sociological science

# Algorithmic Risk Scoring and Welfare State Contact Among US Children

Martin Eiermann

Duke University

**Abstract:** Predictive Risk Modeling (PRM) tools are widely used by governing institutions, yet research on their effects has yielded divergent findings with low external validity. This study examines how such tools influence child welfare governance, using a quasi-experimental design and data from more than one million maltreatment investigations in 121 US counties. It demonstrates that the adoption of PRM tools reduced maltreatment confirmations among Hispanic and Black children but increased such confirmations among high-risk and low-SES children. PRM tools did not reduce the likelihood of subsequent maltreatment confirmations; and effects were heterogeneous across counties. These findings demonstrate that the use of PRM tools can reduce the incidence of state interventions among historically over-represented minorities while increasing it among poor children more generally. However, they also illustrate that the impact of such tools depends on local contexts and that technological innovations do not meaningfully address chronic state interventions in family life that often characterize the lives of vulnerable children.

**Keywords:** predictive risk modeling; algorithms; child welfare; child maltreatment; welfare state; inequality

**Replication Package:** Access to restricted-use NCANDS data can be requested through the National Data Archive on Child Abuse and Neglect (NDACAN). Other data and replication code are available at: https://osf.io/dq3xp/.

**Citation:** Eiermann, Martin. 2024. "Algorithmic Risk Scoring and Welfare State Contact Among US Children" Sociological Science 11: 707-742.

**Received:** May 20, 2024

**Accepted:** July 2, 2024

**Published:** August 23, 2024

**Editor(s):** Arnout van de Rijt, Maria Abascal

**DOI:** 10.15195/v11.a26



THE now-common use of predictive risk modeling (PRM) tools in public administration has divided scholarly and public opinion. On the one hand, such tools promise increased efficiency in the allocation of scarce governmental resources and reductions in institutional bias (Russell 2015; Flores et al. 2016; Kleinberg et al. 2018; Meijer and Wessels 2019). On the other hand, many sociologists have argued that algorithmic governance simply automates the reproduction of inequality and deepens state interventions in the lives of marginalized populations (Eubanks 2017; Rudin, Wang, and Coker 2020; Joyce et al. 2021).

However, strong claims about the effects of PRM tools on state contact and racialized "patterns of inclusion" often rest on a delicate and inconclusive empirical foundation (Gillespie 2013:168). Many PRM tools are proprietary and black-boxed technologies, making it difficult for researchers to access relevant data (Pasquale 2015; Burrell 2016; Christin 2020). Research is dominated by studies that focus on predictive accuracy in very specific settings, which have produced divergent findings that additionally vary with competing definitions of algorithmic fairness (Kleinberg, Mullainathan, and Raghavan 2016; Angwin et al. 2016; Hamilton 2019; Hellman 2020; Wang et al. 2023; Kwegyir-Aggrey et al. 2023; Imai et al. 2023). A smaller strand of research analyzes how PRM tools affect administrative

decisions and patterns of state contact, yet such studies overwhelmingly focus on single jurisdictions in the criminal justice system and are sometimes based on hypothetical vignettes rather than observational data (Marshall and English 2000; Green and Chen 2019; Skeem, Scurich, and Monahan 2020; Garrett and Monahan 2020; Parker et al. 2022; Rittenhouse, Putnam-Hornstein, and Vaithianathan 2022). Their external validity is largely uncertain. Although it is sometimes possible to design randomized trials (Brantingham, Valasik, and Mohler 2018; Imai et al. 2023), this approach is usually unfeasible in settings where the use of PRM tools is already prevalent and subject to strict policy regimes and ethical constraints. The net result is a strong theoretical emphasis on specific social impacts of algorithmic governance, coupled with limited evidence of the effects that PRM tools have across different domains of public administration, communities, and populations (McNellan et al. 2022; Imai et al. 2023; Cuellar 2023).

This study analyzes how the adoption of PRM tools during child welfare investigations affects state contact, with a particular focus on effect heterogeneity across demographic groups, the low-to-high risk spectrum, and jurisdictions. Focusing on the child welfare system is especially pertinent for three reasons. First, involvement with the child welfare system is the first form of state contact for millions of US children and is also more common than juvenile justice system contact (Brame et al. 2014; Kim et al. 2017; Yi et al. 2020; Puzzanchera 2021; Putnam-Hornstein et al. 2021). Second, large adverse consequences of childhood maltreatment imply correspondingly high stakes of Child Protective Services (CPS) interventions in family life (Hussey, Chang, and Kotch 2006; Font and Berger 2015; Jaffee 2017). Third, algorithmic tools are widely embraced by CPS and have been used in millions of child welfare investigations (Cuccaro-Alamin et al. 2017; Brown et al. 2019; Parker et al. 2022).

CPS use algorithmic tools for different purposes. Some jurisdictions rely on such tools while screening incoming maltreatment reports, which determines whether child welfare investigations are initiated. Other jurisdictions use such tools only *during* ongoing investigations. This study focuses on the latter, because such investigations are uniquely consequential for children. When investigators confirm maltreatment, CPS involvement in family life is escalated through interventions that range from family counseling to the permanent termination of parental rights. Additionally, CPS investigations are an important source of institutional bias. One recent study (Baron et al. 2024) estimated that 81 percent of unwarranted racial disparities in foster care placements are due to investigative decisions, with only the remaining 19 percent due to screen-in practices.

Using data from 121 counties in eight states, I show that the adoption of PRM tools during CPS investigations decreased the overall incidence of maltreatment confirmations among investigated children by around 1 percent. This effect size is comparable to changes in maltreatment confirmations that are associated with a 3 percent absolute decrease in county poverty rates.[1] Crucially, such decreases were concentrated among children with relatively low risk profiles and among Hispanic and Black children, which complicates the claim that algorithmic tools increase the exposure of minorities to particularly assertive state interventions in family life. In contrast, confirmed maltreatment did not change significantly for White

children after the adoption of PRM tools; and high-risk children experienced a significant *increase* in confirmed maltreatment. Such high-risk children were especially likely to come from families that depended on public assistance, highlighting the particular significance of PRM tools for the contemporary governance of poverty (Soss, Fording, and Schram 2011; Eubanks 2017; Stevenson 2018). Effect sizes varied across US counties. PRM tools also did not reduce the incidence of subsequent maltreatment confirmations, indicating that such tools (which are partly designed to aid in the prevention of future abuse and neglect) may not be successful at counteracting chronic state contact among vulnerable children (Parker et al. 2022).

These findings, which are robust to different data selection criteria and model specifications, add empirical nuance and theoretical substance to discussions of algorithmic governance. Moving beyond broad but often speculative claims about the perils and promises of PRM tools, they demonstrate the uneven impact of such tools in welfare systems that are already characterized by pervasive racial bias in administrative decision-making (Baron et al. 2024) and are also strained under high employee turnover and low staffing levels (Edwards and Wildeman 2018). In this context, the embrace of auto-generated risk scores has the potential to reduce the incidence of state interventions among historically over-surveilled minorities but increase such interventions among poor children more generally. Understanding how the deployment of algorithmic technologies in public administration shapes patterns of state contact therefore requires an analytic focus on effect heterogeneity across places and populations, combined with a theoretical emphasis on disparate impacts as a core aspect of algorithmic governance in the contemporary United States. Such impacts can manifest in ways that corroborate the common understanding of algorithms as instruments of poverty governance while also complicating broad claims about the algorithmic reproduction of racial inequities.

Yet the findings of this study also occasion skepticism about the efficacy of actuarial tools (Stevenson 2018) and the allure of what Morozov (2013) has called "solutionism": The belief that intractable social problems and state-propagated inequalities can be remedied through technological innovation. The adoption of PRM tools by CPS failed at achieving one of the stated aims of such tools, that is, to reduce the exposure of vulnerable children to recurring confirmed maltreatment. Persistent social vulnerabilities (and inequities therein) that characterize the lives of disadvantaged children and their families in the United States are not amenable to quick technological fixes.

I arrive at these findings by matching more than one million child welfare investigations before and after the adoption of PRM tools in US counties. This quasi-experimental setup allows me to estimate the effects of a treatment of interest (in this case, PRM tool adoption) on child welfare investigations net of other potentially confounding factors, and additionally to assess effect heterogeneity across sub-populations and jurisdictions. It also yields findings that can be subjected to a battery of robustness checks that vary key data selection criteria and model specifications. I additionally use a regression discontinuity design to corroborate core findings by looking for discontinuities in system contact around the PRM tool adoption date.

This research design, which offers a potential framework for studies of algorithmic governance in other domains, adopts a pragmatist focus on the "palpable consequences" of technological innovation in the US welfare system (Solove 2002:1091), and thus makes no claims about the predictive power and performance of the PRM tools themselves (e.g., the accuracy of risk scores or the frequency of false negative or false positive confirmations). This choice is partly motivated by non-negotiable data constraints. But more importantly, it reflects a substantive concern with, and a theoretical emphasis on, the second-order impacts that new techniques of governance have on patterns of state contact (Gillespie 2013; Brayne 2017; Eubanks 2017). The social significance of algorithms depends to a lesser degree on the mysterious happenings within the black box and depends to a greater degree on the effects that propagate outwards into society.

## Competing Perspectives on Algorithmic Governance

The incorporation of algorithmic scores into bureaucratic routines illustrates that the techniques through which the American state manages social vulnerabilities have drastically evolved in recent years (Brayne 2017; Church and Fairchild 2017; Hannah-Moffat 2018; Mau 2018; Katzenbach and Ulbricht 2019; Burrell and Fourcade 2021). Advocates of PRM tool adoption argue that this can improve administrative decision-making by analyzing information more accurately than frontline workers alone (Russell 2015; Cuccaro-Alamin et al. 2017; McNellan et al. 2022), by increasing the ability to identify and correct unfair but difficult-to-detect decision-making (Kleinberg et al. 2018, 113; Brown et al. 2019), and by improving the targeting of state interventions in family life (Schwartz et al. 2017; Duwe and Kim 2017; Rittenhouse, Putnam-Hornstein, and Vaithianathan 2022).

Specifically in the child welfare system, early tests of PRM tools generated predictions that outperformed standard regression models in assessing overall risk levels among system-involved children, showed improved accuracy across different risk levels, predicted injury-related medical encounters among high-risk children, and reduced racial disparities in screen-in decisions (Marshall and English 2000; Daley et al. 2016; Schwartz et al. 2017; Vaithianathan et al. 2020; Rittenhouse, Putnam-Hornstein, and Vaithianathan 2022). Similar benefits have also been reported for recidivism tools in the criminal justice system (Duwe and Kim 2017), although other recent research on criminal justice algorithms has only found limited efficiency gains and small but potentially stratifying effects (Stevenson 2018; Imai et al. 2023). Generally speaking, this research suggests that algorithms may outperform humans in settings where real-time feedback is unavailable (so human decision-makers cannot adjust future decisions based on prior outcomes), where decisions need to incorporate a large number of variables, and where predictive equality across groups is high (Zeng, Ustun, and Rudin 2017; Sun and Gerchick 2019; Lin et al. 2020; Wang et al. 2023).

However, much of recent social science research has adopted a critical stance. In one widely cited work, Eubanks (2017:10) pointedly asks, "How has the digital revolution become a nightmare for so many?" According to this view, PRM tools can constrict opportunities for marginalized groups and substitute human bias with

"machine bias," especially when these tools are fed with data that are patterned by social histories of exclusion or are selectively used to justify biased administrative decision-making (Eubanks 2017:10; Angwin et al. 2016; Meijer and Wessels 2019; Zajko 2021). There is ample evidence that marginalized families already experience disproportionate levels of family surveillance and administrative bias (Harcourt 2006; Eubanks 2006; Maguire-Jack and Font 2017; Eubanks 2017; Fong 2020; Roberts 2022; Baron et al. 2024); and the adoption of PRM tools may simply reinforce such structural inequities (Skeem, Scurich, and Monahan 2020; Samant et al. 2021). For example, risk scores in the criminal justice system may increase the likelihood of incarceration among relatively poor defendants (Skeem, Scurich, and Monahan 2020) or lead to the over-prediction of recidivism risk among Black and Hispanic individuals (Dressel and Farid 2018; Angwin et al. 2016; Hamilton 2019). This has led to charges that algorithms deployed by governing agencies "[see] without knowing" (Ananny and Crawford 2016:973): By focusing on indices that are "easily quantifiable" and by hiding pervasive societal biases behind a veneer of computational neutrality, risk scores can justify exclusionary decisions and increase state interventions in the lives of marginalized populations (Saxena et al. 2020:9; Starr 2014; Eubanks 2017; Rosen, Garboden, and Cossyleon 2021; Brayne and Christin 2021; Bigman et al. 2023).

At stake in these debates is whether algorithmic tools affect (1) the overall reach of the contemporary American state into family life and (2) the unevenness of state interventions along racial and socio-economic lines. These are important sociological questions, yet empirical findings are often divergent and of limited external validity. Research is dominated by studies that focus on the special case of bail and parole decisions in the criminal justice system (Duwe and Kim 2017; Dressel and Farid 2018; Angwin et al. 2016; Stevenson 2018; Skeem, Scurich, and Monahan 2020; Imai et al 2023), are restricted to individual jurisdictions (Garrett and Monahan 2020; Rittenhouse, Putnam-Hornstein, and Vaithianathan 2022; Parker et al. 2022), are reliant on hypothetical vignettes (Green and Chen 2019; Skeem, Scurich, and Monahan 2020), or are based on contested methodologies and definitions of fairness (Flores et al. 2016; Kleinberg, Mullainathan, and Raghavan 2016; Corbett-Davies et al. 2017; Rudin, Wang, and Coker 2020). Recent empirical work has also cast doubt on generalizing claims about algorithmic governance by showing that such PRM tools may have "little overall impact" on administrative decisions (Imai et al. 2023:168), that expected efficiency gains "did not occur" (Stevenson 2018:369), and that their effects are narrowly concentrated among specific sub-groups (Imai et al. 2023).

The present study extends this literature in several ways. First, it broadens the empirical scope beyond the criminal justice system. Although criminal justice contact typically begins during adolescence or early adulthood (Neil and Sampson 2021), the use of PRM tools is also common in welfare agencies that interface with a significant percentage of American children at an earlier and highly consequential stage of the life course. CPS investigate around one third of US children before age 18, and racial and class disparities in maltreatment investigations and confirmations are high (Kim et al. 2017; Yi et al. 2020; Putnam-Hornstein et al. 2021; Edwards et al. 2021). Second, this study directly measures the effects of PRM tools on system

contact and thereby sidesteps definitional problems about what constitutes a "fair" or "accurate" metric of risk (Corbett-Davies et al. 2017). Third, it assesses effect magnitude and heterogeneity across multiple jurisdictions. This overcomes a key limitation of single-jurisdiction studies: CPS contact varies considerably across the US and the impact of algorithmic tools additionally depends on their integration into complex and localized bureaucratic routines (Edwards et al. 2021; Brayne and Christin 2021; Pruss 2023), yet single-jurisdiction studies cannot test for effect heterogeneity across administrative contexts.

## PRM Tools in the Child Welfare System

CPS in 20 US states have used PRM tools since 2013. These tools can be used while screening incoming maltreatment reports (which determine if CPS launch a full investigation), during ongoing maltreatment investigations (which either confirm or fail to confirm maltreatment), during the subsequent placement of children with confirmed maltreatment in foster care, and during decision-making about the potential reunification of fostered children with their biological parents. In this study, I focus specifically on jurisdictions that used PRM tools during ongoing maltreatment investigations (also known as "open case review tools"). CPS investigate all reports of maltreatment that pass an initial screening test, but only children whose maltreatment is subsequently confirmed are exposed to an escalating repertoire of state interventions. The outcomes of such investigations—called "dispositions" by CPS—are informed by data collected during at-home visits and family interviews, data on prior CPS contact, and algorithmically generated risk scores.

The American child welfare system is primarily organized at the state-level. Federal agencies provide funding and establish standards of care, for example, through the Adoption and Safe Families Act (ASFA) and the Child Abuse Prevention and Treatment Act (CAPTA). But state authorities set policies, administer services, and oversee local operations. County agencies then handle direct operations, including case management, although their relative autonomy can vary. Within this system, PRM tool adoption and relevant administrative procedures are commonly determined by the state (although some pilot programs were only implemented in specific counties, for example in Pennsylvania, Florida, and Oklahoma)—and adoption and contract termination dates are generally uniform within states but vary across states—because the development and licensing of such tools is largely funded through state CPS budgets and because counties usually rely on electronic tools provided by state agencies. Although county caseworkers may differ in how they pragmatically interact with risk scores, the computation of risk scores and the administrative events that are triggered by high scores are standardized.

In states and counties that use PRM tools during CPS investigations, risk scores are ordinarily generated for all investigated children. These scores are derived from analyses of historical data and predict the likelihood of future harm (Ruscio 1998; Saxena et al. 2020). The most widely used open case review tool, called the Eckerd Rapid Safety Feedback (ERSF), computes the risk of experiencing an additional maltreatment disposition within a 12-month window and flags all children with a risk prediction score above 50 percent, which commonly triggers two developments:

First, caseworkers perform an in-depth review of information collected from at-home visits (and can request supplemental information). Second, a supervisor independently reviews the investigative files, which can in turn lead to an additional risk review and a consultation with caseworkers before a final disposition is reached and an action plan is formulated (Parker et al. 2022). Another tool, the Severe Harm Predictive Model, similarly predicts the likelihood of future physical or sexual abuse within an 18-month timeframe and flags high-risk children to caseworkers and supervisors during conference meetings about dispositions and recommended action plans.

## Effect Heterogeneity in the Algorithmic Governance of Social Vulnerability

Maltreatment confirmations vary across jurisdictions and are elevated among poor children (Maguire-Jack et al. 2015; Drake, Lee, and Jonson-Reid 2009; Maguire-Jack and Font 2017), in part because underlying risk factors of maltreatment also vary geographically and are more prevalent in poor communities (Drake and Pandey 1996; Maguire-Jack, Font, and Dillard 2020). But local CPS caseworkers also retain considerable discretionary authority to determine if potentially ambiguous signs of abuse and neglect constitute actionable maltreatment (Gaudin 1995; Committee on Child Abuse and Neglect 2002). Their decisions can reflect practical resource and training constraints (DiMario 2022), but recent work also suggests that caseworker authority is exercised in a racialized manner, which in turn increases disparities in state interventions beyond disparities in the prevalence of underlying risk factors (Fong 2020; Roberts 2022; Baron et al. 2024).

Frontline workers additionally retain some degree of autonomy over the use of PRM tools during administrative decision-making. Prior work has shown, based largely on a combination of ethnographic observations and interviews, that officials working in the criminal justice system and the child welfare system can alternatively treat risk scores as particularly objective measures of risk (Brayne and Christin 2021), reject the widespread adoption of such scores as a threat to the autonomy and skills of judges and trained bureaucrats (Burton et al. 2020; Brayne and Christin 2021), question their accuracy and practical utility (Stevenson 2018; Pruss 2023), integrate algorithmic scores into holistic risk assessments (Cheng et al. 2022), exhibit greater risk aversion during consequential administrative decisions (Green and Chen 2021), or otherwise use algorithmic scores in unintended ways (Stevenson and Doleac 2022). Table 1 provides a schematic overview of the various potential mechanisms through which the adoption of open case review tools can affect investigative decision-making. It illustrates that the expected impact of such scores on administrative outcomes (in this case, dispositions at the end of a child welfare investigation) is closely linked to the specific ways in which PRM tools are used at the bureaucratic frontline.

Some studies understand the potentially disparate impacts of PRM tools primarily as a consequence of poor algorithmic calibration, which may prevent predictive equality across populations (Zeng, Ustun, and Rudin 2017; Angwin et al. 2016;

**Table 1:** Potential impacts of PRM tool adoption on CPS investigations.

| Impact on CPS investigations | Expected change in investigative dispositions |
|---|---|
| Increased targeting of high-risk cases | Elevated risk of false positive confirmations; decreased risk of false negative dismissals |
| Reduced targeting of low-risk cases | Elevated risk of false negative dismissals; decreased risk of false positive confirmations |
| Risk scores as scrutineering tools (increasing attention in borderline cases with limited/discordant evidence, e.g. by triggering in-depth reviews and additional evidence collection) | Decreases in false negative dismissals and false positive confirmations |
| Risk scores as discovery tools (identifying potentially unrecognized high-risk cases) | Decreased risk of false negative dismissals; elevated risk of false positive confirmations |
| Risk scores as justificatory tools (legitimizing pre-existing administrative decisions with selective references to PRM) | None |
| Prioritization of risk scores over alternative evidence (deferring to PRM as particularly objective measures of risk) | Contingent on PRM calibration |
| Discordant interpretations (caseworker disagreement over accuracy/utility of PRM) | Contingent on resolution of disagreement |
| Dismissal of risk scores (deprioritizing/ignoring PRM during open case reviews) | None |
| Symbolic compliance (acknowledging PRM without changes in administrative procedures) | None |

Hamilton 2019). Yet the research discussed above also suggests that such impacts can emerge more dynamically when risk scores are thrust into complex processes of administrative decision-making. The latter perspective implies the likelihood of (but does not demonstrate the existence of) heterogeneous effects across institutional contexts and jurisdictions: if caseworkers use risk scores differently across jurisdictions or use them differently depending on the socio-demographic characteristics of investigated children, a singular analytic focus on average effects may obscure substantial effect heterogeneity across jurisdictions or sub-populations of vulnerable children.

Given this demonstrably uneven landscape of state interventions in family life, the present study focuses not just on average effects across all US children but also on effect heterogeneity across jurisdictions and sub-populations of children. This significantly extends studies of algorithmic governance that are based on data from single jurisdictions, because understanding effect heterogeneity and the uneven accrual of algorithmic burdens and benefits across different populations remain central issues in social-scientific research on algorithmic governance (Christin 2018:274). The differential ability to reap the benefits—or carry the burdens—of technological innovation in the governance of social vulnerability is a key mech-

**Table 2:** Effect heterogeneity scenarios.

| Effect heterogeneity scenario | Expected outcome | Potential explanation |
|---|---|---|
| Scenario 1: Effects are heterogeneous across groups of children | 1a: Effects vary across ethno-racial groups | PRM tools affect administrative scrutiny and/or system contact among specific ethno-racial minorities |
| | 1b: Effects vary between low-risk and high-risk children | PRM tools increase scrutiny during high-risk investigations and/or redirect scrutiny away from low-risk investigations |
| Scenario 2: Effects are heterogeneous across jurisdictions | 2a: Effects vary by county socio-demographics | PRM tool effects depend on local conditions like urbanicity and poverty rates |
| | 2b: Effects vary by strain of CPS resources | PRM tool effects depend on front-line bureaucratic workload |

anism through which population-level inequality is maintained (Eubanks 2017; Cuéllar and Huq 2021:342).

I specifically focus on two effect heterogeneity scenarios at the bureaucratic frontline, summarized in Table 2: effect heterogeneity can arise when (1) effects differ across groups of children who either belong to different ethno-racial groups or are alternatively flagged as being high- or low-risk, or when (2) effects differ across jurisdictions. We can expect scenario (1) if the deployment of PRM tools allows CPS frontline caseworkers to direct administrative scrutiny towards, or away from, children who fit specific profiles. For example, PRM tool adoption may have stronger effects for children whose perceived risk of experiencing recurring abuse and neglect is particularly high (if such children experience substantively higher levels of scrutiny after the introduction of algorithmic scores than they would have experienced without the availability of PRM scores to investigators); or it may have stronger effects among children whose perceived risk of experiencing recurring abuse and neglect is relatively low (if the availability of PRM scores makes it more likely that such investigations are resolved without triggering state interventions in family life). PRM tools may also lead to the disproportionate identification of ethno-racial minorities as high-risk (potentially increasing false positives among those children and exacerbating racial inequalities in system contact), or alternatively allow caseworkers to assess maltreatment allegations more accurately in communities that have historically been over-surveilled (e.g., Black families) or are less legible to government officials due to linguistic and other barriers (e.g., Hispanic/Latinx families).

We can expect to find patterns along the lines of scenario (2) if the utility of PRM tools depends primarily on local community characteristics or local administrative needs. For example, PRM tools may offer a greater marginal benefit in large urban counties that process a high volume of annual CPS investigations, or in relatively poor and under-resourced counties. Findings in line with scenario (2) would also

offer suggestive evidence that the social impacts of PRM tools depend less directly on predictive accuracy but depend substantially conditioned on local administrative contexts.

# Data and Methods

## Data Sources

I link county-level data on PRM tool adoption to child-level microdata from the National Child Abuse and Neglect Database System (NCANDS) to estimate the effects of PRM tool adoption on patterns of state contact. Specifically, I analyze 1,028,074 NCANDS reports from 121 counties that used PRM tools during ongoing maltreatment investigations between 2016 and 2019. Three counties in Florida had previously adopted such tools during a pilot program in 2013. However, NCANDS data (which are reported voluntarily by state governments) are only complete for all fiscal years since 2012. Because the analyses below require complete data on CPS investigations for several years prior to CPS tool adoption, I drop these three counties from the analysis.

The 121 counties included in the analysis are located in eight U.S. states (Alaska, Connecticut, Illinois, Indiana, Louisiana, New Hampshire, New York, and Oklahoma). I compile this list of counties from an original data set that includes the start and end dates of PRM tool usage in each US jurisdiction that ever adopted such tools, the adoption stage (e.g., exploratory/pilot program, in-use, terminated), and characteristic tool features, including the unit of risk analysis (person-based vs. place-based) and the relevant stage of CPS contact (e.g. initial call screening, open case review, foster care placement, or family reunification). Data were collected with the help of research assistants from several sources: (1) publicly available administrative documents—including presentations, reports, and handbooks from state Departments of Child Support Services and state Departments of Health & Human Services, as well as Annual Progress and Services Reports filed by state governments with the US Department of Health and Human Services—; (2) newspaper articles related to PRM tools in the child welfare system, accessed through Nexis Uni and Google News; (3) reports and presentations from research institutions and non-governmental organizations like the American Civil Liberties Union; and (4) data requests to state governments. The data set covers all 50 US states from 2013 to 2020, although counties in only 20 states ever adopted PRM tools for regular use during that period (and only nine states adopted open case review tools that are used *during* CPS investigations). Data for each state was validated against multiple sources to confirm that all instances of PRM tool adoption are included in the data set. In several cases, CPS officials considered the adoption of PRM tools but never progressed to a pilot program or implementation. These are not counted in the data. County-level adoption patterns are shown in Figure 1.

I link these data to NCANDS annual Child Files, which cover all 50 states (plus DC and Puerto Rico) for all years since 2012. NCANDS Child Files include information about the source and date of each screened-in maltreatment report received by CPS, the demographic characteristics of each reported child (including race, gender,

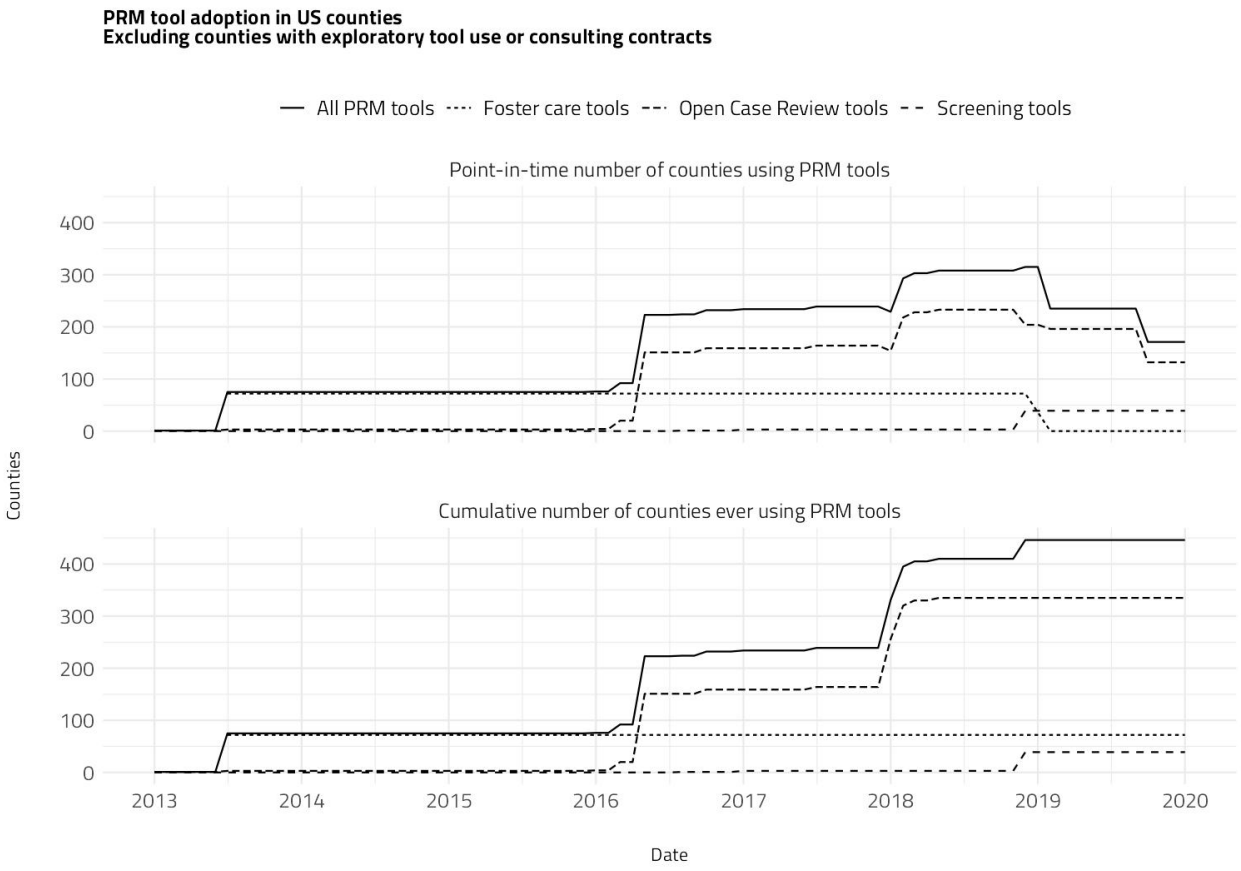


**Figure 1:** Timeline of PRM tool adoption by US counties.

and age-at-report), and information about administrative outcomes, including the report disposition (i.e. whether and when maltreatment was confirmed). County identifiers are masked for counties with fewer than 1000 annual reports to protect the privacy of vulnerable children. As a result, the number of counties included in this study (121) is smaller than the number of counties that ever had access to algorithmically generated risk scores during CPS investigations (335), and findings may not be generalizable to smaller and predominantly rural counties that fall below the masking threshold.

NCANDS data are complete for key administrative variables, including date of report, report disposition, and date of disposition. Information on age and gender is missing for a trivially low number of children (between 0.4 percent and 0.7 percent annually), but ethno-racial information is missing for more children (e.g. 12.6 percent missing in 2019). Around $1/4^{\text{th}}$ of missing values can be inferred directly from adjacent years, using NCANDS alphanumeric identifiers that are unique to each child within each state and stable over time. I impute missing values in the remaining cases, ensuring that the distribution of imputed values matches

the observed distribution within each county. This approach assumes that ethno-racial information is missing at random. I confirm the validity of this assumption with spot checks that compare the distribution of directly observed ethno-racial information to the distribution of ethno-racial information can be inferred from adjacent years (using the alphanumeric identifiers mentioned above), finding close distributional matches. In a separate robustness check, I also drop reports with partially missing demographic data and obtain substantively similar results.

Annual county- and age-specific poverty rates are taken from the Small Area Income and Poverty Estimates (SAIPE) program of the US Census Bureau; and annual race- and age-specific population counts come from the Surveillance, Epidemiology, and End Results Program (SEER) of the National Cancer Institute. I also use SEER population counts (in combination with NCANDS Child Files) to estimate overall CPS workload, defined as the percentage of the under-18 population in each county that was investigated for alleged maltreatment in a given year. Additional data on CPS budgets come from Child Welfare Financing Surveys conducted by the organization Child Trends. Surveys cover state fiscal years (SFY), which run from July 1 to June 30. I average across adjacent pairs of SFYs to match funding (in U.S. dollars) to calendar years; and I then aggregate all funds—federal, state, and local—that are specifically earmarked for CPS maltreatment investigations and all related administrative and personnel costs. I compute the average annual level of CPS funding per maltreatment investigation by dividing total CPS funding in each U.S. state by the number of screened-in maltreatment reports within that state in the same calendar year.

## Outcomes of Interest

I focus on PRM tools used *during* maltreatment investigations, which aim to identify high-risk cases by predicting the likelihood of future harm, allowing such cases to be flagged for additional scrutiny or prioritized during open case reviews. Importantly, the jurisdictions included in this analysis used PRM tools only during investigations and not during screen-in decisions. If a jurisdiction had simultaneously adopted PRM tools during investigations *and* prior screenings, it would be impossible to isolate changes in the outcomes of interest due to PRM effects on investigative dispositions from changes due to PRM effects on screen-in decisions. The analyses focus on the former, without biasing results due to contemporaneous changes in screen-in decisions.

I analyze effects on two child-level outcomes: the likelihood of having one's maltreatment confirmed; and the likelihood of experiencing a separate maltreatment confirmation within 12 months of a prior investigation. The first—and main—outcome of interest measures whether, at the end of an investigation, CPS determine that maltreatment allegations are "substantiated" or "indicated" (the latter disposition is only used in a small number of states; both dispositions are commonly treated as indicators of maltreatment confirmation). These determinations are a key stage of system involvement because they can trigger assertive state interventions in family life, ranging from family counseling to the forcible removal of the child from the home and the permanent termination of parental rights (Yi et al.

2020; Putnam-Hornstein et al. 2021; Parker et al. 2022). CPS involvement usually terminates for children with unconfirmed maltreatment but intensifies for children with confirmed maltreatment, leaving those two groups on divergent trajectories of system contact.

The second outcome of interest—experiencing an additional maltreatment confirmation within a 12-month period of a prior investigation—is central to the design of many open case review tools, which aim in part to reduce child deaths due to neglect and abuse by identifying children who are especially likely to experience chronic maltreatment (Parker et al. 2022). Put more formally, the aim of such tools is to identify children investigated in period $p_1$ who are most likely to be re-investigated and to have their maltreatment confirmed during subsequent periods $p_{1+n}$, and to pursue interim interventions in family life that preempt such outcomes. A key measure of algorithmic efficacy is therefore whether the adoption of PRM tools helps to interrupt recurring cycles of childhood trauma and state intervention (Beebe et al. 2023). I operationalize "additional confirmations" as follows: Among all children who experienced more than one CPS investigation, I identify the subset of children who experienced a maltreatment confirmation through a subsequent investigation within 12 months, conditional on the latter investigation having started after the disposition date of the first investigation. This last condition helps to distinguish between children who experienced multiple independent investigations and a small subset of children whose maltreatment report triggered several concurrent or partially overlapping investigations.

## Matching and Estimation Strategy

I use a quasi-experimental setup that conceptualizes PRM tool adoption as a county-level "treatment": Investigations that occur in each county prior to the adoption date are "untreated", whereas investigations that occur after the adoption date are "treated". Counties that never adopted PRM tools during CPS investigations are excluded from the analysis. Some related research includes such never-treated units as a control group based in part on a "parallel trends assumption" (Goodman-Bacon 2021). But in the present case, where adoption of PRM tools was not randomized across jurisdictions (although the timing was quasi-random) and where CPS policies and decision-making also varied across jurisdictions, this assumption does not hold; and never-treated jurisdictions differ in an unknown number of outcome-relevant ways (e.g., socio-demographics of investigated children, total annual volume of CPS investigations, and CPS investigation guidelines). Focusing on before/after comparisons within treated counties prevents bias due to potential differences between ever-treated and never-treated counties.

Potentially confounding environmental factors within treated counties are unaffected by PRM tool adoption. For example, tool adoption is not causally linked to local poverty rates, residential density, or other factors that have been shown to influence the overall prevalence of, and racial disparities in, system contact (Albert and Barth 1996; Drake and Pandey 1996; Maguire-Jack et al. 2015; Maguire-Jack, Font, and Dillard 2020). There is, additionally, a plausible claim of exogeneity of PRM tool adoption with respect to the frequency and patterning of screened-in

maltreatment reports: Decisions about PRM tool adoption are made away from the administrative frontline are not widely communicated to the public. They are unlikely to affect parental behavior, the calculus of mandated reporters like elementary school teachers or healthcare workers who submit maltreatment reports through phone intake hotlines, or the work of intake screeners. CPS caseworkers are generally informed about PRM tool adoption (and receive relevant training), but access to risk scores from open case review tools is provided only during ongoing investigations; and the scores do not affect whether an investigation is opened in the first place. Put differently, the adoption of PRM tools *during* ongoing investigations is not plausibly linked to the *prior* reporting of suspected maltreatment (which then triggers screen-in decisions) or to screen-in decisions (which then trigger investigations). As a result, the selection of children into investigations is likely unaffected by the treatment.

I study the effects of PRM tool adoption by first comparing potential outcomes for treated and untreated children; second, corroborating results with supplemental robustness checks and a placebo analysis; and third, implementing a regression discontinuity design that uses polynomial point estimators to identify potential jumps in outcome probabilities around the PRM tool adoption date. These methods make different assumptions and employ different strategies to estimate treatment effects. If their respective findings evince close agreement, claims about the effects of PRM tool adoption become more plausible (Legewie 2016).

I begin with a comparison of observed and counterfactual outcomes to identify the average treatment effect on treated children (ATT).[2] My estimation strategy involves: (1) running a nonparametric preprocessing model to match treated investigations to most similar untreated investigations; (2) specifying regression models for observed and counterfactual outcomes as a function of PRM adoption status, child- and county-level characteristics, and a vector of time-variant covariates; and (3) comparing weighted average potential outcomes for treated and untreated samples to compute the overall ATT. This approach assumes that outcomes for matched pairs of children would have been roughly equal in the absence of any treatment (Snowden, Rose, and Mortimer 2011).

I match post-treatment cases to most similar pre-treatment cases with a nonparametric generalized full matching model (Sävje, Higgins, and Sekhon 2021; Ho et al. 2007; Stuart 2010; Austin and Stuart 2017). Matches are exact for a child's county of residence, race, gender, reported maltreatment type, and prior maltreatment history (using a CPS identifier for children with a prior substantiated or indicated maltreatment disposition); and are approximate (using Mahalanobis distances) for age-at-report and month-of-report. This nonparametric preprocessing achieves very good covariate balance between pre- and post-treatment observations, reducing the maximum standardized mean differences from 0.033 (in the unmatched sample) to 0.004 (in the matched sample).

I then estimate potential outcomes for treated and untreated children with regression models that again control for child-level socio-demographics and maltreatment history, add county and state fixed effects, detrend the data by controlling for linear time, and also control for other (potentially non-linear) time-variant factors that could plausibly be linked to the outcomes of interest, including CPS funding levels,

local poverty rates, and the percentage of all children investigated by CPS in a given calendar year. Logit models take the following basic form:

$$\hat{y}_i = \alpha + \lambda p_i + t_i(\beta_1 x_{1i} + \ldots + \beta_j x_{ji}) + \gamma_1 z_{1i} + \ldots + \gamma_j z_{ji} \quad (1)$$

where $t_i$ is a binary indicator equal to 1 if unit $i$ received treatment, and equal to 0 otherwise; $p_i$ is a linear measure of time for unit $i$; $x_{1i}$ to $x_{ji}$ are potential moderators of treatment effects, including factors such as a child's race, age, gender, and state and county of residence; and $z_{1i}$ to $z_{ji}$ are potential time-variant confounders. The effect of PRM tool adoption can then be calculated by averaging across all weighted observations under both treatment conditions and comparing the resulting mean values for treated and untreated populations (Schafer and Kang 2008; Snowden, Rose, and Mortimer 2011). Results are reported as risk differences (RD), each with cluster-robust standard errors (Liang and Zeger 1986). Coefficients from the outcome model should not be interpreted directly and are not reported here (Ho et al. 2007).

The nonparametric matching approach reduces model dependence on the correct covariate specification (Ho et al. 2007). However, I also confirm the robustness of findings to different model permutations by alternatively dropping information on child demographics (including race, gender, and sex), types of reported maltreatment (i.e., physical abuse, psychological abuse, sexual abuse, or neglect), and county and CPS characteristics (i.e., local poverty rates and CPS workload and funding levels) before estimating the ATT. Results remain substantively the same.

## Effect Heterogeneity

I examine effect heterogeneity with ATT analyses that separately match samples for several sub-groups and then compute group-specific average treatment effects. I first compute ATT separately for non-Hispanic White, non-Hispanic Black, and Hispanic children (Scenario 1a in Table 2). I then assess effect heterogeneity across low-risk and high-risk children (Scenario 1b in Table 2) by splitting the sample into risk deciles prior to matching. Obtaining risk scores directly from PRM software providers is unfeasible due to the proprietary nature of such tools, restrictive licensing agreements, and obvious privacy concerns. Instead, I compute proxy risk scores with lasso-regularized regressions that predict the risk of experiencing additional maltreatment confirmations within 12 months of a prior investigation.

Lasso-based risk scores have been validated and are widely used in biomedical research (Pavlou et al. 2015). These scores are computed as follows: I first compile a training data set by identifying, for each of the 121 treated counties, the most similar county among all remaining US counties (i.e., counties without PRM tool adoption) based on a Euclidean distance matrix that includes information on county size (by area and population), demographic composition and racial stratification, urbanicity, poverty levels, and CPS workload (measured as the percentage of all children investigated annually by CPS). I then use reports from those 121 most similar counties to compute the optimal tuning parameter $\lambda$, using k-fold cross-validation (with $k = 10$). Risk predictions are based on information about a child's age, gender, and prior maltreatment history (including reported maltreatment type and the

occurrence or non-occurrence of a prior maltreatment confirmation). Supplemental models also directly use race and ethnicity as predictors; but having experienced a prior maltreatment confirmation, being reported for suspected neglect, and having such neglect confirmed by CPS are the strongest predictors in each lasso model. In a third step, I then use $\lambda$ to compute risk scores for each maltreatment report in the main data set with separate models for each state. Finally, I examine effect heterogeneity across jurisdictions by computing ATTs separately for each county (Scenarios 2a and 2b in Table 2).

## Robustness Checks

I corroborate key findings with additional robustness checks that adjust data selection criteria and model specifications. First, I repeat the analyses with a data set that includes only those children for whom complete demographic data are available in the original NCANDS files. This reduces the sample from around one million investigations to 911,863 but otherwise replicates the nonparametric matching and ATT estimation procedures described above.

Second, I adjust the time periods from which CPS data are selected. My main models compare investigations from the twelve months after PRM tool adoption to the twelve months prior to the adoption date. However, if the routine use of PRM tools was delayed (e.g. because caseworkers first had to receive necessary training), a small percentage of investigations that occurred soon after the official date of PRM tool adoption would falsely be considered "treated" (Burton et al. 2020). Conversely, if caseworkers accessed algorithmic scores only at the end of ongoing investigations, a small percentage of investigations that began shortly prior to the official date of PRM tool adoption and were grandfathered in would falsely be considered "untreated". I address this uncertainty by confirming the robustness of key findings to a different period specification. Specifically, I select maltreatment reports from post-treatment months 3-10 and then match these to the corresponding pre-treatment periods.[3] In effect, this excludes the two months immediately before and after the PRM tool adoption date from the analysis—an exclusion window selected because it is slightly longer than the average length of CPS investigations in the US (around 52 days during the 2019 calendar year). Selecting the same months in pre- and post-treatment years helps to eliminate potential bias from seasonal fluctuations in CPS contact, for example, due to a lower volume and a changed demographic composition of maltreatment reports during school holidays.

Third, I perform a placebo analysis that compares treatment effects after the (true) treatment date to effects observed after a prior (fictitious) treatment date. This placebo analysis can offer suggestive evidence in support of causal claims, based on the assumptions that treatment effects should only be observable after the actual date of PRM tool adoption and not after an alternative placebo date. Specifically, I shift the data selection window for each county one year into the past. In effect, this analysis replaces the true treatment date with a fictitious treatment date one year prior and then analyzes data from the year before and after this placebo treatment.

Fourth, I subset the sample by PRM tool. The basic aim (flagging high-risk children) and methodology (computing the likelihood of future harm, e.g. in a

12-month or 18-month window) of all open case review tools are similar. However, most jurisdictions (116 out of 121 counties) in the sample used a single tool, ERSF. Several large New York counties used alternative tools. Excluding these five non-ERSF counties from the analysis restricts the analysis to a specific and highly prevalent tool while also eliminating a set of counties that are unique in their overall size and CPS caseload.

### Regression Discontinuity Design

In a final step, I implement a sharp regression discontinuity (RD) analysis that yields local-polynomial point estimators and confidence intervals for local average treatment effects around the treatment threshold.[4] Although the matching analysis above focuses on differences between treated and untreated *populations*, the RD design allows me to assess whether the outcomes of interest changed discontinuously around the PRM tool *adoption date* (Hahn et al. 2001; Imbens and Lemieux 2008; Calonico, Cattaneo, and Titiunik 2014, 2015; Ito 2015; Hausman and Rapson 2018). This RD design adapts traditional RD methods by using time as the running variable, with the PRM tool adoption date as a threshold. As in the analyses above, a maltreatment investigation is considered to have been "treated" if it occurred after the date of PRM tool adoption in the respective county (so that $t_i = 1$ if $p_i > c$). The RD model then estimates the local average treatment effect with the following regression:

$$\hat{y}_i = \alpha t_i + \lambda p_i + \beta s_i + \theta_i + \eta_i + \gamma_1 z_{1i} + \ldots + \gamma_j z_{ji} \tag{2}$$

where $p_i$ is a linear measure of time (to control for linear time-series properties of the data), $s_i$ is a measure of seasonality (to control for cyclical fluctuations in maltreatment reports and CPS activity throughout the calendar year), $\theta_i$ and $\eta_i$ are state and county fixed effects, and $z_{1i} \ldots z_{ji}$ control for state- and county-level factors that may also change discontinuously, including CPS funding levels and the percentage of all children investigated by CPS in a given year. Supplemental models include additional dummy controls for two federal laws—the 2018 Family First Prevention Services Act and the 2019 Family First Transition Act—that impacted federal grant-giving under Title IV-B and Title IV-E. The main specification of the RD model uses triangular kernels to estimate local average treatment effects across different data selection bandwidths. Supplemental models use uniform kernels (giving equal weight to all investigations, regardless of how close to the PRM tool adoption date they occurred) as well as a so-called "donut" specification by dropping data from the four weeks surrounding the PRM tool adoption date.

Below, I sequentially present (1) overall ATT estimates and group- and jurisdictionally specific estimates obtained after nonparametric matching, (2) results from robustness checks and placebo analyses that corroborate core findings, and (3) complementary results from the alternative RD analysis.

## Results

The adoption of PRM tools occurred in the context of a high overall prevalence of, and persistent inequalities in, child welfare system contact. During the 2019

**Table 3:** Average treatment effects on the treated (ATT) by ethno-racial group.

| Population | Maltreatment confirmation | Subsequent confirmation within 12 months |
|---|---|---|
| All | −0.008** | 0.013** |
| | (0.002) | (0.001) |
| Non-Hispanic White | −0.005 | 0.003 |
| | (0.003) | (0.001) |
| Non-Hispanic Black | −0.008* | 0.016** |
| | (0.003) | (0.001) |
| Hispanic only | −0.012** | 0.017** |
| | (0.004) | (0.001) |

*Notes*: Estimates are reported in risk differences (RD). Cluster-robust SEs in parentheses.
$^{**}p < 0.01$; $^{*}p < 0.05$.

calendar year—the most recent non-pandemic year for which complete data are available—CPS conducted around 4.2 million investigations and confirmed maltreatment in around 820,000 cases, respectively affecting 4.7 percent and 1.0 percent of all US children and adolescents. To put this into perspective, law enforcement agencies performed around 700,000 juvenile arrests during the same calendar year, highlighting the overall size of the child welfare system relative to the juvenile justice system (Puzzanchera 2021). Black children were almost twice as likely as White children to be investigated for possible maltreatment (Black/White ratio = 1.9) and to have their maltreatment confirmed (ratio = 1.7). Hispanic children were also slightly more likely than White children to experience an investigation (Hispanic/White ratio = 1.1) and a maltreatment confirmation (ratio = 1.2).

The root causes of such disparities are disputed (Drake, Lee, and Jonson-Reid 2009; Drake et al. 2011; Fong 2020; Roberts 2022; Baron et al. 2024), and this study remains agnostic about them. It instead examines the effects of PRM tool adoption within this high prevalence/high disparity context, focusing on two key aspects of child welfare investigations: their final disposition (i.e., whether maltreatment allegations were confirmed by CPS investigators), and the likelihood of experiencing another subsequent maltreatment confirmation. The analyses below estimate average treatment effects among treated units (ATT), which are reported as risk differences $RD = I_t - I_u$, where $I_t$ is the incidence among all treated units and $I_u$ is the incidence among untreated units. Negative values thus indicate a post-treatment reduction in the risk of experiencing a given outcome.

## Effects on Child Welfare System Contact

Across all investigated children, the adoption of PRM tools by CPS led to a statistically significant reduction in maltreatment confirmations (Table 3). Compared to pre-treatment periods, children who were investigated for potential abuse and neglect after the adoption of such tools had an annual incidence of maltreatment confirmations that was 0.8 percent lower (ATT = −0.008, SE = 0.002). This is a non-trivial change when considering that only around one in five CPS investigations in

the US result in a maltreatment confirmation (e.g., 19.5 percent of all investigations during the 2019 calendar year). The probability of experiencing a maltreatment confirmation at the end of a CPS investigation varies considerably by location, ethno-racial identity, gender, and age; however, back-of-envelope calculations show that lowering the absolute incidence of confirmed maltreatment by 0.8 percent is roughly equal to a 4 percent relative decline. This observed change is also noteworthy give the plethora of other factors that can plausibly impact investigative outcomes, including staffing and funding levels (Edwards and Wildeman 2018), investigator discretion (Baron et al. 2024), and institutional and policy environments (Glisson, Green, and Williams 2012). Models used to estimate ATT control for such time-invariant and time-variant factors at both the state-level and the county-level, and ATT estimates can therefore be interpreted as effects that exist net of such potentially confounding factors.

Reductions in the incidence of maltreatment confirmations were concentrated among Black children (ATT $= -0.008$, SE $= 0.003$) and Hispanic children (ATT $= -0.012$, SE $= 0.004$). No statistically significant change was detectable for White children (ATT $= -0.005$, SE $= 0.003$). These findings hint at race-specific effects of PRM tools on administrative decision-making, and they partially contradict claims that the use of risk scoring algorithms in public administration will increase racial stratification. Such research has suggested, based on studies of the criminal justice system, that algorithmic tools are "biased against blacks" because they falsely assign minority individuals to high-risk categories and thereby trigger intensive and often punitive state interventions (Angwin et al. 2016; Flores et al. 2016). Perhaps surprisingly, I do not observe such effects in the child welfare system but instead find that populations which have long been over-represented in CPS investigations and among confirmed maltreatment dispositions—Non-Hispanic Black children and Hispanic children—were less likely to have their maltreatment confirmed once CPS incorporated algorithmic tools into investigations. Yet this finding accords with prior research on algorithms used during initial screen-in decisions, which also found that the adoption of such tools reduced racial disparities in system contact (Rittenhouse, Putnam-Hornstein, and Vaithianathan 2022). In the context of persistent and very high racial disparities, the introduction of algorithmic scores may partially offset well-documented institutional biases and racialized decision-making by frontline workers, which are pervasive (Baron et al. 2014).

However, overall and race-specific reductions in maltreatment confirmations occurred alongside an *increased* likelihood of maltreatment confirmation for one specific subset of children (Table 4): Among those in the highest decile of the risk distribution—based on lasso-regularized regression models that predict the risk of experiencing future maltreatment confirmations within a 12-month period—I observe a statistically significant increase in maltreatment confirmations after the adoption of PRM tools (ATT $= 0.021$, SE $= 0.007$). As shown in Figure 2, risk score deciles are significantly correlated with the direction and magnitude of ATTs, with effects equal in statistical significance but opposite in direction at the two ends of the low-to-high risk spectrum. Put differently, the adoption of PRM tools lowered the incidence of maltreatment confirmations among historically

**Table 4:** ATT by risk decile, using risk scores from lasso-regularized regressions.

| Risk decile | Maltreatment confirmation | Subsequent confirmation within 12 months |
|---|---|---|
| 1 (lowest) | −0.007 | 0.000 |
| | (0.005) | (0.004) |
| 2 | −0.013** | 0.009** |
| | (0.004) | (0.002) |
| 3 | 0.000 | 0.010** |
| | (0.003) | (0.001) |
| 4 | 0.002 | 0.010** |
| | (0.003) | (0.001) |
| 5 | 0.005 | 0.012** |
| | (0.003) | (0.002) |
| 6 | 0.003 | 0.017** |
| | (0.003) | (0.002) |
| 7 | 0.004 | 0.018** |
| | (0.003) | (0.002) |
| 8 | 0.004 | 0.015** |
| | (0.004) | (0.002) |
| 9 | 0.007 | 0.013** |
| | (0.004) | (0.003) |
| 10 (highest) | 0.021** | 0.019** |
| | (0.007) | (0.003) |

*Notes*: Estimates are reported in risk differences (RD). Cluster-robust SEs in parentheses.
$^{**}p < 0.01$; $^{*}p < 0.05$.

over-represented minorities but also contributed to a growing stratification in the incidence of maltreatment confirmations along the low-to-high risk spectrum.

This is the expected outcome if (1) lasso-derived risk scores approximate CPS assessments of maltreatment risk and (2) PRM tools are appropriately calibrated to flag high-risk children. But it also highlights the specific significance of PRM tools for the contemporary governance of poverty (Eubanks 2017). In lasso-regularized regressions, three factors were especially predictive of higher risk scores: Having already experienced a prior maltreatment confirmation ($\beta = 0.031$), being reported for suspected at-home neglect ($\beta = 0.005$), and having at-home neglect confirmed by CPS ($\beta = 0.015$). Previous research has found that such child-level characteristics—for example, experiencing chronic system contact and at-home neglect—correlate strongly with family and neighborhood poverty (Drake and Pandey 1996; Maguire-Jack and Font 2017). Information included in NCANDS reports also demonstrates that higher risk deciles were significantly associated with familial welfare reliance even after controlling for children's race and ethnicity. Children in the tenth decile of the risk distribution were 2.2 times as likely as children in the first decile of the risk distribution, 1.2 times as likely as children in the ninth decile, and 1.4 times as likely as the average CPS-involved child to come from families that received public assistance, including Temporary Assistance for Needy Families (TANF), Supplemental Security Income (SSI), food stamps, or Medicaid. This strongly

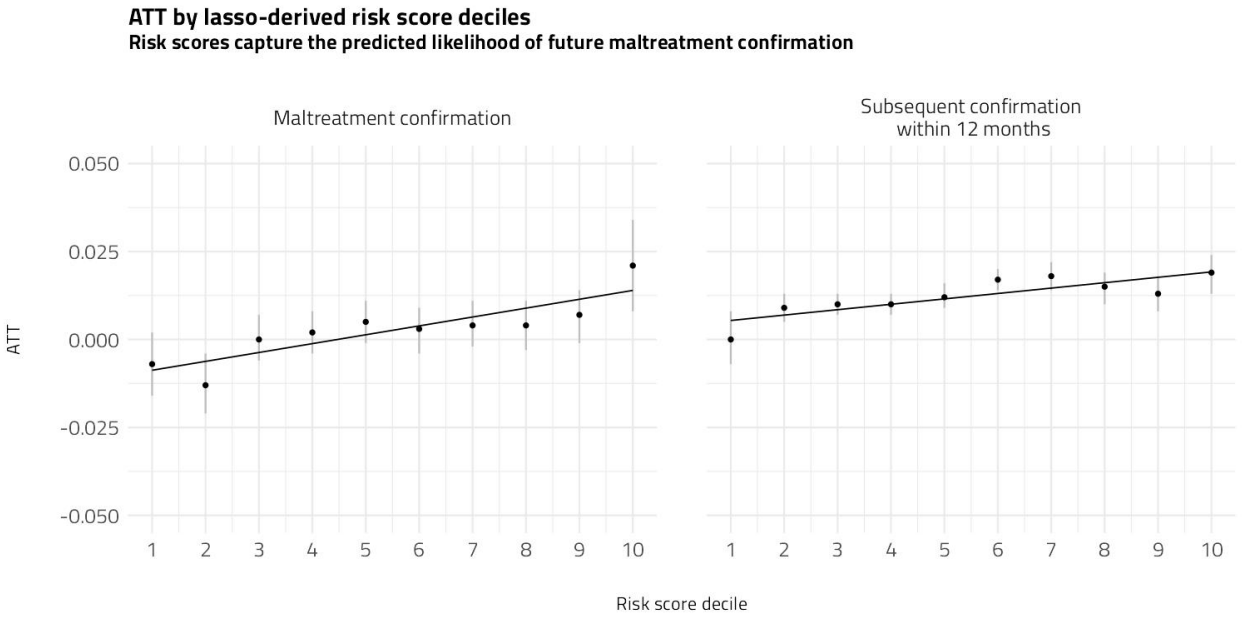


**Figure 2:** ATT by risk score decile, shown with 95 percent CI.

suggests that a stratification of maltreatment confirmations along the risk spectrum after the adoption of PRM tools is, substantively, a stratification of CPS interventions along socio-economic lines.

Focusing on the second outcome of interest—the risk of experiencing subsequent maltreatment confirmations—I find that children investigated after the adoption of PRM tools had a higher risk of experiencing future maltreatment confirmations (ATT $= 0.013$, SE $= 0.001$), with effects once again statistically significant among Black (ATT $= 0.016$, SE $= 0.001$) and Hispanic (ATT $= 0.017$, SE $= 0.001$) children. Crucially, however, the incidence of such subsequent confirmations was elevated across the entire risk spectrum (Fig. 2) except for the lowest-risk decile. Results from placebo tests (presented below) indicate that some caution is warranted when interpreting these effects as causal. As Parker et al. (2022) suggest, even a 12-month timeframe may be too long for strong causal claims and include outcomes that are too distal for targeted interventions in family life. However, across all models and robustness checks, I find no evidence that the adoption of PRM tools *reduced* the risk of future maltreatment confirmations among vulnerable children—which is a key aim behind the adoption of PRM tools.

Taken together, these findings indicate that shorter-term effects of PRM tools differed significantly across a diverse population of children—decreasing maltreatment confirmations among historically over-represented minorities but increasing the incidence of such confirmations among poor children more generally—and that such tools did not reduce the longer-term exposure of vulnerable children to maltreatment confirmations and the state interventions in family life that are commonly triggered by such dispositions.

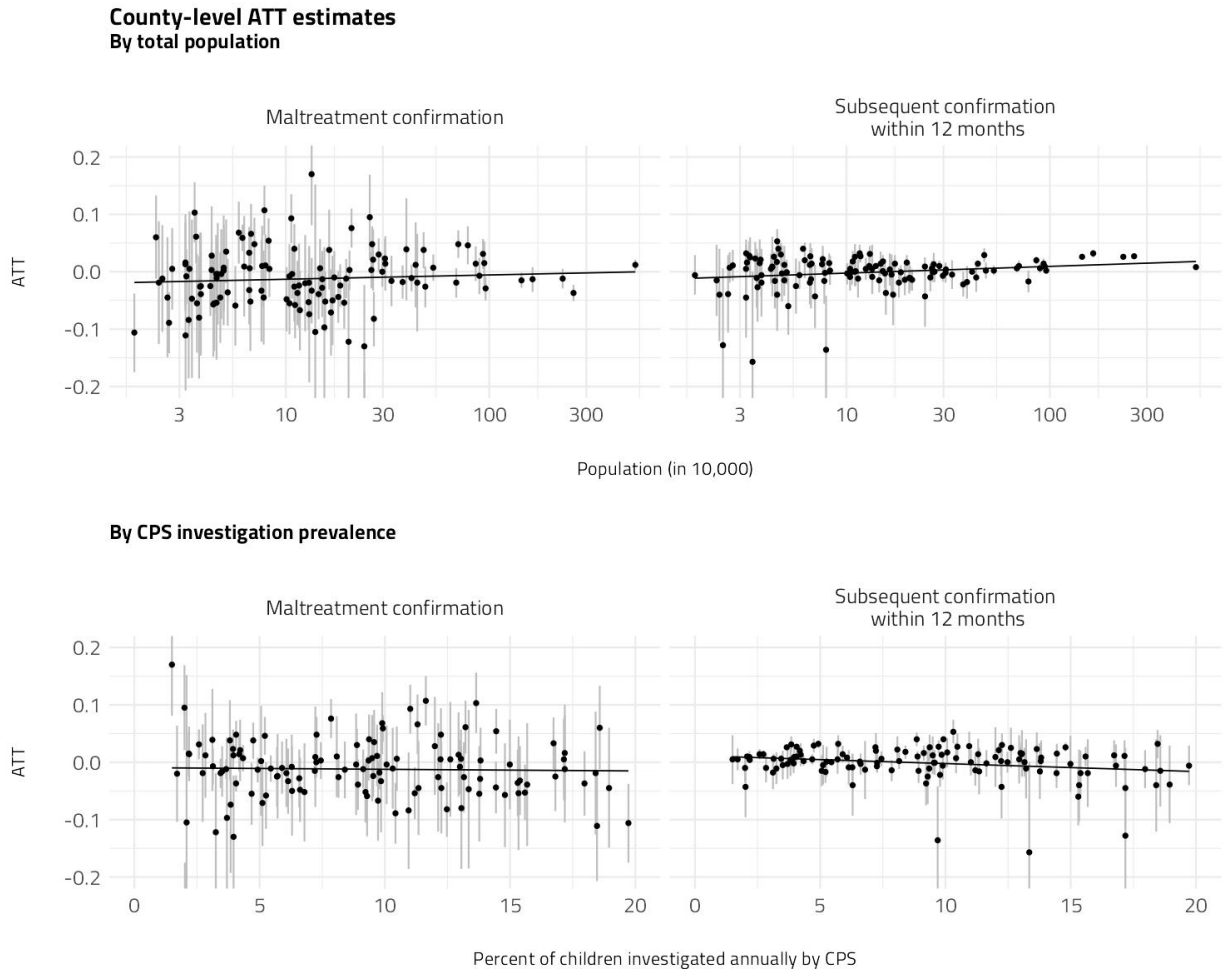


**Figure 3:** ATT by selected county characteristics, shown with 95 percent CI.

In Figure 3, I show results from additional decomposition analyses that separately estimate ATTs for each county included in the analysis. These analyses allow me to assess if effect heterogeneities observed across ethno-racial groups and across the risk spectrum are fundamentally affected by the clustering of populations (e.g., Hispanic children or poor children) in particular jurisdictions, or are alternatively affected by local administrative strain, such as the workload experienced by CPS frontline workers (i.e., heterogeneity scenarios 2a and 2b from Table 2). Figure 3 plots ATTs by total county population and by the percentage of all children investigated in each county during the 12 months after PRM tool adoption (as a proxy for administrative workload). Counties with larger total populations had marginally higher ATT for both outcomes of interest, and counties with greater administrative workloads had marginally lower ATT, but neither association is statistically significant. Results are substantively the same when using other environmental variables (including commonly used variables in research on child welfare governance), such

as local poverty rates, local average household incomes, the ethno-racial minority share of the local under-18 population, or CPS funding levels. I return to this finding in the discussion below, because it leaves open the possibility that variation in effect sizes across jurisdictions results from the uneven incorporation of risk scores into administrative decision-making by frontline caseworkers in child welfare offices.

## Robustness Checks

I confirm the robustness of key findings to several data selection and modeling choices, focusing on four alternative specifications: (1) I drop all children for whom information on gender or ethno-racial identity was missing in the original NCANDS files; (2) I restrict data selection to months 3-10 after PRM tool adoption (and select the corresponding pre-treatment months) to address uncertainties about the date at which PRM tools were de-facto integrated into routine CPS procedures; (3) I perform a placebo analysis that replaces the true treatment date with a fictitious prior treatment date; and (4) I subset the sample by PRM tool. Broadly speaking, robustness checks 1, 2, and 4 should replicate the results presented above in their direction, approximate magnitude, and statistical significance. Robustness check 3—the placebo analysis—should fail to replicate these results.

I find that results are substantively similar in robustness checks (1) and (2), as shown in Table 5. In both cases, overall ATT closely match estimates presented above. In particular, dropping reports with imputed socio-demographic information from the analysis has no effect on core findings. Restricting data selection to months 3-10 after PRM tool adoption also yields similar overall effect estimates, still shows largest effects among Hispanic children, and (more generally) shows significant differences in the effects of PRM tool adoption between White and non-White children. Robustness check (3)—the placebo test—largely fails to replicate these results, as hypothesized. Focusing on the first outcome of interest (maltreatment confirmations), I observe no statistically significant overall effects and no statistically significant effects for White and Black children separately. ATT are significant for Hispanic children, but the direction of the effect is reversed relative to the estimates presented above.

However, focusing on the second outcome of interest (subsequent maltreatment confirmations within a 12-month period) shows that results from the placebo test are similar to results obtained from the main models. This reduces the plausibility of strictly causal claims about the effects of PRM tools on future maltreatment confirmations and suggests that longer-term outcomes may be too distal to be reliably linked to the availability of PRM tools at the bureaucratic frontline (Parker et al. 2022). However, the placebo test still produces no evidence of a *decreased* risk of future maltreatment confirmations. Indeed, across all analyses and using different model specifications and data selection criteria, I find no evidence that PRM tools were successful in forestalling subsequent maltreatment and thereby reducing chronic state interventions in family life.

Results are directionally similar regardless of PRM tool, which suggests that the specific technological solution adopted by CPS has no major impact on the overall patterning of results. The incidence of maltreatment confirmations was reduced in

**Table 5:** ATT of PRM tool adoption from three robustness checks.

| Robustness check | Population | Maltreatment confirmation | Subsequent confirmation within 12 months |
|---|---|---|---|
| (1) Only reports with complete demographic data | All | −0.007** | 0.014** |
| | | (0.002) | (0.001) |
| | Non-Hispanic White | −0.006 | 0.001 |
| | | (0.003) | (0.002) |
| | Non-Hispanic Black | −0.007* | 0.017** |
| | | (0.004) | (0.001) |
| | Hispanic | −0.009* | 0.018** |
| | | (0.004) | (0.001) |
| (2) Data from post-treatment months 3-10 | All | −0.008* | 0.013** |
| | | (0.003) | (0.001) |
| | Non-Hispanic White | 0.010* | 0.007* |
| | | (0.005) | (0.004) |
| | Non-Hispanic Black | 0.000 | 0.016** |
| | | (0.006) | (0.002) |
| | Hispanic | −0.012* | 0.019** |
| | | (0.006) | (0.001) |
| (3) Placebo treatment date | All | 0.002 | 0.013** |
| | | (0.002) | (0.001) |
| | Non-Hispanic White | −0.003 | −0.001 |
| | | (0.003) | (0.002) |
| | Non-Hispanic Black | 0.005 | 0.014** |
| | | (0.004) | (0.001) |
| | Hispanic | 0.013* | 0.019** |
| | | (0.006) | (0.001) |

Notes: Estimates are reported in risk differences (RD). Cluster-robust SEs in parentheses.
$^{**}p < 0.01$; $^{*}p < 0.05$.

the subset of counties that used ERSF (ATT $= -0.005$, $SE = 0.002$) as well as in New York jurisdictions that used alternative tools (ATT $= -0.016$, SE $= 0.004$). ATTs for individual ethno-racial groups are also directionally similar to those presented above. The larger overall magnitude of effects in New York may be due to the unique challenges faced by CPS in America's largest metropolitan area, although testing that hypothesis is beyond the scope of this study.

## Regression Discontinuity

The analyses above estimate the effects of PRM tool adoption by comparing potential outcomes between untreated and treated cases. In a final step, I corroborate core findings by implementing an RD model that tests whether outcome probabilities jumped discontinuously around the PRM tool adoption date in each county. This RD design uses time (in weeks) as the running variable and the date of adoption

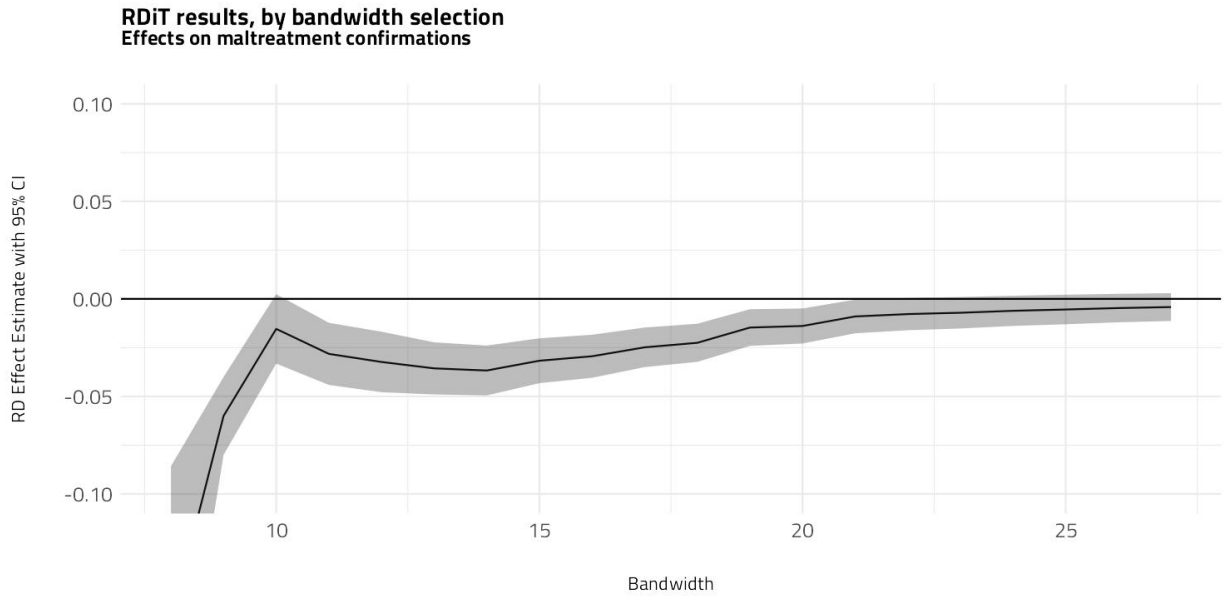


**Figure 4:** Regression discontinuity in time (RDiT) results, estimated using triangular kernels. Effects are shown across bandwidths 8 to 27 with 95 percent CI.

as the cutoff. Such discontinuity-in-time adaptations of traditional RD models are potentially sensitive to the time-series properties of the data-generating process. I address this by controlling for linear time as well as seasonality, and additionally control for child- and county-level factors that have been shown in prior research to affect the likelihood of maltreatment confirmations (Drake and Pandey 1996; Maguire-Jack et al. 2015; Edwards et al. 2021). I focus only on discontinuities in maltreatment confirmation, because the second outcome of interest (experiencing subsequent maltreatment confirmations within a 12-month period) implicitly draws on data far away from the cutoff, making this outcome poorly suited to RD analyses.

This analytic approach imposes a high bar, because the effects of PRM tools are potentially mediated by policy environments, staffing levels and administrative workloads, or buy-in from frontline staff, all of which can shape and potentially delay the integration of algorithmically generated scores into CPS investigations. Effects may manifest gradually, unevenly, and belatedly, instead of manifesting as clear discontinuities around the adoption date. However, despite this relatively high conceptual bar, core findings are replicated in the RD analysis. I find that the risk of maltreatment confirmations dropped after PRM tool adoption. This drop is observable among all children and also separately for different ethno-racial groups of children. As shown in Figure 4, the observed effect also holds across different bandwidth choices, which is a key parameter in RD analyses. Bandwidth selection determines the range of data from either side of the cutoff that are included during the RD estimation. Figure 4 shows that results are similar regardless of the number of periods (between 8 and 27 weeks, i.e. between 2 and 6 months) that are used to estimate RD effects. Results are also closely replicated in alternative RD

specifications that use uniform kernels or omit data from the four weeks preceding the date of PRM tool adoption (to account for uncertainties about the treatment of investigations that started close to the official adoption date). Overall, these results corroborate claims about the causal effect of PRM tools on maltreatment confirmations.

## Limitations

This study examines the effects of PRM tool adoption on patterns of system contact—or, as Christin (2020:906) puts it, the "reconfigurations that occur when algorithms, people, and institutions interact." As such, the study makes no claims about the predictive power and scoring accuracy of PRM tools or about the relative utility of different algorithmic performance metrics (Kwegyir-Aggrey et al. 2023). It also cannot determine if overall reductions in the incidence of maltreatment confirmations were due to a decrease in false positive confirmations (e.g., if the introduction of PRM tools increased the accuracy of caseworker decisions by directing administrative scrutiny towards truly high-risk cases) or due to an increase in false negative dismissals (e.g., if CPS failed to substantiate maltreatment among children who were assigned low risk scores and falsely de-prioritized). The fact that the incidence of confirmed maltreatment increased among children in the highest decile of the risk distribution offers suggestive evidence in favor of an explanation that emphasizes a re-targeting towards particularly vulnerable (and disproportionately poor) children. But PRM tools are constantly evolving; and it is therefore possible that differently calibrated or (perhaps more importantly) differently used PRM tools would have other effects.

This study also sidesteps important but substantially different questions about human-computer interactions in governing agencies (Stevenson 2018; Christin 2020; Burton et al. 2020; Brayne and Christin 2021; Cheng et al. 2022). Observed effect heterogeneity across US counties is not obviously correlated with environmental factors, such as local poverty rates, demographic composition, and CPS workload. This leaves open the possibility that such heterogeneity is best explained by local administrative contexts, especially because dispositions are usually determined at the bureaucratic frontline in conference meetings. But did CPS staff in different jurisdictions vary in their interpretation of risk scores? Did they assign more or less importance to those scores, or weigh them differently against potentially discordant evidence from at-home visits? Answering those questions may shed light on the specific mechanisms through which PRM tools influence patterns of state contact, already summarized in Table 1. It is possible, for example, that particularly large effects among Hispanic children reflect the disproportionately beneficial impact of such tools in situations where linguistic barriers complicate caseworkers' ability to obtain relevant information from family members, or that easily interpretable risk scores are disproportionately influential when buy-in from frontline workers is high. Testing such hypotheses would require fundamentally different kinds of data based on ethnographic observations, interviews, or transcripts of caseworker discussions (Barley 1986; Miller and Maloney 2013; Brayne 2017; MacKenzie 2018; Pruss 2023). Future extensions of this research will pursue that line of inquiry and

specifically investigate the integration of computationally derived knowledge into existing CPS decision-making structures.

Additionally, this study does not examine the effects of PRM tools on one additional high-stakes intervention in family life: the court-sanctioned placement of maltreated children in the foster care system. This omission is necessary because foster care placements are not consistently reported in NCANDS Child Files, and conversations with NCANDS staff have revealed that available data are coded unevenly by CPS. It is still possible to link NCANDS files to the Adoption and Foster Care Analysis and Reporting System (AFCARS) using child-specific alphanumeric identifiers; however, for all children who experience more than one investigation (e.g. around 13 percent in the 2019 calendar year), AFCARS data cannot be used to reliably link foster care placements to specific investigations—which is a key requirement for the analyses presented above. But given the well-documented importance of such placement for health and educational outcomes (Doyle 2013), understanding the effects of PRM tools on foster care placements remains a key future research aim.

## Discussion

The routine use of PRM tools is increasingly central to the governance of social vulnerability and poverty in the "metric society" of the twenty-first century (Mau 2018; Brayne 2017). Theories of algorithmic governance alternatively emphasize the promise of those technological innovations or, more commonly in sociology, highlight their adverse effects on marginalized communities and social inequalities. In particular, sociologists have posited that variations in predictive accuracy may yield racially disparate impacts (Angwin et al. 2016). However, adjudicating between these competing perspectives is made difficult by a scarcity of generalizable findings and by competing and possibly incompatible conceptions of fairness and equality (Corbett-Davies et al. 2017). Recent empirical research has also found evidence of heterogeneous and small effects (Stevenson 2018; Imai et al. 2023), which further complicates strong general claims about the social significance of algorithmic tools in welfare systems that are already shaped by histories of systemic bias and often starved of necessary resources (Edwards and Wildeman 2018; DiMario 2022).

This study offers a partial consilience between optimistic and critical views on algorithmic governance and emphasizes the potentially disparate impacts of PRM tools in public administration. It finds that the adoption of PRM tools reduced the overall risk of maltreatment confirmation, reduced the risk of maltreatment confirmation among minority populations that have historically been over-surveilled and over-represented among system-involved children, and reduced the risk of state interventions among low-risk children. This suggests that PRM tools can have directionally favorable effects that reduce chronic but potentially unwarranted state interventions. However, the observed risk of maltreatment confirmations concurrently increased for a subset of ("high-risk") children with histories of prior CPS contact and a greater familial reliance on public assistance. On the one hand, this indicates that PRM tools can increase administrative attention on a particularly vul-

nerable subset of children. On the other hand, it also hints at the stratifying effects of algorithmic tools along the socio-economic spectrum and highlights the impact that PRM tools can have on the governance of poverty in an age of widespread precarity (Wacquant 2009; Eubanks 2017). More intensive state contact may not ultimately result in improved outcomes for those children. Results also point toward considerable effect heterogeneity across jurisdictions that cannot readily be explained with reference to local socio-demographic environments.

Analyses that focus on the impact of algorithmic tools on future maltreatment confirmations warrant a more cautious interpretation. However, across a wide range of model specifications, these analyses produce no evidence that PRM tools were effective at preventing such confirmations and reduced the chronic involvement of governing agencies in family life. Recent CPS practice already suggests as much. Since late 2018, a larger number of jurisdictions have phased out open case review tools than have newly adopted such tools, with reports and newspaper coverage showing that contract terminations are commonly justified by the limited long-term efficacy of PRM tools, rather than by budget constraints.[5] This assessment also accords with prior research on algorithmic governance that has highlighted the small magnitude of effects in other domains of public administration (Imai et al. 2023)—where PRM tools "may not have provided as large a gain in predictive power as expected" (Stevenson 2018:369)—and has found null effects on the prevention of recurring state contact (Parker et al. 2022).

More generally, these findings suggest an imperfect parallel to other administrative domains—like the criminal justice system—where concerns about intractable social inequalities and pervasive discrimination have spurred enthusiasm about technological innovations but have also generated pushback against the algorithmically-augmented surveillance of marginalized communities (Eubanks 2017; Stevenson 2018; Skeem, Scurich, and Monahan 2020). But although some studies of risk scoring in the criminal justice system have found evidence of racially stratifying effects, I do not observe such effects in the algorithmic governance of child welfare. Effect heterogeneity across jurisdictions also offers suggestive evidence that disparate impacts of PRM tools are not simply driven by differences in predictive accuracy, which has been the focus of research on criminal justice algorithms (Hamilton 2019; Kwegyir-Aggrey et al. 2023). Instead, such disparate impacts may depend most directly on how PRM tools are integrated into administrative decision-making at the local level (Stevenson 2018; Parker et al. 2022).

This holds important lessons for the sociological theorization of algorithmic governance. The routine use of risk scores by the "machine-learning state" (Cuéllar and Huq 2021) can—in the context of pervasive administrative bias—reduce disparities in system contact along one axis (across ethno-racial groups) but lead to a simultaneous stratification of state involvement in family life along another axis (by socio-economic status). As technologies of the governmental frontier, PRM tools can decisively shape the contemporary governance of social vulnerability through disparate impacts on administrative decision-making, attenuating some unwanted patterns in state interventions while exacerbating others and thereby affecting how race, ethnicity, and poverty correlate with state interventions in family life. Yet their significance for the long-term governance of chronic social vulnerabilities is

unproven. Recurring state interventions and persistent state-propagated inequities ultimately have no technological fixes.

## Notes

1 In 2019—the last pre-pandemic year—, the overall US poverty rate was around 12.3%.

2 ATT estimations were performed using the MatchIt and Marginaleffects packages in R.

3 No county adopted PRM open case review tools after January 2019, so analyses do not include any maltreatment reports filed after January 2020. This also prevents potential distortions introduced by the onset of the COVID-19 pandemic in March 2020.

4 RD analyses were performed using the Rdrobust and Rddtools packages in R.

5 See, for example: "Illinois child welfare to end use of predictive program." *AP News*, 12/06/2017. https://apnews.com/article/de09af9fdc8843b8ab2dbcff513d261c. Accessed 11/04/2023; "Oregon is dropping an artificial intelligence tool used in child welfare system. *NPR*, 06/02/2022. https://www.npr.org/2022/06/02/1102661376/oregon-drops-artificial-intelligence-child-abuse-cases. Accessed 11/04/2023; "Examination of Using Structured Decision-Making and Predictive Analytics in Assessing Safety and Risk in Child Welfare." *LA County Office of Child Protection*. Letter to the Board of Supervisors, 05/04/2017.

## References

Albert, Vicky N., and Richard P. Barth. 1996. "Predicting Growth in Child Abuse and Neglect Reports in Urban, Suburban, and Rural Counties." *Social Service Review* 70(1):58–82. https://doi.org/10.1086/604165

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Krichner. 2016. "Machine Bias." *ProPublica*, May 23, 2016. Retrieved November 9, 2023 (https://www.propublica.org/Article/Machine-Bias-Risk-Assessments-In-Criminal-Sentencing).

Ananny, Mike, and Kate Crawford. 2016. "Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability." *New Media & Society* 20(3):973–989. https://doi.org/10.1177/1461444816676645

Austin, Peter C., and Elizabeth A. Stuart. 2017. "Estimating the Effect of Treatment on Binary Outcomes Using Full Matching on the Propensity Score." *Statistical Methods in Medical Research* 26(6):2505–2525. https://doi.org/10.1177/0962280215601134

Barley, Stephen R. 1986. "Technology as an Occasion for Structuring: Evidence from Observations of CT Scanners and the Social Order of Radiology Departments." *Administrative Science Quarterly* 31(1):78–108. https://doi.org/10.2307/2392767

Baron, E. Jason, Joseph J. Doyle Jr, Natalia Emanuel, Peter Hull, and Joseph P. Ryan. 2024. "Discrimination in Multi-Phase Systems: Evidence from Child Protection." National Bureau of Economic Research Working Paper No. W31490. https://doi.org/10.3386/w31490

Beebe, Rebecca, Meghan C. Fish, Damion Grasso, Bruce Bernstein, Susan DiVietro, and Carla Smith Stover. 2023. "Reducing Family Violence Through Child Welfare Intervention: A Propensity Score-matched Study of Fathers for Change." *Journal of Interpersonal Violence* 38(21-22):11666–11691. https://doi.org/10.1177/08862605231186121

Bigman, Yochanan E., Desman Wilson, Mads N. Arnestad, Adam Waytz, and Kurt Gray. 2023. "Algorithmic Discrimination Causes Less Moral Outrage Than Human Discrimination." *Journal Of Experimental Psychology: General* 152(1):4–27. https://doi.org/10.1037/xge0001250

Brame, Robert, Shawn D. Bushway, Ray Paternoster, and Michael G. Turner. 2014. "Demographic Patterns of Cumulative Arrest Prevalence by Ages 18 and 23." *Crime & Delinquency* 60(3): 471–486. https://doi.org/10.1177/0011128713514801

Brantingham, P. Jeffrey, Matthew Valasik, and George O. Mohler. 2018. "Does Predictive Policing Lead to Biased Arrests? Results From a Randomized Controlled Trial." *Statistics and Public Policy* 5(1):1–6. https://doi.org/10.1080/2330443X.2018.1438940

Brayne, Sarah. 2017. "Big Data Surveillance: The Case of Policing." *American Sociological Review* 82(5):977–1008. https://doi.org/10.1177/0003122417725865

Brayne, Sarah, and Angèle Christin. 2021. "Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts." *Social Problems* 68(3):608–624. https://doi.org/10.1093/socpro/spaa004

Brown, Anna, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. "Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-Making in Child Welfare Services." *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*:1–12. https://doi.org/10.1145/3290605.3300271

Burrell, Jenna. 2016. "How The Machine 'Thinks:' Understanding Opacity In Machine Learning Algorithms." *Big Data & Society* 3(1):1–12. https://doi.org/10.1177/2053951715622512

Burrell, Jenna, and Marion Fourcade. 2021. "The Society of Algorithms." *Annual Review of Sociology* 47:213–237. https://doi.org/10.1146/annurev-soc-090820-020800

Burton, Jason W., Mari-Klara Stein, and Tina Blegind Jensen. 2020. "A Systematic Review of Algorithm Aversion in Augmented Decision Making." *Journal of Behavioral Decision Making* 33(2):220–239. https://doi.org/10.1002/bdm.2155

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82(6):2295–2326. https://doi.org/10.3982/ECTA11757

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2015. "Rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs." *The R Journal* 7(1):38–51. https://doi.org/10.32614/RJ-2015-004

Cheng, Hao-Fei, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. "How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions." *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*:1–22. https://doi.org/10.1145/3491102.3501831

Christin, Angèle. 2018. "Predictive algorithms and criminal sentencing." Pp. 272–294 in: *The Decisionist Imagination: Sovereignty, Social Science and Democracy in the 20th Century*, edited by Daniel Bessner and Nicolas Guilhot. New York: Berghahn. https://doi.org/10.2307/j.ctvw04b7q.14

Christin, Angèle. 2020. "The Ethnographer And The Algorithm: Beyond The Black Box." *Theory And Society* 49(5-6):897–918. https://doi.org/10.1007/s11186-020-09411-3

Church, Christopher E., and Amanda J. Fairchild. 2017. "In Search of a Silver Bullet: Child Welfare's Embrace of Predictive Analytics." *Juvenile And Family Court Journal* 68 (1): 67–81. https://doi.org/10.1111/jfcj.12086

Committee on Child Abuse and Neglect. 2002. "When Inflicted Skin Injuries Constitute Child Abuse." *Pediatrics* 110(3):644–45 https://doi.org/10.1542/peds.110.3.644

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. "Algorithmic Decision Making And The Cost Of Fairness." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*:797–806. https://doi.org/10.1145/3097983.3098095

Cuccaro-Alamin, Stephanie, Regan Foust, Rhema Vaithianathan, and Emily Putnam-Hornstein. 2017. "Risk Assessment And Decision Making in Child Protective Services: Predictive Risk Modeling in Context." *Children and Youth Services Review* 79:291–298. https://doi.org/10.1016/j.childyouth.2017.06.027

Cuellar, Maria. 2023. "Proposer of the vote of thanks and contribution to the Discussion of 'Experimental evaluation of algorithm- assisted human decision-making: application to pretrial public safety assessment' by Imai et al." *Journal of the Royal Statistical Society Series A: Statistics in Society* 186:190–191. https://doi.org/10.1093/jrsssa/qnad011

Cuéllar, Mariano-Florentino, and Aziz Z. Huq. 2021. "Privacy's Political Economy and the State of Machine Learning: An Essay in Honor of Stephen J. Schulhofer." *N.Y.U. Annual Survey of American Law* 76:317–354.

Daley, Dyann, Michael Bachmann, Brittany A. Bachmann, Christian Pedigo, Minh-Thuy Bui, and Jamye Coffman. 2016. "Risk Terrain Modeling Predicts Child Maltreatment." *Child Abuse & Neglect* 62:29–38. https://doi.org/10.1016/j.chiabu.2016.09.014

DiMario, Anthony. 2022. "To Punish, Parent, or Palliate: Governing Urban Poverty Through Institutional Failure." *American Sociological Review* 87 (5):860–888. https://doi.org/10.1177/00031224221116145

Doyle, Joseph J. 2013. "Causal Effects of Foster Care: An Instrumental-Variables Approach." *Children and Youth Services Review* 35(7):1143–1151. https://doi.org/10.1016/j.childyouth.2011.03.014

Drake, Brett, Sang Moo Lee, and Melissa Jonson-Reid. 2009. "Race and Child Maltreatment Reporting: Are Blacks Overrepresented?" *Children And Youth Services Review* 31(3):309–316. https://doi.org/10.1016/j.childyouth.2008.08.004

Drake, Brett, and Shanta Pandey. 1996. "Understanding the Relationship Between Neighborhood Poverty and Specific Types of Child Maltreatment." *Child Abuse & Neglect* 20(11):1003–1018. https://doi.org/10.1016/0145-2134(96)00091-9

Dressel, Julia, and Hany Farid. 2018. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances* 4(1):Eaao5580. https://doi.org/10.1126/sciadv.aao5580

Duwe, Grant, and Kideuk Kim.2017. "Out With the Old and in With the New? An Empirical Comparison of Supervised Learning Algorithms to Predict Recidivism." *Criminal Justice Policy Review* 28(6):570–600. https://doi.org/10.1177/0887403415604899

Edwards, Frank, and Christopher Wildeman. 2018. "Characteristics of the Front-Line Child Welfare Workforce." *Children And Youth Services Review* 89:13–26. https://doi.org/10.1016/j.childyouth.2018.04.013

Edwards, Frank, Sara Wakefield, Kieran Healy, and Christopher Wildeman. 2021. "Contact With Child Protective Services is Pervasive but Unequally Distributed by Race and Ethnicity in Large US Counties." *Proceedings of the National Academy of Sciences* 118(30):E2106272118. https://doi.org/10.1073/pnas.2106272118

Eubanks, Virginia. 2006. "Technologies of Citizenship: Surveillance and Political Learning in the Welfare System. Pp. 89-108 in: *Surveillance & Security*, edited by Torin Monahan. London: Routledge.

Eubanks, Virginia. 2017. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Flores, Anthony W., Kristin Bechtel, and Christopher T. Lowenkamp. 2016. "False Positives, False Negatives, and False Analyses: A Rejoinder To 'Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks.'" *Federal Probation Journal* 80(2):38–46.

Fong, Kelley. 2020. "Getting Eyes in the Home: Child Protective Services Investigations and State Surveillance of Family Life." *American Sociological Review* 85(4):610–638. https://doi.org/10.1177/0003122420938460

Font, Sarah A., and Lawrence M. Berger. 2015. "Child Maltreatment and Children's Developmental Trajectories in Early to Middle Childhood." *Child Development* 86(2):536–556. https://doi.org/10.1111/cdev.12322

Garrett, Brandon L., and John Monahan. 2020. "Judging risk." *California Law Review* 108:439–493.

Gaudin, James. 1995. Child Neglect: A Guide for Intervention. Washington, DC: US Department of Health and Human Services.

Gillespie, Tarleton. 2013. "The Relevance of Algorithms." Pp. 167-194 in: *Media Technologies*, edited by Tarleton Gillespie, Pablo Boczkowski, and Kirsten Foot. Cambridge: MIT Press.

Glisson, Charles, Philip Green, and Nathaniel J. Williams. 2012. "Assessing the organizational social context (OSC) of child welfare systems: Implications for research and practice." *Child Abuse & Neglect* 36(9):621–632. https://doi.org/10.1016/j.chiabu.2012.06.002

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics* 225(2):254–277. https://doi.org/10.1016/j.jeconom.2021.03.014

Green, Ben, and Yiling Chen. 2019. "Disparate Interactions: An Algorithm-in-the-loop Analysis of Fairness in Risk Assessments." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*: 90–99. https://doi.org/10.1145/3287560.3287563

Green, Ben, and Yiling Chen. 2021. "Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts." *Proceedings of the ACM on Human-Computer Interaction* 5, No. CSCW2:1–33. https://doi.org/10.1145/3479562

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica* 69 (1): 201–209. https://doi.org/10.1111/1468-0262.00183

Hamilton, Melissa. 2019. "The Biased Algorithm: Evidence of Disparate Impact on Hispanics." *American Criminal Law Review* 56:1553–1577.

Hannah-Moffat, Kelly. 2018. "Algorithmic Risk Governance: Big Data Analytics, Race and Information Activism in Criminal Justice Debates." *Theoretical Criminology* 23(4):453–470. https://doi.org/10.1177/1362480618763582

Harcourt Bernard E. 2006. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Chicago: The University of Chicago Press. https://doi.org/10.7208/chicago/9780226315997.001.0001

Hausman, Catherine, and David S. Rapson. 2018. "Regression Discontinuity in Time: Considerations for Empirical Applications." *Annual Review of Resource Economics* 10:533–552. https://doi.org/10.1146/annurev-resource-121517-033306

Hellman, Deborah. 2020. "Measuring Algorithmic Fairness." *Virginia Law Review* 106(4):811–866.

Hill, Jennifer, and Jerome P. Reiter. 2006. “Interval Estimation for Treatment Effects Using Propensity Score Matching.” *Statistics In Medicine* 25(13):2230–2256. https://doi.org/10.1002/sim.2277

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference.” *Political Analysis* 15(3):199–236. https://doi.org/10.1093/pan/mpl013

Hussey, Jon M., Jen Chang, and Jonathan B. Kotch. 2006. "Child Maltreatment in the United States: Prevalence, Risk Factors, and Adolescent Health Consequences." *Pediatrics* 118(3):933–942. https://doi.org/10.1542/peds.2005-2452

Imai, Kosuke, Zhichao Jiang, D. James Greiner, Ryan Halen, and Sooahn Shin. 2023. "Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." *Journal of the Royal Statistical Society Series A: Statistics in Society* 186(2):167–189. https://doi.org/10.1093/jrsssa/qnad010

Imbens, Guido, and Thomas Lemieux. 2008. “Regression Discontinuity Designs: A Guide to Practice.” *Journal of Econometrics* 142:615–635. https://doi.org/10.1016/j.jeconom.2007.05.001

Ito, Koichiro. 2015. "Asymmetric Incentives in Subsidies: Evidence from a Large-Scale Electricity Rebate Program." *American Economic Journal: Economic Policy* 7(3):209–237. https://doi.org/10.1257/pol.20130397

Jaffee, Sara R. 2017. "Child Maltreatment and Risk for Psychopathology in Childhood and Adulthood." *Annual Review of Clinical Psychology* 13:525–551. https://doi.org/10.1146/annurev-clinpsy-032816-045005

Joyce, Kelly, Laurel Smith-Doerr, Sharla Alegria, Susan Bell, Taylor Cruz, Steve G. Hoffman, Safiya Noble, and Benjamin Shestakofsky. 2021. “Toward a Sociology of Artificial Intelligence: A Call for Research on Inequalities and Structural Change.” *Socius* 7:1–11. https://doi.org/10.1177/2378023121999581

Katzenbach, Christian, and Lena Ulbricht. 2019. "Algorithmic Governance." *Internet Policy Review* 8(4):1–18. https://doi.org/10.14763/2019.4.1424

Kim, Hyunil, Christopher Wildeman, Melissa Jonson-Reid, and Brett Drake. 2017. "Lifetime Prevalence of Investigating Child Maltreatment Among US Children." *American Journal of Public Health* 107(2):274–280. https://doi.org/10.2105/AJPH.2016.303545

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. 2018. "Discrimination in the Age of Algorithms." *Journal of Legal Analysis* 10:113–174. https://doi.org/10.1093/jla/laz001

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." Arxiv Preprint:1609.05807.

Kwegyir-Aggrey, Kweku, Marissa Gerchick, Malika Mohan, Aaron Horowitz, and Suresh Venkatasubramanian. 2023. "The Misuse of AUC: What High Impact Risk Assessment Gets Wrong." *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*:1570–1583. https://doi.org/10.1145/3593013.3594100

Legewie, Joscha. 2016. "Racial Profiling and Use of Force in Police Stops: How Local Events Trigger Periods of Increased Discrimination." *American Journal of Sociology* 122(2):379–424. https://doi.org/10.1086/687518

Liang, Kung-Yee, and Scott L. Zeger. 1986. “Longitudinal Data Analysis Using Generalized Linear Models.” *Biometrika* 73(1):13–22. https://doi.org/10.1093/biomet/73.1.13

Miller, Joel, and Carrie Maloney. 2013. "Practitioner Compliance with Risk/Needs Assessment Tools: A Theoretical and Empirical Assessment." *Criminal Justice and Behavior* 40(7):716–736. https://doi.org/10.1177/0093854812468883

Lin, Zhiyuan, Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. "The Limits of Human Predictions of Recidivism." *Science Advances* 6(7):Eaaz0652. https://doi.org/10.1126/sciadv.aaz0652

Mackenzie, Donald. 2018. "Material Signals: A Historical Sociology Of High-Frequency Trading." *American Journal of Sociology* 123(6):1635–1683. https://doi.org/10.1086/697318

Maguire-Jack, Kathryn, Paul Lanier, Michelle Johnson-Motoyama, Hannah Welch, and Michael Dineen. 2015. "Geographic Variation in Racial Disparities in Child Maltreatment: The Influence of County Poverty and Population Density." *Child Abuse & Neglect* 47:1–13. https://doi.org/10.1016/j.chiabu.2015.05.020

Maguire-Jack, Kathryn, and Sarah A. Font. 2017. "Community and individual risk factors for physical child abuse and child neglect: Variations by poverty status." *Child Maltreatment* 22(3):215–226. https://doi.org/10.1177/1077559517711806

Maguire-Jack, Kathryn, Sarah A. Font, and Rebecca Dillard. 2020. "Child Protective Services Decision-Making: The Role of Children's Race and County Factors." *American Journal of Orthopsychiatry* 90(1):48–62. https://doi.org/10.1037/ort0000388

Marshall, David B., and Diana J. English. 2000. "Neural Network Modeling of Risk Assessment in Child Protective Services." *Psychological Methods* 5(1):102–124. https://doi.org/10.1037/1082-989X.5.1.102

Mau, Steffen. 2018. *The Metric Society: On The Quantification of the Social*. Cambridge: Polity.

McNellan, Claire R., Daniel J. Gibbs, Ann S. Knobel, and Emily Putnam-Hornstein. 2022. "The Evidence Base for Risk Assessment Tools Used in US Child Protection Investigations: A Systematic Scoping Review." *Child Abuse & Neglect* 134:105887. https://doi.org/10.1016/j.chiabu.2022.105887

Meijer, Albert, and Martijn Wessels. 2019. "Predictive Policing: Review of Benefits and Drawbacks." *International Journal of Public Administration* 42(12):1031–1039. https://doi.org/10.1080/01900692.2019.1575664

Morozov, Evgeny. 2013. *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: Public Affairs.

Neil, Roland, and Robert J. Sampson. 2021. "The Birth Lottery of History: Arrest Over the Life Course of Multiple Cohorts Coming of Age, 1995–2018." *American Journal of Sociology* 126(5):1127–1178. https://doi.org/10.1086/714062

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money And Information*. Cambridge: Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Parker, Elizabeth M., Jason R. Williams, Peter J. Pecora, and Daniel Despard. 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." *Child Abuse & Neglect* 133:105856. https://doi.org/10.1016/j.chiabu.2022.105856

Pavlou, Menelaos, Gareth Ambler, Shaun R. Seaman, Oliver Guttmann, Perry Elliott, Michael King, and Rumana Z. Omar. 2015. "How to develop a more accurate risk prediction model when there are few events." *The BMJ* 351:h3868. https://doi.org/10.1136/bmj.h3868

Pruss, Dasha. 2023. "Ghosting the machine: Judicial resistance to a recidivism risk assessment instrument." *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*:312–323. https://doi.org/10.1145/3593013.3593999

Putnam-Hornstein, Emily, Eunhye Ahn, John Prindle, Joseph Magruder, Daniel Webster, and Christopher Wildeman. 2021. "Cumulative Rates of Child Protection Involvement and

Terminations of Parental Rights in a California Birth Cohort, 1999–2017." *American Journal of Public Health* 111(6):1157–1163. https://doi.org/10.2105/AJPH.2021.306214

Puzzanchera, Charles. 2021. "Juvenile Arrests, 2019." U.S. Department of Justice, Office of Justice Programs. Retrieved February 02, 2024 (https://ojjdp.ojp.gov/publications/juvenile-arrests-2019.pdf.)

Rittenhouse, Katherine, Emily Putnam-Hornstein, and Rhema Vaithianathan. 2022. "Algorithms, Humans, and Racial Disparities in Child Protective Services: Evidence from the Allegheny Family Screening Tool." Working Paper.

Roberts, Dorothy E. 2022. *Torn Apart: How The Child Welfare System Destroys Black Families—And How Abolition Can Build A Safer World*. New York: Basic Books.

Rosen, Eva, Philip M.E. Garboden, and Jennifer E. Cossyleon. 2021. "Racial Discrimination in Housing: How Landlords Use Algorithms and Home Visits to Screen Tenants." *American Sociological Review* 86(5):787–822. https://doi.org/10.1177/00031224211029618

Rudin, Cynthia, Caroline Wang, and Beau Coker. 2020. "The Age of Secrecy and Unfairness in Recidivism Prediction." *Harvard Data Science Review* 2(1):2–55. https://doi.org/10.1162/99608f92.6ed64b30

Ruscio, John. 1998. "Information Integration in Child Welfare Cases: An Introduction to Statistical Decision Making." *Child Maltreatment* 3(2):143–156. https://doi.org/10.1177/1077559598003002008

Russell, Jesse. 2015. "Predictive Analytics and Child Protection: Constraints and Opportunities." *Child Abuse & Neglect* 46:182–189. https://doi.org/10.1016/j.chiabu.2015.05.022

Samant, Anjana, Aaron Horowitz, Kath Xu, and Sophie Beiers. 2021. "Family Surveillance by Algorithm: The Rapidly Spreading Tools Few Have Heard Of." *American Civil Liberties Union*, September 28, 2021. Retrieved October 2, 2023 (https://www.aclu.org/news/womens-rights/family-surveillance-by-algorithm-the-rapidly-spreading-tools-few-have-heard-of).

Sävje, Fredrik, Michael J. Higgins, and Jasjeet S. Sekhon. 2021. "Generalized Full Matching." *Political Analysis* 29(4):423–447. https://doi.org/10.1017/pan.2020.32

Saxena, Devansh, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2020. "A Human-Centered Review of Algorithms Used Within the US Child Welfare System." *Proceedings of the 2020 CHI Conference On Human Factors In Computing Systems*:1–15. https://doi.org/10.1145/3313831.3376229

Schafer, Joseph L., and Joseph Kang. 2008. "Average Causal Effects from Nonrandomized Studies: A Practical Guide and Simulated Example." *Psychological Methods* 13(4):279–313. https://doi.org/10.1037/a0014268

Schwartz, Ira M., Peter York, Eva Nowakowski-Sims, and Ana Ramos-Hernandez. 2017. "Predictive and Prescriptive Analytics, Machine Learning and Child Welfare Risk Assessment: The Broward County Experience." *Children and Youth Services Review* 81:309–320. https://doi.org/10.1016/j.childyouth.2017.08.020

Skeem, Jennifer, Nicholas Scurich, and John Monahan. 2020. "Impact of Risk Assessment on Judges' Fairness in Sentencing Relatively Poor Defendants." *Law and Human Behavior* 44(1):51–59. https://doi.org/10.1037/lhb0000360

Snowden, Jonathan M., Sherri Rose, and Kathleen M. Mortimer. 2011. "Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique." *American Journal of Epidemiology* 173(7):731–738. https://doi.org/10.1093/aje/kwq472

Solove, Daniel J. 2002. "Conceptualizing Privacy." *California Law Review* 90(4):1087–1155. https://doi.org/10.2307/3481326

Soss, Joe, Richard C. Fording, and Sanford Schram. 2011. *Disciplining The Poor: Neoliberal Paternalism And The Persistent Power Of Race*. Chicago: The University of Chicago Press. https://doi.org/10.7208/chicago/9780226768786.001.0001

Starr, Sonja B. 2014. "Evidence-Based Sentencing and the Scientific Rationalization of Discrimination." *Stanford Law Review* 66(4):803–872.

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

Stevenson, Megan and Jennifer L. Doleac. 2022. "Algorithmic Risk Assessment in the Hands of Humans." SSRN Working Paper. Retrieved June 28, 2024 (http://dx.doi.org/10.2139/ssrn.3489440).

Stuart, Elizabeth A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science* 25(1):1–21. https://doi.org/10.1214/09-STS313

Sun, Matthew, and Marissa Gerchick. 2019. "The Scales of (Algorithmic) Justice: Tradeoffs and Remedies." *AI Matters* 5(2):30–40. https://doi.org/10.1145/3340470.3340478

Vaithianathan, Rhema, Emily Putnam-Hornstein, Alexandra Chouldechova, Diana Benavides-Prado, and Rachel Berger. 2020. "Hospital Injury Encounters of Children Identified By a Predictive Risk Model For Screening Child Maltreatment Referrals: Evidence From the Allegheny Family Screening Tool." *JAMA Pediatrics* 174(11):E202770-E202770. https://doi.org/10.1001/jamapediatrics.2020.2770

Wacquant, Loïc. 2009. *Punishing the Poor: The Neoliberal Government of Social Insecurity*. Durham: Duke University Press. https://doi.org/10.1215/9780822392255

Wang, Caroline, Bin Han, Bhrij Patel, and Cynthia Rudin. 2023. "In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction." *Journal Of Quantitative Criminology* 39(2):519–581. https://doi.org/10.1007/s10940-022-09545-w

Yi, Youngmin, Frank R. Edwards, and Christopher Wildeman. 2020. "Cumulative Prevalence of Confirmed Maltreatment and Foster Care Placement for US Children by Race/Ethnicity, 2011–2016." *American Journal of Public Health* 110(5):704–709. https://doi.org/10.2105/AJPH.2019.305554

Zajko, Mike. 2021. "Conservative AI and Social Inequality: Conceptualizing Alternatives to Bias Through Social Theory." *AI & Society* 36(3):1047–1056. https://doi.org/10.1007/s00146-021-01153-9

Zeng, Jiaming, Berk Ustun, and Cynthia Rudin. 2017. "Interpretable Classification Models for Recidivism Prediction." *Journal of the Royal Statistical Society Series A: Statistics In Society* 180(3):689–722. https://doi.org/10.1111/rssa.12227

**Acknowledgements:** The author thanks Olivia Kim and Henry Zapata for invaluable research assistance, and thanks Garrett Baker, Alexandra Gibbons, Sarah Sernaker, and Christopher Wildeman for constructive feedback.

**Martin Eiermann:** Department of Sociology, Duke University.
E-mail: martin.eiermann@duke.edu.