triaging process, in order to determine which cases require heightened attention and resources. Traditionally, such risk assessment was based on unaided clinical judgment and was gradually formalized into team-based procedures, following the empirical observation that teams tend to perform better in risk prediction (Drake et al. 2020). In more recent times, there has been a steady shift to actuarial methods, which include instruments empirically demonstrated

[1] Specific requirements vary between states but are of similar characteristics.

to be predictive and outperform clinical methods (Baird and Wagner 2000). Actuarial methods are anticipated to improve clinical accuracy in risk prediction and thus improve agency response, and increase consistency across case-workers and other decision-makers which is a lingering problem in CPS (Benbenishty et al. 2015; Haan and Connolly 2014; Russell 2015).

Currently, the most common system in use is Structured Decision-Making (SDM), a decision-aid tool which uses worker input and standardized items to assist decisions on child risk and safety in multiple critical decision-points (initiation, investigation disposition, case planning, ongoing case evaluation, reunification, and case closure) (Drake et al. 2020). SDM is based on a simple additive procedure which adds and deducts points according to different attributes and yields a risk score (Gillingham and Humphreys 2010). Unlike modern machine learning tools, the SDM weights are not typically validated for the jurisdiction in which it is used - rather the same weights are used across multiple States and Counties.

2.3 Predictive Risk Modeling

In the last decade, with the move to electronic child welfare case management systems, research began to demonstrate the advantages of Predictive Risk Modelling (PRM). PRM methods predict adverse outcomes, such as re-referral, removal, or severe injury following a referral (Vaithianathan et al. 2013). Unlike SDM, PRM does not rely on reported concerns or clinical assessment, but rather on a broad set of historical information regarding family members, alleged perpetrator, and children that are already held in the case management systems; and so is less susceptible to subjective assertions and mistakes as it only requires a call-screener to input basic intake data.

Beyond the theoretical proof of concept, in recent years PRM systems have been incorporated in the field by several CPS agencies around the U.S. and worldwide (Church and Fairchild 2017; Cuccaro-Alamin et al. 2017; Schwartz et al. 2017). For instance, one of the notable cases is in Allegheny County in Pennsylvania, where the Allegheny Family Screening Tool (AFST) was developed by a team of researchers and deployed in 2016 (Brown et al. 2019; Chouldechova et al. 2018; Goldhaber-Fiebert and Prince 2019). The AFST aids hotline center workers in screening cases, by categorizing the reported case on a risk score scale, and other agencies are working on adopting similar tools (Vaithianathan 2019). The use of PRM in CPS has sparked fierce public and academic debate. Some voices are essentially skeptical, viewing it as part of increased surveillance or an attempt to solve deeper problems with the use of technology (Eubanks 2018; Keddell 2019). Other critics have focused on the risk that these tools displace human judgement and effort and emphasized the need to maintain the workers' autonomy as the final arbiter (Cheng et al. 2022; De-Arteaga, Fogliato, and Chouldechova 2020; Kawakami et al. 2022).

Recent empirical studies have added nuance to this debate, finding mixed evidence. Rittenhouse et al (2023), looking at the AFST implementation, find the tool reduced racial disparities in screening decisions, case openings, and

home removal rates for Black children compared to White children. Grimon and Mills (2025) find in a similar setting to ours, using a field randomized controlled trial, that overall accuracy of CPS intervention has improved when a PRM tool was introduced, leading to a decrease in hospital-reported injuries. Moreover, they find the tool lowered racial and socio-economic disparities in intervention rates.

On the other hand, Parker et al find mixed results between counties in the same state (2022). While PRMs led to an increase in investigation of families from lower socio-economic strata in some counties, others saw a decrease in disparities. Likewise, Eierman (2024), looking at aggregate data from 121 counties around the country finds no clear effect of reducing future maltreatment assessments, but decreasing gaps across racial groups, while potentially increasing interventions for children of low socio-economic families. Notably, both studies focused on the use of PRMs in the investigation stage, that is after a referral was screened in; as opposed to the previous studies discussed, and our own study, in which the PRM tool is utilized at the screening stage.

2.4 Screening Process in the Study County

The present study focused on a County in a State in the Western United States. The County is a semi-urban county, one of the largest in the State in size of population. Child welfare issues are handled by the Department of Human Services, and specifically by their Child and Adult Protection Services Division (CAPS or the Division). The Division is in charge of all allegations of child abuse and neglect.

Allegations of child abuse or neglect are mostly reported to the Division through the County or State hotline. A hotline worker receives the call, collects needed information from the reporter, and generates a referral in the case management system. A hotline supervisor then reviews the referral and makes a preliminary triage decision – immediate response cases are promptly transferred to a caseworker on duty or the police department (e.g. a reporter who saw an infant locked in a car). Calls that are clearly not allegations of abuse or neglect are dismissed or referred to other, more relevant government functions (e.g. a person calling the hotline in request for information). All other cases are transferred to the Division screening process.[2] Figure 1 shows the sequence of decisions involved in a referral.

[2]To be eligible for screening, a referral must meet one of the following conditions by state law: (1) There is a vulnerable child in the home (either under 6 years, or with physical/developmental/mental disability); (2) There are two or more screen outs within the last 12 months (in any county in the state); (3) There are three or more assessments (i.e. assignments) within the last 12 months (in any county); (4) Relevant criminal history for adults in the home (e.g. felony/misdemeanor conviction of child abuse; any conviction of a violent crime; any conviction of unlawful sexual behavior); or (5) meets criteria for Family Assessment Response (FAR). Referrals that don't meet these criteria are not routed to screening teams, but instead are disposed by the supervisor in consultation with her colleague. These amount to roughly 25 percent of all hotline calls. Since screening teams are considered best practice, the hotline supervisor tries to route most referrals to them.

As part of the Differential Response model used by the County, screening decisions are made by a screening team comprised of a facilitator (who is a supervisor or administrator in the Division) and up to three caseworkers.[3] For simplicity, we call the facilitator of a screening team a *lead supervisor*. On business working days, several teams meet in the morning for a one-hour session, and several meet in the afternoon. Screening referrals is not the main task of Division workers, but rather a duty they perform several times a month, such that there are varying degrees of experience and familiarity with the procedure. Case-workers and supervisors are required to regularly sign up to teams according to their own scheduling constraints., such that teams are not consistently comprised by the same members.

In each screening team meeting, the team discusses up to four referrals, assigned by the hotline supervisor.[4] The team reviews several sources of information for each referral: the narrative and details collected by the hotline from the reporter; any available data in the State case management system on prior referrals, including dispositions; criminal history and other court involvement (e.g. family disputes). The team then makes a disposition decision on the referral, based on State law and guidelines pertaining to the reported child or children being in imminent risk of serious abuse or neglect.

By and large, the screening team can choose from one of three responses available: they may screen out the referral which means no further action, direct it to an informal assessment track called the Family Assessment Response (FAR) or assign it to a formal investigation called High Risk Assessment (HRA). Screening teams do not conduct the investigation, but rather direct it to the relevant team in the Division. Only allegations that are assigned to the HRA track require a formal finding by the assigned caseworker, while FAR cases are handled less formally in order to reduce invasive involvement. Some referrals are mandated by the State to be assigned to the HRA track (e.g. sex abuse allegation or an institutional referral). Additionally, the team needs to decide between a three-day and a five-day response window, which determines the timeframe in which a caseworker assigned the case must go out to meet the reported child.[5]

Referrals that are assigned to either HRA or FAR (which we refer to as screened-in) are then handled by intake caseworkers, who meet the child and the family, collect information, assess the situation, and make a determination about the appropriate response within 60 days. The caseworker can substantiate an allegation or find it to be unsubstantiated. Importantly, referrals that are assigned to the HRA track and are founded, are mostly listed on the State registry and may impose restrictions for perpetrators, such as debarring cer-

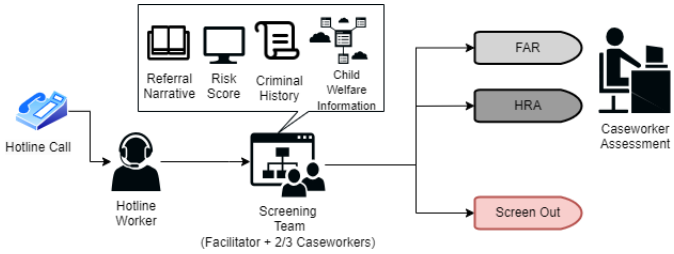
[3] Officially, screening teams are called RED teams: Review, Evaluate, and Direct.

[4] The team only has one hour to discuss the referrals. Any spill-over referrals are handled by the lead supervisor. The Division differentiates between busy days and less busy day. On busy days, all referrals must be disposed within that same day, so the lead supervisor in charge disposes them on their own in what is called direct disposition. On less busy days, referrals can be reallocated to afternoon or next day screening teams.

[5] A one-day immediate response assignment exists, but is mostly not relevant for screening teams as it is handled by the hotline supervisor and workers.

tain jobs. Furthermore, based on a founded allegation and an assessment, the Division can move to execute further actions, mostly through court proceedings, such as a temporary removal of a child, placement in foster care or with relatives, and in extreme cases even to terminate parental rights.

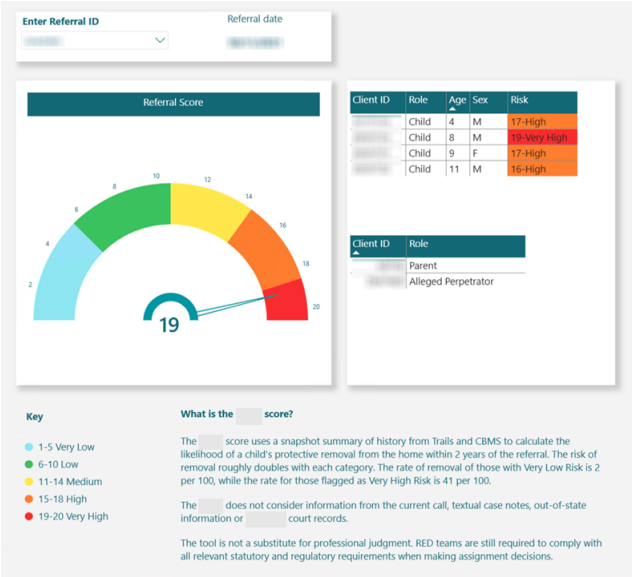
Figure 1: Referral Screening Process



On December 16, 2021 the Division deployed a PRM tool for its screening teams. The tool uses machine learning methods to predict the risk probability. The model uses a LASSO regularized logistic regression, with an outcome variable of removal from home within 24 months. The model was trained on a dataset of past child welfare referrals in the county and other statewide counties using historical data from 2010-2014 on approximately 221,519 client-referral observations (i.e. unique child-referral observations). It takes into account available information in the county case management system, such as previous allegations and their assessments, but does not rely - by design - on the current referral. A total of 460 features are included in the model, relating to demographics (excluding direct indicators of race/ethnicity and disability), past child welfare involvement of household members, as well as limited case management data with social services. The initial choice of which variables to include for the model was made in agreement with local implementing partners. Although many variables are included in the model, most of the predictive power of the tool comes from child welfare history of household members. The area under the receiver operating characteristic curve (AUC-ROC) of this model is 76.2 percent, which indicates good discrimination.

When participating in the screening team meeting, the facilitator has access to a dashboard that presents the score for that referral, ranging from one to twenty and accompanied by a color (green to red) and a risk category quantile (very low to very high) (see Figure 2). The Division instructed facilitators to use the tool in screening meetings as an additional piece of information, but emphasized that the team is required to otherwise conduct its deliberations in the same manner as before the introduction of the tool.

Figure 2: PRM Tool Interface



3 Methodology

3.1 Data

The data for this study comes from the state child welfare data system. The system is a case management system which stores referrals, investigations and cases arising from child abuse or neglect allegations. Administrative data for referrals associated with the Study County were extracted and after cleaning the data, we had 129,479 client-referral observations (that is unique child-referral observations). The data spans referrals that occurred between October 1, 2020 to September 1, 2023 (approximately one year before deployment of the PRM tool on December 2021, and two years post deployment). In this period, the Division handled an average of 186 referrals per day. Each referral may have multiple children named, and each child may be named on multiple referrals over this period, as subsequent referrals may indicate the same child or family. We had a total of 65,436 unique children in the dataset.[6]

The data includes information about each referral, including demographic data about the child and family, general administrative aspects (e.g. timestamps, codes for dispositions etc.), the type of reporter, reason for referral, and the screening outcome. Additionally, the dataset includes the predicted risk score by the PRM tool for any given referral (a raw probability, a numerical prediction between one and twenty, and the risk category ranging from very

[6]The data was further restricted according to several other criteria, to ensure proper estimation, as described below in section 3.3.

low to very high risk). Importantly, the data includes the prediction for any given referral, *regardless* of whether the tool was deployed and available for the screening team in real time, which allows us to control for predicted risk levels.

Additionally, we received anonymized data from the Division about the case-workers who participated in the screening team screening meetings for each referral. After cleaning the data and identifying supervisors, we combined the datasets such that for each referral that went through a screening meeting, there was information about the supervisor that led the screening meeting.[7] Since not all meetings had data on the participants, or that none were listed as supervisors on the county's lists, we were left with 22,804 observations, which are used to estimate supervisor-level effects (see section 3.3).

3.2 Descriptive Statistics

Table 1 presents descriptive statistics from the sample of analysis. The sample includes 65,436 children listed on 37,079 referrals. Of these, 14 percent were referrals about "active" families - namely those that had an existing CPS case. For the remaining 86 percent, the referrals where for non-active families.

In terms of race, the sample was predominantly White with 39 percent of children being identified as such, followed by 17 percent identified as Black and 18 percent as Hispanic in ethnicity (less than few percent of the sample cumulatively had other racial designations such as Indian or Native American, Asian or Pacific Islander). A significant portion of the sample, 44 percent, had unknown race demographic categories, accounting for the remainder of the data.[8]

Referrals are categorized at the hotline with reasons for referral coded in up to 34 categories. Injurious environment, which is a form of neglect, is the most common reason, appearing for 28 percent of children. Two of the other largest reason categories are physical abuse (8 percent) and parental substance abuse (20 percent). Other common reasons for referral include sexual abuse, emotional abuse, and medical neglect.

Referrals are also categorized on the type of reporter, out of 51 possibilities. The most common reporters are medical staff (18 percent) and schools (16 percent). Most of the reports come from mandated reporters (74 percent), such as medical staff, law enforcement, or school staff.

[7] Some meetings were listed as having more than one supervisor present. This could arise either a variety of situations such as a specialized supervisor may have been asked to attend if a referral required their input, a supervisor shadowing for training or quality assurance purposes, a coding error or some other unknown reasons. We defined the lead supervisor in a meeting as the supervisor that participated in all referrals in that meeting on the basis that when supervisors were shadowing or providing input for a specific case, they are unlikely to have. If there was more than one, we randomly assigned the lead supervisor tag. The estimates in the analysis section are robust to different randomization seeds.

[8] Since the data is collected at the hotline stage, race is identified by reporters, and not by the clients themselves. Race can be missing if the reporter is unsure about it, or for coding errors. In open cases racial designations are confirmed with the family, such that recurring clients are more likely to have their correct racial designation reflected in the data.

We also report the average outcomes for the referrals in the data, with 42 percent of all referrals being screened in for further investigation, and 19 percent are characterized as HRA (a little less than half of screened-in referrals). The average risk score predicted by the PRM tool in our sample is 11 (SD = 6.5).

Table 1: Summary Statistics

Variable	**Mean**	**SD**	**Variable**	**Mean**	**SD**
Screen In	0.42	0.49	Alleged Substance (Parent)	0.20	0.42
HRA	0.19	0.39	Alleged Sexual Abuse	0.035	0.18
FAR	0.24	0.42	Allegation Missing	0	0
Active Case	0.14	0.34	Mandated Reporter	0.74	0.44
Risk Score	11	6.5	Reporter: Community	0.026	0.16
White	0.39	0.49	Reporter: Counselor	0.024	0.15
Black	0.17	0.38	Reporter: CYF	0.34	0.47
Hispanic	0.18	0.39	Reporter: Family	0.07	0.26
Indian or Native American	0.0079	0.089	Reporter: Law Enforcement	0.031	0.17
Asian	0.014	0.12	Reporter: Medical Staff	0.18	0.38
Hawaiian or Pacific Islander	0.0059	0.077	Reporter: Other	0.063	0.24
Race Missing	0.44	0.5	Reporter: School	0.16	0.37
Alleged Injurious Environment	0.28	0.49	Self Report (Perp)	0.036	0.19
Alleged Physical Abuse	0.085	0.28	Self Report (Victim)	0.0017	0.041
Alleged Emotional Abuse	0.056	0.23	Anonymous Report	0.025	0.16
			Reporter Missing	0	0

Note: This table contains summary statistics for the sample, covering the full dataset (both pre- and post-deployment of the PRM tool). $N = 129{,}405$.

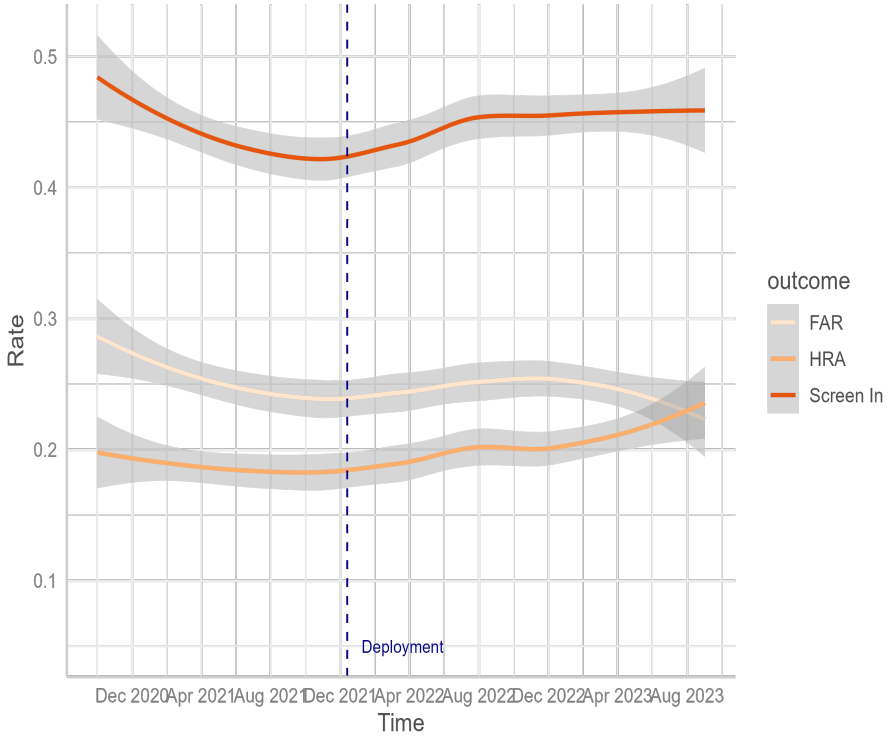
Figure 3 shows the average rates of the outcomes of interest over time, both before and after the deployment of the PRM tool (deployment is marked by a blue dotted line). Overall, there do not seem to be substantial trend changes emanating from the tool being deployed.

3.3 Empirical Strategy

Our goal is to estimate the effect of access to the PRM tool on screening outcomes. Because tool access varied by time of day only after deployment, our design yields a natural intention-to-treat (ITT) framework. We leverage this structure using a canonical difference-in-differences (DiD) specification.

Setup. Screening teams meet twice daily: morning teams (where the PRM score could be displayed after deployment) and afternoon teams (where the score could not). Before deployment, neither group had access to the tool. Let $Morning_j$ indicate that referral r was assigned to a morning screening team j, and let $Post_t$ indicate that the referral occurred after the tool deployment date. Access to the PRM is thus determined by the interaction $Morning_j \times Post_t$.

Figure 3: Decision Rates over Time



Difference-in-Differences Model. We estimate:

$$Y_{r(i)jt} = \beta_0 + \beta_1 Morning_j + \beta_2 Post_t + \beta_3 (Morning_j \times Post_t) + X_{r(i)}\gamma + \delta_j + \tau_t + \varepsilon_{r(i)jt}, \quad (1)$$

where $Y_{r(i)jt}$ is the relevant screening outcome for referral r; δ_j and τ_t are supervisor and calendar-day fixed effects; and $X_{r(i)}$ includes demographic characteristics, allegation types, reporter type, and prior involvement.

The coefficient β_3 captures the ITT effect of PRM availability. This specification addresses the concern raised by reviewers that earlier models omitted main effects for *Morning* and *Post*, rendering interaction terms difficult to interpret.

There could be systematic differences between morning and afternoon teams, which will invalidate the identification assumptions. Table 2 presents the balance between morning and afternoon teams. [9] In terms of overall risk, the sessions are on par, but screening rates are higher in the morning. Generally,

[9] This is restricted to the post-deployment period, since as mentioned we do not have the appropriate labels for the pre-deployment period. Nevertheless, there is no reason to believe these differences were affected by the deployment since community behavior is independent.

there is a fair balance across covariates, but there are some statistically significant differences, such as schools and law enforcement reports showing up more often in the afternoon (likely meaning they call in more often in the morning), as opposed to medical staff which appear more often in the morning. Neglect and substance abuse referrals are present more frequently in the afternoon (thus likely being called in more in the morning), and physical and sexual abuse referrals are more common in the morning (i.e. mostly reported in the prior evening). Afternoon teams see slightly more referrals identified as including black and native american children.

Table 2: Balance Table for Morning and Afternoon Screening Sessions

Screening Session	Afternoon			Morning			
Variable	N	Mean	SD	N	Mean	SD	Test
Screen In	6379	0.37	0.48	5864	0.48	0.5	F= 158.693***
HRA	6379	0.17	0.38	5864	0.14	0.34	F= 28.467***
Risk Score	6379	11	6.5	5864	11	6.2	F= 1.269
Risk Category	6379			5864			X2= 40.515***
... High	1094	17%		1071	18%		
... Low	1333	21%		1381	24%		
... Medium	889	14%		927	16%		
... Very High	1144	18%		927	16%		
... Very Low	1919	30%		1558	27%		
White	6379	0.47	0.5	5864	0.51	0.5	F= 24.27***
Black	6379	0.21	0.41	5864	0.19	0.39	F= 5.396**
Hispanic	6379	0.22	0.42	5864	0.24	0.43	F= 6.357**
Indian or Native American	6379	0.0094	0.097	5864	0.014	0.12	F= 6.362**
Asian	6379	0.023	0.15	5864	0.014	0.12	F= 12.812***
Hawaiian or Pacific Islander	6379	0.0077	0.087	5864	0.0084	0.091	F= 0.175
Race Missing	6379	0.31	0.46	5864	0.29	0.45	F= 8.842***
Alleged Injurious Environment	6379	0.13	0.33	5864	0.12	0.32	F= 1.615
Alleged Physical Abuse	6379	0.039	0.19	5864	0.036	0.19	F= 1.161
Alleged Emotional Abuse	6379	0.056	0.23	5864	0.042	0.2	F= 11.654***
Alleged Substance (Parent)	6379	0.21	0.41	5864	0.19	0.39	F= 6.644***
Alleged Sexual Abuse	6379	0.0096	0.097	5864	0.0087	0.093	F= 0.252
Mandated Reporter	6379	0.8	0.4	5864	0.8	0.4	F= 0.103
Reporter: Community	6379	0.018	0.13	5864	0.018	0.13	F= 0.004
Reporter: Counselor	6379	0.019	0.14	5864	0.023	0.15	F= 2.072
Reporter: CYF	6379	0.4	0.49	5864	0.42	0.49	F= 4.736**
Reporter: Family	6379	0.08	0.27	5864	0.096	0.29	F= 10.474***
Reporter: Law Enforcement	6379	0.013	0.12	5864	0.011	0.1	F= 1.906
Reporter: Medical Staff	6379	0.18	0.38	5864	0.15	0.36	F= 14.706***
Reporter: Other	6379	0.054	0.23	5864	0.048	0.21	F= 2.375
Reporter: School	6379	0.19	0.39	5864	0.2	0.4	F= 1.21
Self Report (Perp)	6379	0.008	0.089	5864	0.008	0.089	F= 0
Self Report (Victim)	6379	0.0038	0.061	5864	0.00068	0.026	F= 12.715***
Anonymous Report	6379	0.0034	0.059	5864	0.00051	0.023	F= 12.947***

Note: Statistical significance markers: * p<0.1; ** p<0.05; *** p<0.01. This table shows the balance between morning screening sessions and afternoon screening sessions in the post-deployment period. N=12243.

Tool Use and Instrumental Variables. Actual use of the tool (denoted *Acknowledged*) is non-random: facilitators were encouraged, but not required, to consult the score. We therefore treat $Morning \times Post$ as an instrument for actual tool use and present two-stage least squares estimates as a robustness check:

$$Acknowledged_{r(i)jt} = \pi_0 + \pi_1(Morning_j \times Post_t) + X_{r(i)}\rho + \delta_j + \tau_t + u_{r(i)jt},$$

$$Y_{r(i)jt} = \alpha_0 + \alpha_1 \widehat{Acknowledged}_{r(i)jt} + X_{r(i)}\theta + \delta_j + \tau_t + \eta_{r(i)jt}.$$

Because the exclusion restriction may be imperfect—e.g., morning teams may differ systematically—we interpret IV results cautiously.

Risk-Heterogeneous Effects. To assess whether PRM availability affects referrals differently across risk categories, we estimate:

$$Y_{r(i)jt} = \beta_0 + \beta_1(Morning_j \times Post_t) + \sum_k \beta_k RiskCat_{k,r(i)} + \sum_k \beta_k^{Int}(Morning_j \times Post_t \times RiskCat_{k,r(i)}) + X_{r(i)}\gamma + \tag{2}$$

where $RiskCat_k$ indexes the PRM's five risk categories.

Functional Form. Binary outcomes (screen-in, HRA) are modeled using logistic regression in robustness checks; the main multinomial outcome (screen-out, FAR, HRA) is modeled using multinomial logit. Linear probability models are retained for interpretability but not emphasized.

Assumptions and Limitations. Our design assumes parallel trends: absent the tool, differences between morning and afternoon screening decisions would evolve similarly over time. Because we lack precise morning/afternoon labels in the pre-deployment period, we include supervisor fixed effects and rich time controls to mitigate bias. We interpret effects as local to the morning-versus-afternoon contrast and to the early months of deployment.

4 Findings

4.0.1 Tool Validity

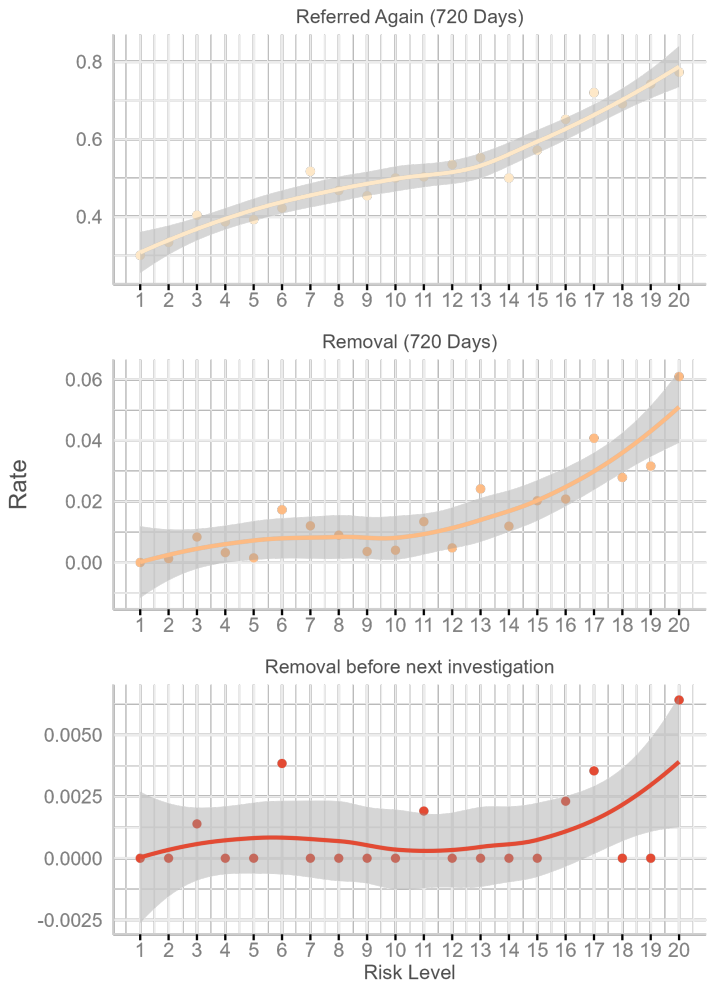
The tool must be predictive of relevant outcomes for it to be useful for CPS workers. The algorithm was trained on statewide historical data to predict a child's foster care placement (removal from home due to maltreatment) within two years of the referral. We examine the correlation between the score and removals with a follow up period of 720 days (almost 2 years) after the initial referral.

Figure 4 shows the calibration of the algorithm within the time period available in our data. All outcome trends show a correlation with risk levels. Re-referrals are predicated on community behavior and show higher-risk children

tend to be re-referred in higher quantities. This is in line with similar recent findings leveraging hospital records and validating that the tool is predictive of injury in children (Grimon and Mills 2025).

Furthermore, when we consider the predictive power of the PRM algorithm compared to a baseline of the actual screening decisions made by staff, we see that real decisions fail to predict a removal outcome over a 2 year period. A screening policy based on the risk score provided by the algorithm would have achieved a better performance, as Figure 13 shows.

Figure 4: Calibration by Risk Level for Different Outcomes



We take a deeper look at re-referrals of children to the hotline, especially those that end up getting screened in, to examine whether the tool predictions

comport with community behavior.[10] Table 3 shows regression estimates for tool use, with community re-referrals and additional screen-in within a 720-day window. We find that using the tool has lead to decreased re-referrals within the longer time window.

Table 3: Tool Use Slightly Affects Re-referral by the Community

	Dependent variable:		
	Referred Again (75 Days)	Referred Again (720 Days)	Referred Again Before Next Investigation (720 Days)
	(1)	(2)	(3)
Tool Used	−0.002	−0.032***	−0.008
	(0.005)	(0.006)	(0.006)
Constant	0.306***	0.486***	0.481***
	(0.004)	(0.005)	(0.005)
Observations	40,213	40,213	40,213
R^2	0.00000	0.001	0.00005
Adjusted R^2	−0.00002	0.001	0.00002
Residual Std. Error (df = 40211)	0.460	0.498	0.499
F Statistic (df = 1; 40211)	0.150	32.815***	1.985
Note:	*p<0.1; **p<0.05; ***p<0.01		

4.0.2 Community Outcomes

We investigate the outcomes of the introduction of the tool on the community by and large. In general, introducing the tool did not vary the number of investigated referrals within the study period, however cases in which workers consulted the tool show resulted in higher investigation rates (see Table 4). The share of children re-referred within two years increases from 51% to 56%. The principal trade-off is a small but significant rise in removals from the home: within 75 days removals increase from 0.13% to 0.36%, and within 720 days from 1.7% to 2.1% - still rare events, but representing relative increases. It did so, however, in better alignment between risk levels and removal outcomes (see Figure 5), with riskier referrals being more likely to be eventually result in removal (despite results being statistically insignificant, due to sample restrictions).

[10] Other margins, such as founded allegations or removals may be more reliable, but are too rare to yield discernible results within the study period. It is unlikely that tool predictions themselves, as opposed to objective risk levels, lead to increased re-referrals by the community within such a time span.

Figure 5: Removal from Home within Two Years

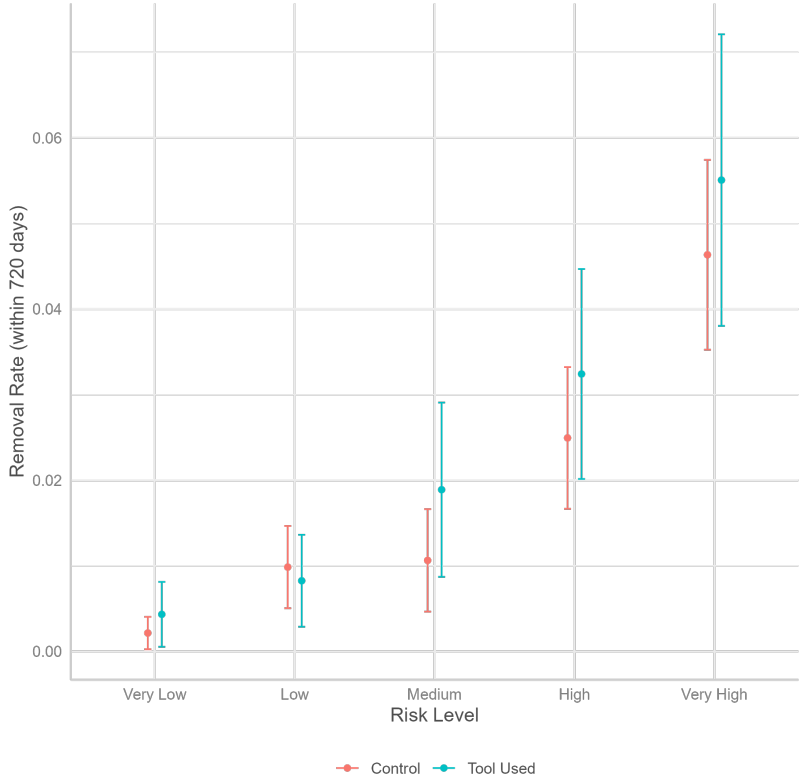


Table 4: Differences in Outcomes based on Tool Use

Tool Use	Not Used			Used			
Variable	N	Mean	SD	N	Mean	SD	Test
Screen In	7819	0.39	0.49	4424	0.46	0.5	F= 51.398***
Referred Again (720 Days)	7819	0.51	0.5	4424	0.56	0.5	F= 32.256***
Removed (75 Days)	7819	0.0013	0.036	4424	0.0036	0.06	F= 7.289***
Removed (720 Days)	7819	0.017	0.13	4424	0.021	0.14	F= 2.31

Statistical significance markers: * p<0.1; ** p<0.05; *** p<0.01.

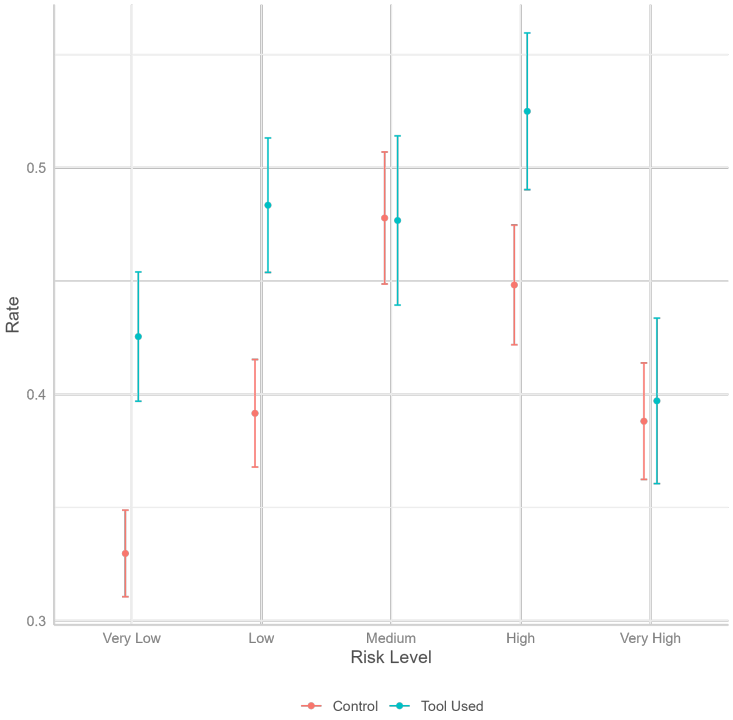
4.1 Main Effects

4.1.1 Effects on Screening In

We find that overall, using the tool slightly decreases the screening rate, depending on the specification (see Table 5), but this effect is largely statistically insignificant. The effect seems to be driven mainly by increasing screening rates for very low and low risk referrals (see Figure 6). Since interaction terms are not

easily understood, we report marginal effects (see Figure 7, which present the effect of using the tool on screening outcomes, interacted with risk categories. Under this specification, the tool yields a negative effect on low risk and medium risk levels.

Figure 6: Screen In Rates by Risk Levels



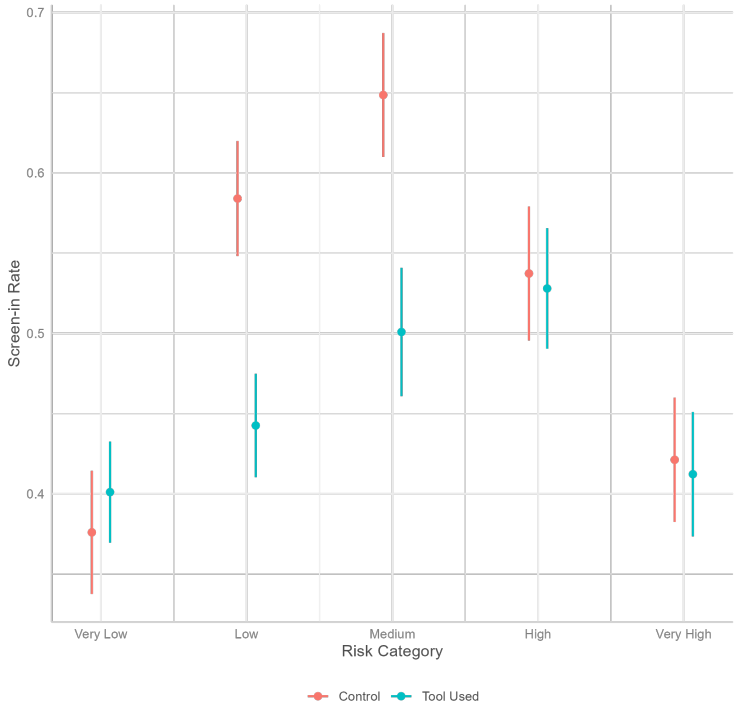
Note: This figure presents differences in means for the use of the tool across risk categories.

4.1.2 Effects on Investigation track

We find slight positive effects overall on the formal HRA track (High Risk Assessment) (see Figure 8 and Table 6). Zooming in, however, the tool seems to decreasing the likelihood of being assigned to the HRA track for the lower risk categories, with some differences by specification.

With regards to the informal FAR track (Family Assessment Response), referrals were less likely to be sent to FAR overall, but the substantial effect was increased rates for lower risk categories (see Figure 9 and Table 6).

Figure 7: Marginal Effects on Screening Rate with Risk Categories



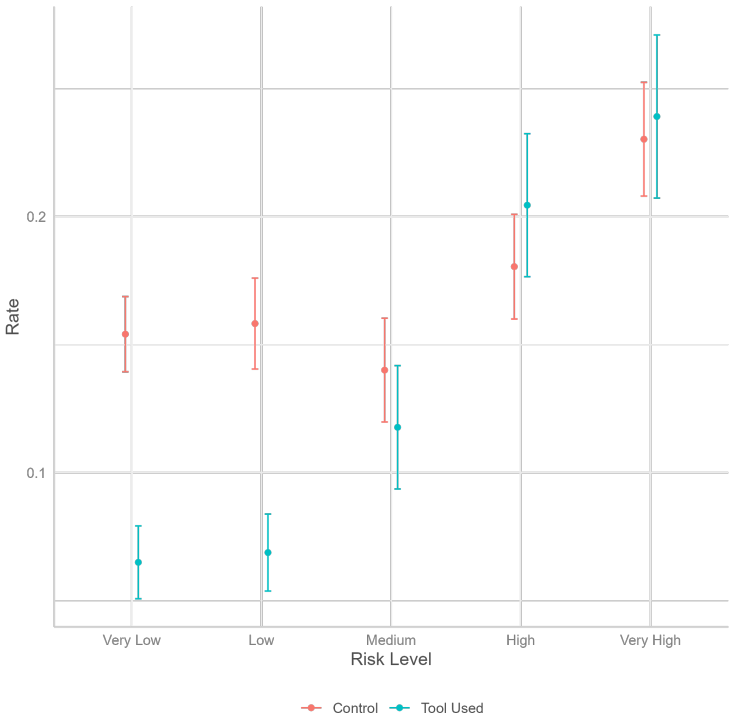
Note: This figure presents the marginal effects of Tool Use on Screen-out, but with an interaction term with risk categories, as specified in model (3). Marginal effects are used to visualize the complex results of interaction terms in a linear regression table.

Table 5: Effect of Tool Use on Screen-in: Linear Regression Model

	Dependent variable:				
	Screen-in				
	(1)	(2)	(3)	(4)	(5)
Tool Use	−0.063***	−0.009	−0.028	−0.012	−0.018
	(0.012)	(0.029)	(0.029)	(0.028)	(0.029)
Low Risk		0.047*	0.043	0.031	0.007
		(0.028)	(0.029)	(0.028)	(0.030)
Medium Risk		0.111***	0.083***	0.085***	0.060**
		(0.029)	(0.029)	(0.028)	(0.029)
High Risk		−0.116***	−0.110***	−0.113***	−0.112***
		(0.029)	(0.030)	(0.029)	(0.029)
Very High Risk		−0.161***	−0.149***	−0.132***	−0.122***
		(0.029)	(0.029)	(0.029)	(0.029)
Tool Use X Low Risk		−0.132***	−0.110***	−0.105***	−0.073*
		(0.038)	(0.039)	(0.038)	(0.039)
Tool Use X Medium Risk		−0.138***	−0.090**	−0.101**	−0.068*
		(0.040)	(0.040)	(0.039)	(0.040)
Tool Use X High Risk		0.0003	0.017	−0.031	−0.024
		(0.040)	(0.040)	(0.039)	(0.039)
Tool Use X Very High Risk		0.034	0.028	0.014	0.001
		(0.038)	(0.038)	(0.038)	(0.037)
Constant	0.515***	0.537***	0.464***	0.609***	0.754***
	(0.009)	(0.022)	(0.045)	(0.054)	(0.064)
Supervisor FE	No	No	Yes	Yes	Yes
Time FE	No	No	No	No	Yes
Controls	No	No	No	Yes	Yes
Observations	6,845	6,845	6,845	6,845	6,845
R^2	0.004	0.028	0.073	0.124	0.196
Adjusted R^2	0.004	0.027	0.066	0.116	0.177
Residual Std. Error	0.499	0.493	0.483	0.470	0.453
	(df = 6843)	(df = 6835)	(df = 6794)	(df = 6778)	(df = 6691)
F Statistic	27.489***	22.016***	10.658***	14.566***	10.633***
	(1; 6843)	(9; 6835)	(50; 6794)	(66; 6778)	(153; 6691)

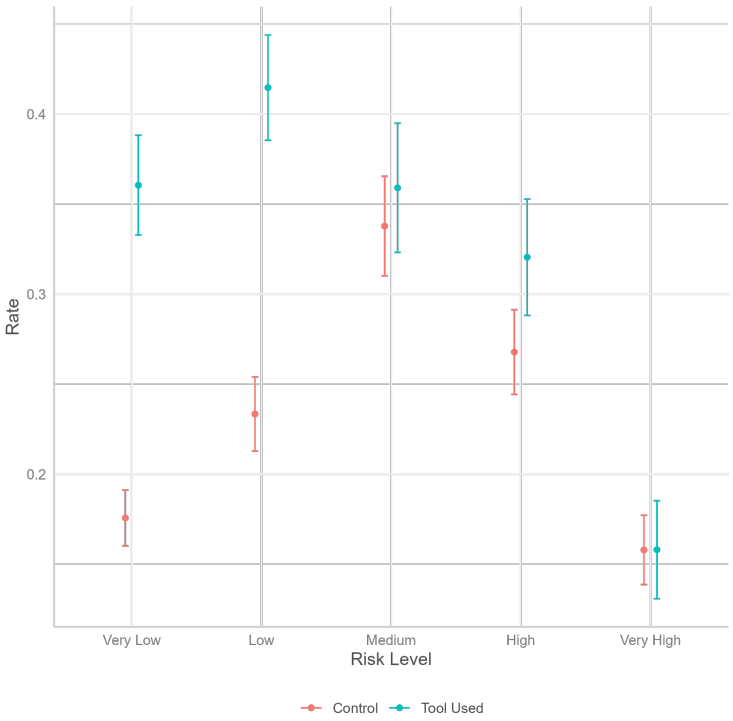
Note: Statistical significance markers: * p<0.1; ** p<0.05; *** p<0.01. Control variables include race, open case, reason for referral, type of reporter, and whether the reporter is mandated. Standard errors are clustered on both the supervisor and time (days before and after deployment).

Figure 8: HRA Rates by Risk Levels



Note: This figure presents differences in means for the use of the tool over risk levels.

Figure 9: FAR Rates by Risk Levels



Note: This figure presents differences in means for the use of the tool over risk levels.

Table 6: Effect of Tool Use on HRA and FAR: Linear Regression Models

	Dependent variable: HRA					*Dependent variable: FAR*				
	1	2	3	4	5	6	7	8	9	10
Tool Use	−0.010	0.052**	0.036	0.046**	0.062***	−0.054***	−0.061**	−0.064**	−0.058**	−0.080***
	(0.008)	(0.022)	(0.022)	(0.022)	(0.021)	(0.012)	(0.028)	(0.027)	(0.028)	(0.028)
Low Risk		0.061***	0.020	0.006	−0.051***		−0.015	0.023	0.025	0.058**
		(0.022)	(0.019)	(0.018)	(0.019)		(0.028)	(0.027)	(0.028)	(0.029)
Medium Risk		−0.081***	−0.073***	−0.070***	−0.056***		0.192***	0.156***	0.155***	0.117***
		(0.018)	(0.019)	(0.019)	(0.019)		(0.029)	(0.028)	(0.028)	(0.029)
High Risk		0.031	0.022	0.027	0.041**		−0.147***	−0.132***	−0.140***	−0.153***
		(0.022)	(0.021)	(0.021)	(0.021)		(0.027)	(0.027)	(0.027)	(0.027)
Very High Risk		−0.091***	−0.100***	−0.077***	−0.084***		−0.071**	−0.049*	−0.055*	−0.038
		(0.018)	(0.018)	(0.018)	(0.018)		(0.028)	(0.028)	(0.028)	(0.028)
Tool Use X Low Risk		−0.214***	−0.162***	−0.136***	−0.076***		0.082**	0.052	0.031	0.003
		(0.028)	(0.026)	(0.025)	(0.025)		(0.037)	(0.037)	(0.037)	(0.038)
Tool Use X Medium Risk		−0.002	−0.008	−0.005	−0.006		−0.136***	−0.082**	−0.097**	−0.062
		(0.028)	(0.028)	(0.028)	(0.027)		(0.040)	(0.039)	(0.039)	(0.039)
Tool Use X High Risk		0.024	0.031	−0.009	−0.005		−0.024	−0.014	−0.022	−0.019
		(0.032)	(0.031)	(0.031)	(0.030)		(0.036)	(0.036)	(0.036)	(0.036)
Tool Use X Very High Risk		−0.048*	−0.032	−0.052**	−0.056**		0.082**	0.060	0.065*	0.058
		(0.025)	(0.025)	(0.025)	(0.024)		(0.037)	(0.036)	(0.036)	(0.036)
Constant	0.135***	0.149***	0.147***	0.250***	0.268***	0.380***	0.388***	0.317***	0.359***	0.486***
	(0.006)	(0.015)	(0.029)	(0.042)	(0.048)	(0.009)	(0.021)	(0.041)	(0.053)	(0.063)
Supervisor FE	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Time FE	No	No	No	No	Yes	No	No	No	No	Yes
Controls	No	No	No	Yes	Yes	No	No	No	Yes	Yes
Observations			6,845					6,845		
R^2	0.0002	0.046	0.102	0.170	0.277	0.003	0.044	0.102	0.129	0.198
Adjusted R^2	0.0001	0.045	0.096	0.162	0.261	0.003	0.043	0.095	0.120	0.180
Residual Std. Error	0.336 (df = 6843)	0.329 (df = 6835)	0.320 (df = 6794)	0.308 (df = 6778)	0.289 (df = 6691)	0.476 (df = 6843)	0.467 (df = 6835)	0.454 (df = 6794)	0.448 (df = 6778)	0.432 (df = 6691)
F Statistic	1.469 (1; 6843)	36.537*** (9; 6835)	15.518*** (50; 6794)	21.030*** (66; 6778)	16.779*** (153; 6691)	21.456*** (1; 6843)	34.864*** (9; 6835)	15.402*** (50; 6794)	15.207*** (66; 6778)	10.829*** (153; 6691)

Note: Statistical significance markers: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Control variables include race, open case, reason for referral, type of reporter, and whether the reporter is mandated. Standard errors are clustered on both the supervisor and time (days before and after deployment).

4.2 Effects of Recommendations

As Albright (2024) has suggested, the design of algorithmic recommendations can have an independent effect on margins of interest. That is, the design of algorithmic aids can reflect different policies, such as what risk threshold constitutes a positive or negative recommendation (depending on the outcome). Other research in human-computer interaction (HCI) shed light on design choices of user interface (UI) such as the use of colors or dashboards as potentially influential in swaying decisions (Q. Yang et al. 2020).

Recall that the interface that screening supervisors use has an indication for a risk category with an indicative color, as well as a numeric score (see Figure 2). Although the Division did not, intentionally, specify what is the recommended course of action was for each risk category, it is plausible that screeners interpret the shift from one category to another as signifying a recommendation. For example, a high or very high risk designation could be seen in practice as a recommendation to screen in a referral; or a very high risk designation can be interpreted as an HRA recommendation.

We leverage the specific UI choices, along with the raw model probabilities that were available to us post hoc as a running variable. We thus employ a Regression Discontinuity Design (RDD), a quasi-experimental method that leverages a discontinuity in the assignment of a treatment or policy, based on a predetermined threshold of an assignment variable, to compare outcomes just above and below the threshold. The assumption is that units near the cutoff are similar in all respects except for treatment status. In our case, we assume that referrals scored around 14 and those scored around 15 are very similar, yet some are assigned a medium risk category (associated with a yellow color) and others assigned to be high risk (associated with orange).

We first map the LASSO probability thresholds to the categorical changes (where 0.25 "VL→L", 0.40 "L→M", 0.55 "M→H", 0.73 "H→VH"). We then normalize the distances from the cutoff points for ease of interpretation. We use a bandwidth of 0.10 from both sides of a cutoff point. We test other bandwidths for robustness but results are largely consistent, substantially smaller bandwidths lack enough data points.

Figures 10 and 11 present local linear regression discontinuity plots estimating the probability of a case being screened in as a function of the risk score distance from four decision thresholds. The vertical dashed line marks the threshold, with linear fits estimated separately on either side. In most cut-off points, we do not find any effect. One exception is the shift from Very Low to Low risk, which unexpectedly leads to lower screening rates. Second, when risk probability jumps from High to Very High (Orange to Red), there are both higher screen-in and HRA rates. We essentially find modest effects of recommendations.

Figure 10: Regression Discontinuity: Screen In

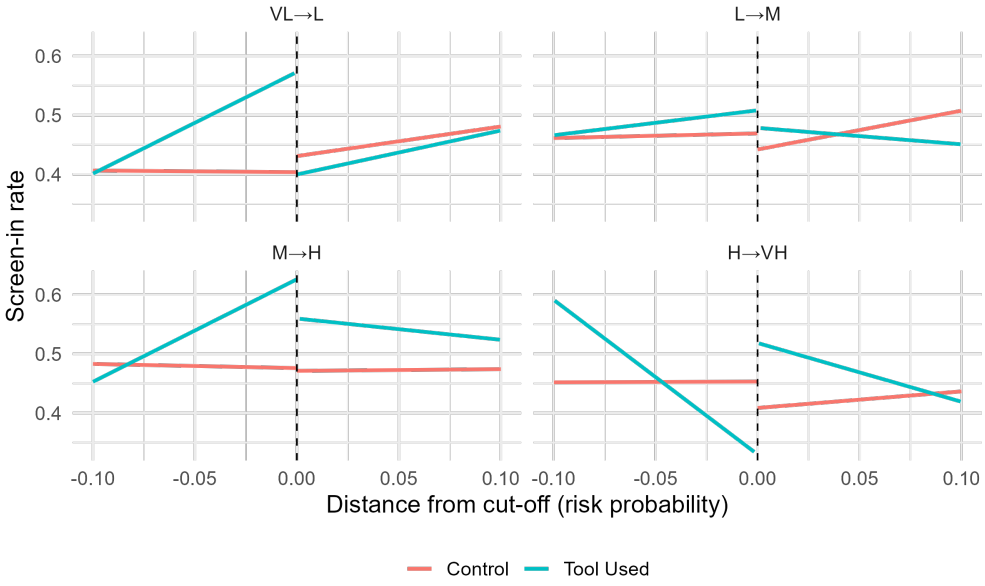
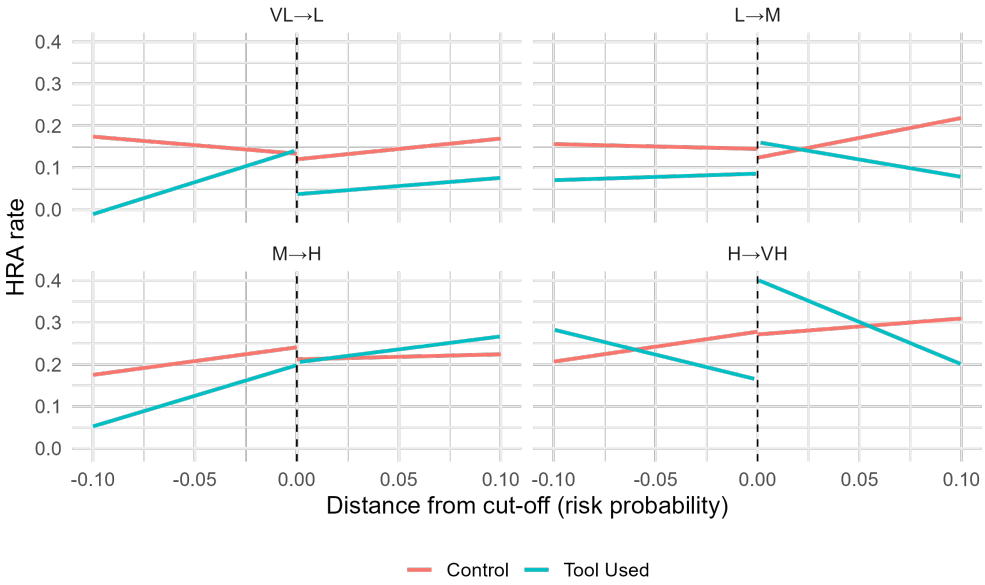


Figure 11: Regression Discontinuity: HRA



5 Discussion

The main findings of this study are that first, the tool adopted by the county is predictive of relevant outcomes, and a policy based on its predictions would have performed better on average than human workers. Second, the tool did not reduce investigated referrals overall, but reduced screening-in rates for low-risk referrals, without increasing screening probability for higher-risk referrals. Regarding the investigation track, we find effects for the lower risk categories - such referrals were less likely to be assigned to the formal investigation track and more likely to be assigned to the informal and less intrusive track. Third, we find that the tool is associated with fewer investigations and fewer subsequent referrals despite slightly higher initial risk ratings, paired with a modest uptick in removals that may reflect sharper targeting of the most serious cases. Finally, we find only modest independent effects of shifts in categories of risk, where moving from high to very high (orange to red) leads to higher screening in and formal investigations.

PRMs and other decision aids are an important prospect for the CPS and similar domains, and will thus benefit from evaluation and assessment of implementation in the field. This study contributes to this endeavor. It also highlights that algorithms have the potential to serve as internal governance mechanisms, supporting administrative agencies in improving the accuracy and consistency of decisions but without the costs of external regulation.

Considering the overall effects, the tool modestly shifts screening patterns in ways that increase concordance between predicted and assigned investigation tracks. It improves the risk-to-outcome fit, mainly by decreasing false positives (screened in children that are not in risk), and thus reducing the bureaucratic burden on the agency.

In general, the tool studied in this article exemplifies a potential avenue for improvement in bureaucratic street-level decision-making in the public sector, especially in the social services domain. It seems to allow more accurate and less noisy decisions, and removes some of the inaccuracies involved in distributed decision-making. It improves overall decisions, without too burdensome interventions in the day-to-day operations of the agency, that more training or review techniques entail. Essentially, it functions as a sort of contemporaneous peer review or quality assurance and yields some of the benefits of this approach (Ho and Sherman 2017), without the direct costs involved. This highlights its potential as an internal governance mechanism. The challenges that agencies face with the quality of decisions are a major hindrance in the administrative state, and algorithms should be considered in this context.

However, the results tell a cautionary tale about deployment and implementation. Despite showing promise in risk prediction, the tool yields modest effects and those could be subsiding over time, as screeners have either accustomed and internalized its predictions, or have grown to disregard it, as organizational studies have shown before (Lipsky 2010; Kellogg, Valentine, and Christin 2020). This suggests that the novelty of having an online risk prediction, which may have led supervisors to question their tendencies in cases where the prediction did not

comport with their own judgment, is not self-enforcing. This finding raises two important issues. First, it is important to evaluate interventions over time and in the field. Second, implementation in administrative decision-making is an ongoing effort, and agencies may have to consider the options available to them to ensure such tools do not become obsolete. This highlights a problem that may be dubbed an the street-level bureaucracy adherence paradox or dilemma, where interventions are either superfluous and do not produce effects; or have to be more forcefully enforced by integration to a policy or rule.

Agencies will be faced with the dilemma of whether to instigate a stronger mechanism to induce reliance, such as devising a policy or creating behavioral or institutional design mechanisms. So long as the agency refrains from making screening decisions solely based on the tool, current legal standards are unlikely to prevent a hybrid decision-making model (Engstrom and D. E. Ho 2020; Coglianese and Lehr 2017), with increased weight for automated risk prediction. Additionally, the superior predictive performance of the tool compared to the average screener (See section 4.0.2 *supra*), suggests a simple policy may better perform; however, since the tool is not privy to the current allegation, it lacks a legal criteria to act upon. Thus policymakers could consider adopting stronger recommendations based on raw predictions.

Agencies could consider instigating override protocols, such that for instance every referral classified as high or very high risk is presumed to be screened-in, unless the screening team decides to override and provides its reasoning. This may allow for more friction in the process and induce reliance, yet mitigating concerns about taking humans out of the loop and creating accountability and algorithmic biases risks. Moreover, since at the study county every referral is discussed at a screening group meeting, there is less concern that automation bias will creep in and effectively make the PRM the primary decision-maker. However, that is a unique situation which varies from county to county (and between states, see *Allegheny Methodological Report* 2019), which highlights the need to tailor implementation to specific organizational and legal circumstances.

The study indicates the need to develop the research on human-machine interaction as part of public administration and administrative law more generally (Engstrom and Haim 2023; Haim 2024). The legal criteria and standards for CPS screening are complex and multi-faceted, while the PRM tool merely predicts one decision margin (i.e. removal from home within 24 months). While it is true that screeners have many factors to consider when making a screening decision, they do not have more than 15 minutes on average to discuss a referral, and realistically cannot invest many resources in a preliminary triage decision. Referrals that are screened-in, on either assignment track, will be considered in more detail and the caseworker in charge will execute fact-finding actions to better assess the risk of imminent harm. When decision-makers are constrained, they tend to rely on coping mechanisms (Bednar 2023). This may be a heuristic or a salient feature of a referral, such as the race of the family or their address. Since the tool appears to be more predictive than the average screener, it essentially nudges screening decisions to be more in line with the risk distribution. The professional experience of screeners and other variables may mediate their

use of the tool, and it is possible that less experienced screeners rely more on the tool (although, that does not mean that experienced screeners are better at predicting risk), yet the scope of this paper does not allow us to examine these questions and we leave this to future work. Moreover, risk is the main legal standard guiding the screening process, and since the tool is trained to predict a downstream decision that is based on a more elaborate factual and legal basis, it may implicitly incorporate more complex standards. Nevertheless, screening teams retain the option to act on a legal basis that they believe to be relevant, even if the score does not comport to it.

Screeners retain the discretion to make a decision based on all available data at their disposal. Since the tool does not take into account the present referral, there may be cases where the screeners think that the referral merits further investigation (i.e. screen-in) despite having a low-risk score (the inverse situation is also possible). These are the cases where discretion, experience, and discussion among the team are best utilized. Yet when new information does not substantially change the picture, the tool seems to yield better results, and not relying on its recommendation is sub-optimal. Additionally, in line with Albright (2024) we find that recommendations, or in our case graphic design choices, may have effects independent of the underlying risk score. This notion should be taken into account when designing and deploying algorithmic systems in the field. Related to this issue, is that utilization of the tool was not perfect. Contrary to what could be believed based on conversations with screeners, uptake seemed to increase over time (see Figure 15 in the appendix), suggesting that users have gotten accommodated and more trusting of the tool, yet the effects subsided suggesting effective desertion or internalization. Alternatively, it is possible that predictive accuracy has decreased over time, due to contextual changes (such as lagged community behavior changes emanating from the Covid-19 pandemic). This highlights again the need for continuous evaluation.

Notably, this study is not free of limitations. First of all, it does not adhere to the gold standard of evaluation methodology and cannot ensure randomized and controlled assignment of treatment. Instead, it leverages quasi-experimental designs and observational methods to recover causal effects. While this approach has gained traction, it still has limitations. Furthermore, it suffers from the follies of administrative and real-world data, which tends to be messier than in controlled environments. One important problem that arose was the untrustworthiness of timestamps, which prevented us from differentiating morning and afternoon screening sessions throughout the period before and after deployment, and also made simpler designs - such as a difference-in-difference estimator - impossible to implement. Another potential limitation and an issue agencies will have to contend with, is that screeners may be updating implicitly their assessment of risk when using the PRM tool. Nevertheless, we do not see this as a serious threat to identification within the study period, since it is unlikely that screeners were able to extrapolate based on a sparse interaction with the tool in a relatively short period of time.

6 Conclusion

In conclusion, this study highlights the potential of data-driven tools, such as predictive risk models, to improve decision-making processes in administrative settings, particularly within child protection services. The findings demonstrate that the implementation of such tools can lead to a reduction in inconsistent screening decisions and influence the assignment of investigation resources. Moreover, it highlights the importance of evaluating algorithmic tools in the field, rather than assessing them in the lab, irregardless of implementation. However, it is important to address challenges related to tool adoption and ensure that frontline workers perceive them as supportive rather than substitutive. Future research should continue to explore the implementation and effectiveness of data-driven tools in administrative decision-making contexts.

References

Albright, Alex (2024). "The Hidden Effects of Algorithmic Recommendations". In.

Allegheny Methodological Report (Apr. 2019). Allegheny County Department of Human Services. URL: https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-2.pdf (visited on 03/19/2021).

Ames, David et al. (2020). "Due Process and Mass Adjudication: Crisis and Reform". In: *Stanford Law Review* 72, p. 1.

Angelova, Victoria, Will Dobbie, and Crystal S Yang (2023). "Algorithmic Recommendations and Human Discretion".

De-Arteaga, Maria, Riccardo Fogliato, and Alexandra Chouldechova (2020). "A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12. DOI: 10.1145/3313831.3376638. arXiv: 2002.08035. URL: http://arxiv.org/abs/2002.08035 (visited on 11/12/2020).

Baird, Christopher and Dennis Wagner (Nov. 1, 2000). "The relative validity of actuarial- and consensus-based risk assessment systems". In: *Children and Youth Services Review* 22.11, pp. 839–871. ISSN: 0190-7409. DOI: 10.1016/S0190-7409(00)00122-5. URL: https://www.sciencedirect.com/science/article/pii/S0190740900001225 (visited on 10/17/2022).

Bednar, Nicholas (2023). "The Public Administration of Justice". In: *Cardozo Law Review* __, __. ISSN: 1556-5068. DOI: 10.2139/ssrn.4189963. URL: https://www.ssrn.com/abstract=4189963 (visited on 09/30/2022).

Benbenishty, Rami et al. (2015). "Decision making in child protection: An international comparative study on maltreatment substantiation, risk assessment and interventions recommendations, and the role of professionals' child welfare attitudes | Elsevier Enhanced Reader". In: *Child Abuse & Neglect* 49, p. 63. DOI: 10.1016/j.chiabu.2015.03.015. URL: https://reader.elsevier.com/reader/sd/pii/S014521341500112X?token=E93BDF0E2BA1C7AE2E3DF3958B0D021777357F originRegion=us-east-1&originCreation=20210422234404 (visited on 04/23/2021).

Brown, Anna et al. (2019). "Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. the 2019 CHI Conference. Glasgow, Scotland Uk: ACM Press, pp. 1–12. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300271. URL: http://dl.acm.org/citation.cfm?doid=3290605.3300271 (visited on 10/20/2019).

Cheng, Hao-Fei et al. (Apr. 27, 2022). "How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions". In: *CHI Conference on Human Factors in Computing Systems*. CHI '22: CHI Conference on Human Factors in Computing Systems. New Orleans LA USA: ACM, pp. 1–22. ISBN: 978-1-

4503-9157-3. DOI: 10.1145/3491102.3501831. URL: https://dl.acm.org/doi/10.1145/3491102.3501831 (visited on 07/15/2022).

Chouldechova, Alexandra et al. (2018). "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions". In: *Proceedings of Machine Learning Research* 81, p. 1.

Church, Christopher E. and Amanda J. Fairchild (2017). "In Search of a Silver Bullet: Child Welfare's Embrace of Predictive Analytics". In: *Juvenile and Family Court Journal* 68.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jfcj.12086, pp. 67–81. ISSN: 1755-6988. DOI: https://doi.org/10.1111/jfcj.12086. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jfcj.12086 (visited on 11/12/2020).

Coglianese, Cary and David Lehr (2017). "Regulating by Robot: Administrative Decision Making in the Machine-Learning Era". In: *Georgetown Law Journal* 105, p. 1147.

Cordella, Antonio and Francesco Gualdi (2024). "Algorithmic Formalization: Impacts on Administrative Processes". In: *Public Administration*. DOI: 10.1111/padm.13030. URL: https://onlinelibrary.wiley.com/doi/10.1111/padm.13030.

Cuccaro-Alamin, Stephanie et al. (2017). "Risk assessment and decision making in child protective services: Predictive risk modeling in context". In: *Children and Youth Services Review* 79, p. 291. DOI: 10.1016/j.childyouth.2017.06.027. URL: https://reader.elsevier.com/reader/sd/pii/S0190740917300452?token=116A15C729EFC2C118A259AB9F0BF43764FC7910E537547CD216884995383B5D2 (visited on 07/22/2020).

Davis, Kenneth Culp (1969). *Discretionary Justice: A Preliminary Inquiry*. LSU Press.

Drake, Brett et al. (2020). "A Practical Framework for Considering the Use of Predictive Risk Modeling in Child Welfare". In: *The Annals of the American Academy of Political and Social Science* 692.1. Publisher: SAGE Publications Inc, pp. 162–181. ISSN: 0002-7162. DOI: 10.1177/0002716220978200. URL: https://doi.org/10.1177/0002716220978200 (visited on 04/10/2021).

Eiermann, Martin (2024). "Algorithmic Risk Scoring and Welfare State Contact Among US Children". In: *Sociological Science* 11, pp. 707–742. DOI: 10.15195/v11.a26. URL: https://sociologicalscience.com/articles-v11-26-707/.

Engstrom, David Freeman and Amit Haim (2023). "Regulating Government AI and the Challenge of Sociotechnical Design". In: *Annual Review of Law and Social Science* 19.1. _eprint: https://doi.org/10.1146/annurev-lawsocsci-120522-091626, null. DOI: 10.1146/annurev-lawsocsci-120522-091626. URL: https://doi.org/10.1146/annurev-lawsocsci-120522-091626.

Engstrom, David Freeman and Daniel E. Ho (2020). "Algorithmic Accountability in the Administrative State". In: *Yale Journal on Regulation* 37, p. 800.

Engstrom, David Freeman, Daniel E. Ho, et al. (2020). *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*. URL: https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf (visited on 11/03/2020).

Eubanks, Virginia (2018). *Automating Ineqality: How High-Tech Tools Profile, Police, and Punish the Poor*.

Gillingham, Phillp and Cathy Humphreys (Dec. 1, 2010). "Child Protection Practitioners and Decision-Making Tools: Observations and Reflections from the Front Line". In: *British Journal of Social Work* 40.8, pp. 2598–2616. ISSN: 0045-3102, 1468-263X. DOI: 10.1093/bjsw/bcp155. URL: https://academic.oup.com/bjsw/article-lookup/doi/10.1093/bjsw/bcp155 (visited on 05/05/2021).

Gillis, Talia, Bryce McLaughlin, and Jann Spiess (Oct. 28, 2021). *On the Fairness of Machine-Assisted Human Decisions*. arXiv: 2110.15310[cs,econ,q-fin,stat]. URL: http://arxiv.org/abs/2110.15310 (visited on 06/13/2023).

Glaze, Kurt et al. (2022). "Artificial Intelligence for Adjudication: The Social Security Administration and AI Governance". In: *The Oxford Handbook on AI Governance*. Ed. by Justin B. Bullock et al., p. 18.

Goel, Sharad et al. (2021). "The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment". In: *Big Data Handbook*. URL: https://www.ssrn.com/abstract=3306723 (visited on 07/29/2020).

Goldhaber-Fiebert, Jeremy D and Lea Prince (2019). *Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office*, p. 97. URL: https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Impact-Evaluation-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-6.pdf.

Grimon, Marie-Pascale and Christopher Mills (2025). *Better Together? A Field Experiment on Human-Algorithm Interaction in Child Protection*. SOFI Working Papers in Labour Economics 2/2025. Swedish Institute for Social Research. URL: https://ideas.repec.org/p/hhs/sofile/2025_002.html.

Grove, William M. and Paul E. Meehl (1996). "Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy". In: *Psychology, Public Policy, and Law* 2. Place: US Publisher: American Psychological Association, pp. 293–323. ISSN: 1939-1528. DOI: 10.1037/1076-8971.2.2.293.

Haan, Irene de and Marie Connolly (Dec. 2014). "Another Pandora's box? Some pros and cons of predictive risk modeling". In: *Children and Youth Services Review* 47, pp. 86–91. ISSN: 01907409. DOI: 10.1016/j.childyouth.2014.07.016. URL: https://linkinghub.elsevier.com/retrieve/pii/S0190740914002655 (visited on 12/11/2020).

Haim, Amit (2024). "The Administrative State and Artificial Intelligence: Toward an Internal Law of Administrative Algorithms". In: *UC Irvine Law Review* 14.1, p. 103.

Hausman, David K (2021). "Reviewing Administrative Review". In: *Yale Journal on Regulation* 38, p. 1059.

Ho (2017). "Does Peer Review Work: An Experiment of Experimentalism". In: *Stanford Law Review* 69, p. 1.

Ho, David Marcus, and Gerald K Ray (Nov. 30, 2021). *Quality Assurance Systems in Agency Adjudication: Emerging Practices and Insights*. Report to the Admin. Conf. of the U.S, p. 29.

Ho and Sam Sherman (2017). "Managing Street-Level Arbitrariness: The Evidence Base for Public Sector Quality Improvement". In: *Annual Review of Law and Social Science* 13.1. _eprint: https://doi.org/10.1146/annurev-lawsocsci-110316-113608, pp. 251–272. DOI: 10.1146/annurev-lawsocsci-110316-113608. URL: https://doi.org/10.1146/annurev-lawsocsci-110316-113608 (visited on 11/11/2020).

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein (2020). *Noise*. ISBN: 978-0-316-45140-6.

Kawakami, Anna et al. (Apr. 5, 2022). "Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support". In: *arXiv:2204.02310 [cs]*. arXiv: 2204.02310. URL: http://arxiv.org/abs/2204.02310 (visited on 04/26/2022).

Keddell, Emily (Oct. 8, 2019). "Algorithmic Justice in Child Protection: Statistical Fairness, Social Justice and the Implications for Practice". In: *Social Sciences* 8.10, p. 281. ISSN: 2076-0760. DOI: 10.3390/socsci8100281. URL: https://www.mdpi.com/2076-0760/8/10/281 (visited on 01/12/2021).

Kellogg, Katherine C., Melissa A. Valentine, and Angéle Christin (Jan. 2020). "Algorithms at Work: The New Contested Terrain of Control". In: *Academy of Management Annals* 14.1, pp. 366–410. ISSN: 1941-6520, 1941-6067. DOI: 10.5465/annals.2018.0174. URL: http://journals.aom.org/doi/10.5465/annals.2018.0174 (visited on 10/31/2022).

Kleinberg, Jon et al. (Feb. 1, 2018). "Human Decisions and Machine Predictions". In: *The quarterly journal of economics* 133.1, pp. 237–293. ISSN: 0033-5533. DOI: 10.1093/qje/qjx032. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5947971/ (visited on 06/13/2023).

Lipsky, Michael (2010). *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. 2nd.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan (2024). "The Unreasonable Effectiveness of Algorithms". In: *NBER Working Paper*.

Parker, E. M. et al. (2022). "Examining the effects of the Eckerd rapid safety feedback process on the occurrence of repeat maltreatment among children involved in the child welfare system". In: *Child Abuse & Neglect* 133, p. 105856. DOI: 10.1016/j.chiabu.2022.105856.

Ramji-Nogales, Jaya, Andrew I. Schoenholtz, and Philip G. Schrag (2007). "Refugee Roulette: Disparities In Asylum Adjudication". In: *Stanford Law Review* 60, p. 295.

Redden, Joanna, Lina Dencik, and Harry Warne (Sept. 2, 2020). "Datafied child welfare services: unpacking politics, economics and power". In: *Policy Studies* 41.5, pp. 507–526. ISSN: 0144-2872, 1470-1006. DOI: 10.1080/01442872.2020.1724928. URL: https://www.tandfonline.com/doi/full/10.1080/01442872.2020.1724928 (visited on 04/14/2021).

Rittenhouse, Katherine, Emily Putnam-Hornstein, and Rhema Vaithianathan (2023). "Algorithms, Humans, and Racial Disparities in Child Protection

Systems: Evidence from the Allegheny Family Screening Tool". Unpublished manuscript. URL: https://krittenh.github.io/katherine-rittenhouse.com/Algorithms%20Humans%20Racial%20Disparities.pdf.

Russell, Jesse (Aug. 2015). "Predictive analytics and child protection: Constraints and opportunities". In: *Child Abuse & Neglect* 46, pp. 182–189. ISSN: 01452134. DOI: 10.1016/j.chiabu.2015.05.022. URL: https://linkinghub.elsevier.com/retrieve/pii/S0145213415002197 (visited on 12/11/2020).

Scherz, China (2011). "Protecting Children, Preserving Families: Moral Conflict and Actuarial Science in a Problem of Contemporary Governance". In: *PoLAR: Political and Legal Anthropology Review* 34.1. _eprint: https://anthrosource.onlinelibrary.wiley.com/d 2934.2011.01137.x, pp. 33–50. ISSN: 1555-2934. DOI: https://doi.org/10.1111/j.1555-2934.2011.01137.x. URL: https://anthrosource.onlinelibrary.wiley.com/doi/abs/10.1111/j.1555-2934.2011.01137.x (visited on 05/05/2021).

Schwartz, Ira M. et al. (Oct. 2017). "Predictive and prescriptive analytics, machine learning and child welfare risk assessment: The Broward County experience". In: *Children and Youth Services Review* 81, pp. 309–320. ISSN: 01907409. DOI: 10.1016/j.childyouth.2017.08.020. URL: https://linkinghub.elsevier.com/retrieve/pii/S0190740917303523 (visited on 07/27/2020).

Starr, Sonja B (2014). "Evidence-Based Sentencing and the Scientific Rationalization of Discrimination". In: *Stanford Law Review* 66, p. 803.

Stevenson, Megan (2018). "Assessing Risk Assessment in Action". In: *MINNESOTA LAW REVIEW*, p. 83.

Stevenson, Megan T. and Jennifer L. Doleac (Nov. 2024). "Algorithmic Risk Assessment in the Hands of Humans". en. In: *American Economic Journal: Economic Policy* 16.4, pp. 382–414. ISSN: 1945-7731, 1945-774X. DOI: 10.1257/pol.20220620. URL: https://pubs.aeaweb.org/doi/10.1257/pol.20220620 (visited on 12/10/2024).

Vaithianathan, Rhema (2019). *Implementing a Child Welfare Decision Aide in Douglas County Methodology Report*. Centre for Social Data Analytics. URL: https://csda.aut.ac.nz/__data/assets/pdf_file/0009/347715/Douglas-County-Methodology_Final_3_02_2020.pdf (visited on 04/23/2021).

Vaithianathan, Rhema et al. (Sept. 2013). "Children in the Public Benefit System at Risk of Maltreatment". In: *American Journal of Preventive Medicine* 45.3, pp. 354–359. ISSN: 07493797. DOI: 10.1016/j.amepre.2013.04.022. URL: https://linkinghub.elsevier.com/retrieve/pii/S0749379713003449 (visited on 12/11/2020).

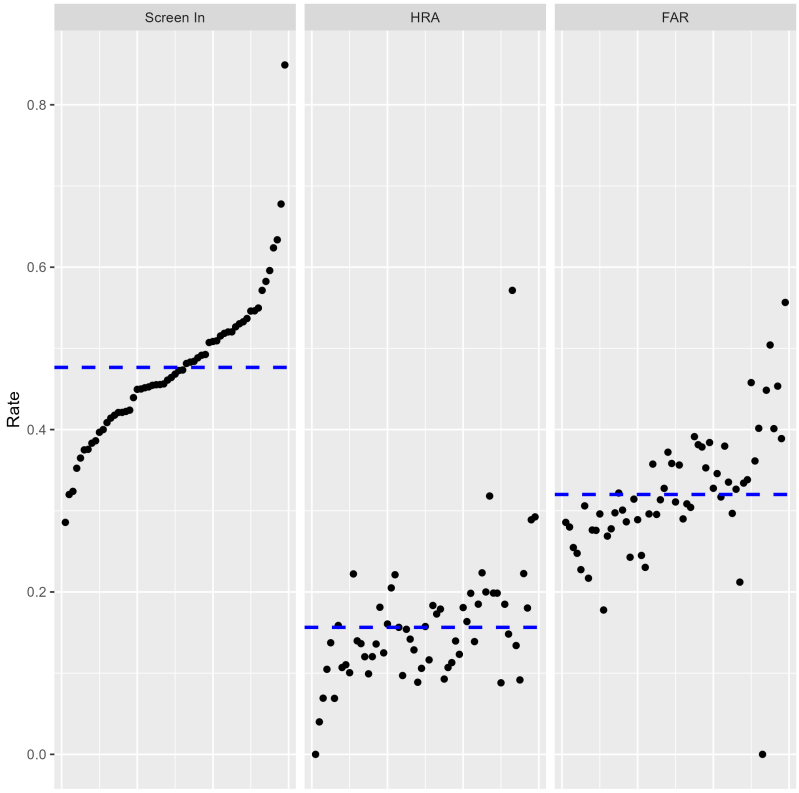
Yang, Qian et al. (Apr. 21, 2020). "Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20: CHI Conference on Human Factors in Computing Systems. Honolulu HI USA: ACM, pp. 1–13. ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.

3376301. URL: https://dl.acm.org/doi/10.1145/3313831.3376301 (visited on 10/15/2022).

A Appendix

A.1 Supervisor Tendencies

Figure 12: Variation Across Supervisors



A.2 Tool Accuracy

Figure 13: Receiver Operating Curve: Prediction of Removal within Two Years

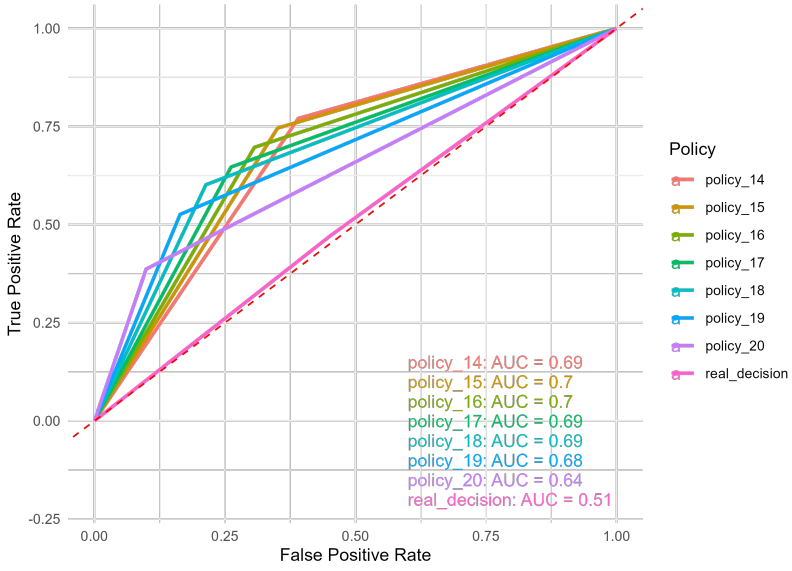


Figure 14: Receiver Operating Curve: Prediction of Removal within Two Years

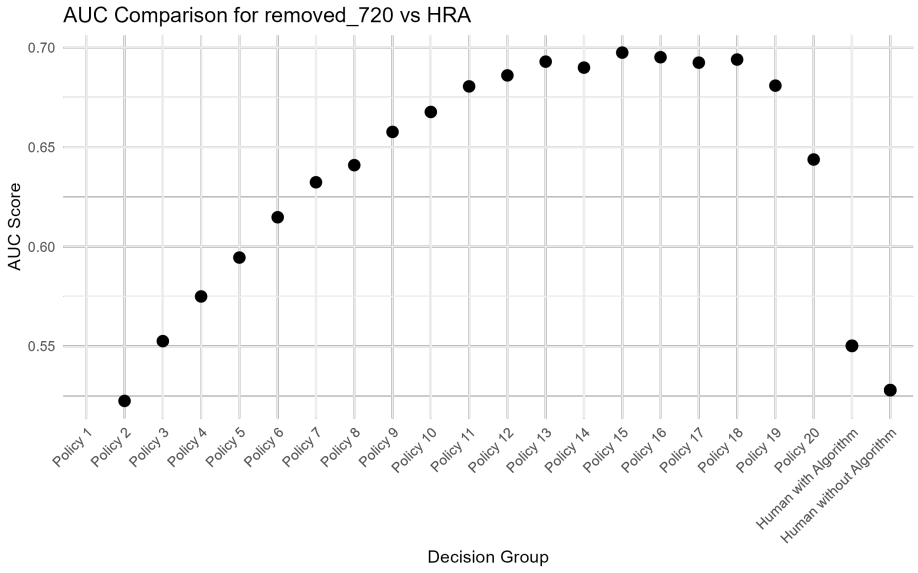
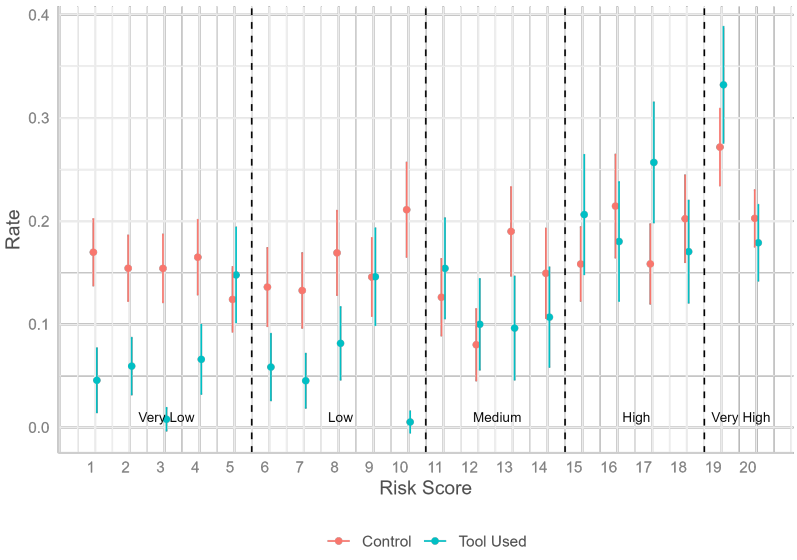


Figure 15: Tool Uptake Over Time

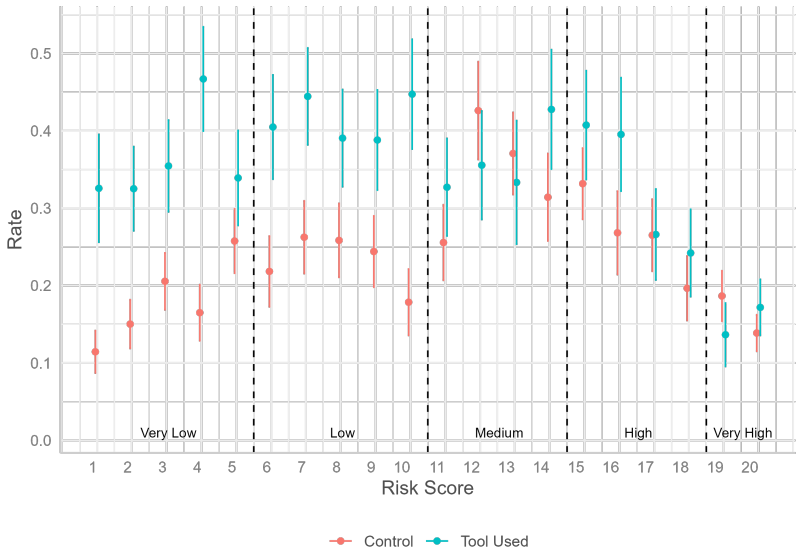
A.3 Outcomes by Risk Scores

Figure 16: HRA Rates by Risk Scores



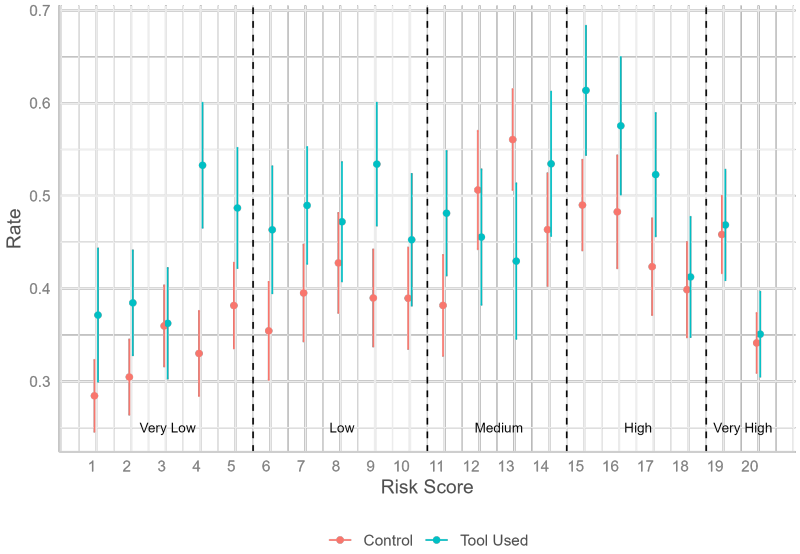
Note: This figure presents differences in means for the use of the tool over risk scores.

Figure 17: FAR Rates by Risk Scores



Note: This figure presents differences in means for the use of the tool over risk scores.

Figure 18: Screen-Out Rates by Risk Score



Note: This figure presents differences in means for the use of the tool over risk scores.

A.4 Effect on Tool Availability

Table 7: Effect of Tool Availability on Screening: Linear Regression Model

	Dependent variable:				
	Screen-in				
	(1)	(2)	(3)	(4)	(5)
Tool Available	−0.036***	−0.058***	−0.051***	−0.036*	−0.037
	(0.011)	(0.019)	(0.019)	(0.019)	(0.023)
Low Risk		0.058***	0.057***	0.056***	0.060***
		(0.010)	(0.010)	(0.009)	(0.009)
Medium Risk		0.090***	0.087***	0.080***	0.081***
		(0.011)	(0.011)	(0.011)	(0.011)
High Risk		0.022**	0.024**	0.044***	0.043***
		(0.011)	(0.011)	(0.011)	(0.011)
Very High Risk		−0.115***	−0.111***	−0.062***	−0.066***
		(0.011)	(0.011)	(0.012)	(0.012)
Tool Available X Low Risk		0.008	0.010	−0.001	−0.003
		(0.030)	(0.030)	(0.029)	(0.030)
Tool Available X Medium Risk		0.010	0.007	0.007	0.015
		(0.034)	(0.033)	(0.033)	(0.033)
Tool Available X High Risk		0.076**	0.074**	0.075**	0.085***
		(0.033)	(0.033)	(0.032)	(0.032)
Tool Available X Very High Risk		0.069**	0.060*	0.043	0.053*
		(0.031)	(0.031)	(0.031)	(0.031)
Constant	0.473***	0.458***	0.406***	0.333***	0.235***
	(0.003)	(0.006)	(0.029)	(0.037)	(0.073)
Supervisor FE	No	No	Yes	Yes	Yes
Time FE	No	No	No	No	Yes
Controls	No	No	No	Yes	Yes
Observations	22,792	22,792	22,792	22,792	22,792
R^2	0.0005	0.016	0.033	0.073	0.140
Adjusted R^2	0.0005	0.015	0.030	0.070	0.116
Residual Std. Error	0.499 (df = 22790)	0.495 (df = 22782)	0.492 (df = 22720)	0.481 (df = 22699)	0.469 (df = 22171)
F Statistic	11.381*** (df = 1; 22790)	39.904*** (df = 9; 22782)	10.902*** (df = 71; 22720)	19.559*** (df = 92; 22699)	5.830*** (df = 620; 22171)

Note: Statistical significance markers: * p<0.1; ** p<0.05; *** p<0.01. Control variables include race, open case, reason for referral, type of reporter, and whether the reporter is mandated. Standard errors are clustered on both the supervisor and time (days before and after deployment).

Table 8: Effect of Tool Availability on HRA: Linear Regression Model

	Dependent variable:				
	HRA				
	(1)	(2)	(3)	(4)	(5)
Tool Available	−0.003	−0.020**	−0.011	−0.004	0.012
	(0.007)	(0.010)	(0.010)	(0.010)	(0.014)
Low Risk		0.012**	0.012**	0.011*	0.016***
		(0.006)	(0.006)	(0.006)	(0.006)
Medium Risk		0.066***	0.063***	0.062***	0.071***
		(0.008)	(0.008)	(0.007)	(0.008)
High Risk		0.095***	0.096***	0.094***	0.101***
		(0.008)	(0.008)	(0.008)	(0.008)
Very High Risk		0.132***	0.133***	0.125***	0.132***
		(0.009)	(0.008)	(0.009)	(0.009)
Tool Available X Low Risk		0.008	0.004	0.004	0.013
		(0.017)	(0.017)	(0.016)	(0.017)
Tool Available X Medium Risk		−0.012	−0.018	−0.024	−0.018
		(0.021)	(0.022)	(0.021)	(0.021)
Tool Available X High Risk		0.067***	0.063**	0.060**	0.059**
		(0.025)	(0.025)	(0.025)	(0.024)
Tool Available X Very High Risk		0.034	0.026	0.006	0.018
		(0.025)	(0.024)	(0.024)	(0.024)
Constant	0.142***	0.093***	0.053***	0.157***	0.162***
	(0.002)	(0.004)	(0.018)	(0.025)	(0.050)
Supervisor FE	No	No	Yes	Yes	Yes
Time FE	No	No	No	No	Yes
Controls	No	No	No	Yes	Yes
Observations	22,804	22,804	22,804	22,804	22,804
R^2	0.00001	0.022	0.037	0.076	0.151
Adjusted R^2	−0.00003	0.022	0.034	0.072	0.127
Residual Std. Error	0.348 (df = 22802)	0.344 (df = 22794)	0.342 (df = 22732)	0.336 (df = 22711)	0.325 (df = 22183)
F Statistic	0.218 (df = 1; 22802)	58.210*** (df = 9; 22794)	12.357*** (df = 71; 22732)	20.161*** (df = 92; 22711)	6.355*** (df = 620; 22183)

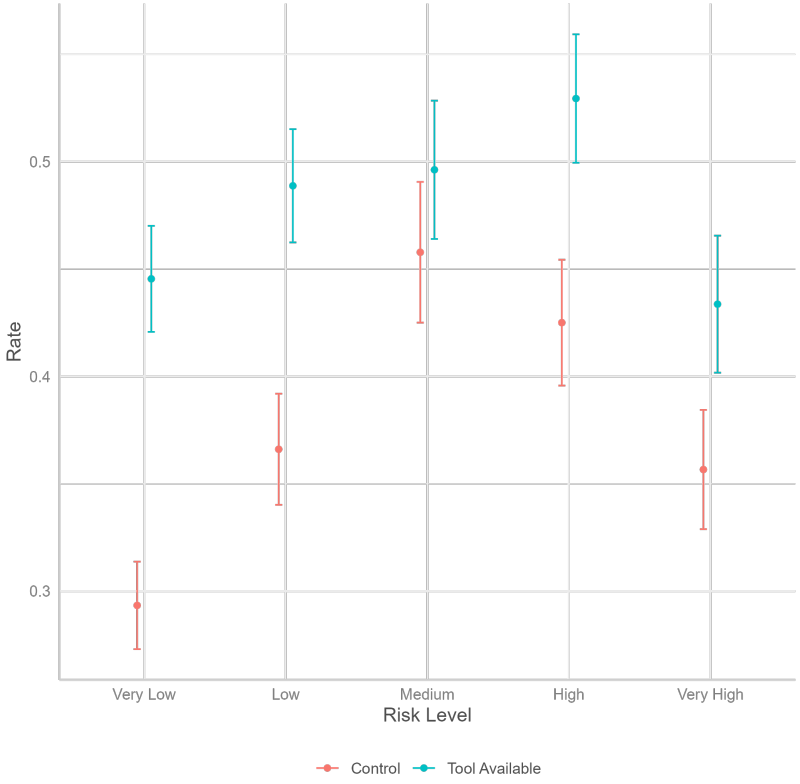
Note: Statistical significance markers: * p<0.1; ** p<0.05; *** p<0.01. Control variables include race, open case, reason for referral, type of reporter, and whether the reporter is mandated. Standard errors are clustered on both the supervisor and time (days before and after deployment).

Table 9: Effect of Tool Availability on FAR: Linear Regression Model

	Dependent variable:				
	FAR				
	(1)	(2)	(3)	(4)	(5)
Tool Available	−0.034***	−0.039**	−0.041**	−0.033*	−0.049**
	(0.010)	(0.018)	(0.018)	(0.018)	(0.022)
Low Risk		0.046***	0.045***	0.045***	0.043***
		(0.009)	(0.009)	(0.009)	(0.009)
Medium Risk		0.025**	0.023**	0.017*	0.010
		(0.011)	(0.011)	(0.011)	(0.011)
High Risk		−0.073***	−0.073***	−0.050***	−0.059***
		(0.010)	(0.010)	(0.010)	(0.010)
Very High Risk		−0.247***	−0.244***	−0.188***	−0.198***
		(0.009)	(0.009)	(0.010)	(0.010)
Tool Available X Low Risk		0.001	0.006	−0.003	−0.016
		(0.029)	(0.029)	(0.028)	(0.028)
Tool Available X Medium Risk		0.022	0.024	0.029	0.030
		(0.032)	(0.032)	(0.032)	(0.032)
Tool Available X High Risk		0.007	0.010	0.014	0.023
		(0.030)	(0.030)	(0.029)	(0.029)
Tool Available X Very High Risk		0.035	0.034	0.036	0.033
		(0.024)	(0.025)	(0.025)	(0.026)
Constant	0.331***	0.365***	0.354***	0.177***	0.075
	(0.003)	(0.006)	(0.028)	(0.034)	(0.061)
Supervisor FE	No	No	Yes	Yes	Yes
Time FE	No	No	No	No	Yes
Controls	No	No	No	Yes	Yes
Observations	22,804	22,804	22,804	22,804	22,804
R^2	0.0005	0.042	0.054	0.097	0.163
Adjusted R^2	0.0004	0.042	0.052	0.094	0.140
Residual Std. Error	0.469 (df = 22802)	0.459 (df = 22794)	0.457 (df = 22732)	0.447 (df = 22711)	0.435 (df = 22183)
F Statistic	11.144*** (df = 1; 22802)	111.281*** (df = 9; 22794)	18.452*** (df = 71; 22732)	26.572*** (df = 92; 22711)	6.989*** (df = 620; 22183)

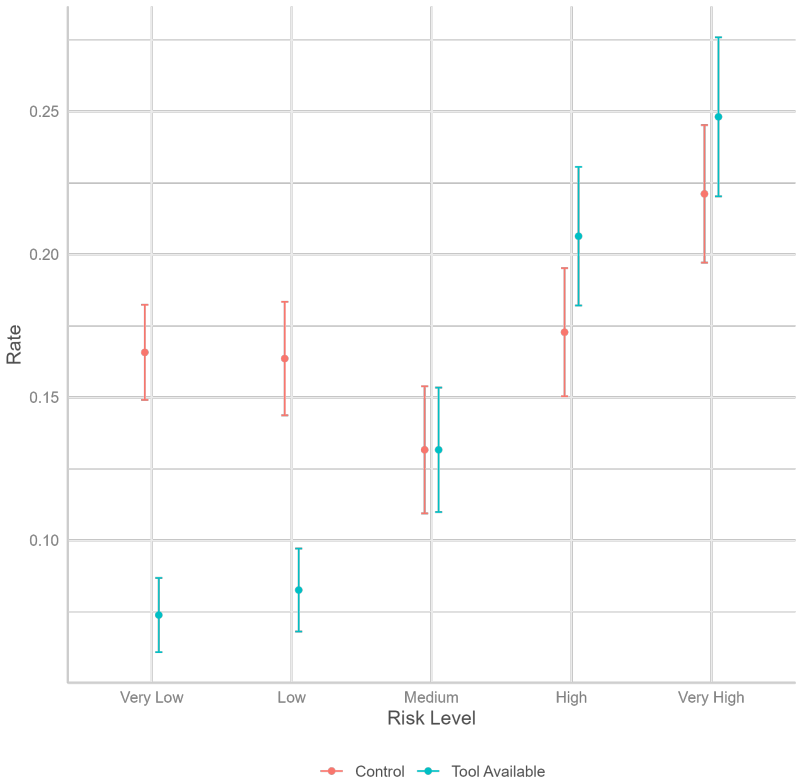
Note: Statistical significance markers: * p<0.1; ** p<0.05; *** p<0.01. Control variables include race, open case, reason for referral, type of reporter, and whether the reporter is mandated. Standard errors are clustered on both the supervisor and time (days before and after deployment).

Figure 19: Screen-Out Rates by Risk Levels



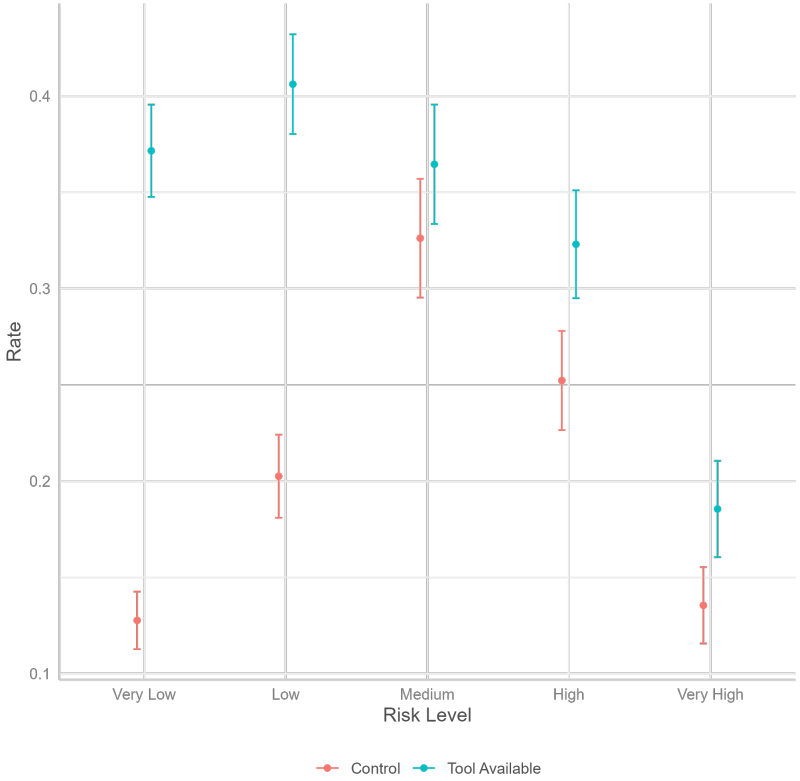
Note: This figure presents differences in means for the use of the tool across risk categories, with Tool Available as treatment.

Figure 20: HRA Rates by Risk Levels



Note: This figure presents differences in means for the use of the tool across risk categories, with Tool Available as treatment.

Figure 21: FAR Rates by Risk Levels



Note: This figure presents differences in means for the use of the tool across risk categories, with Tool Available as treatment.