Human Decisions and Machine Predictions in Multistage Systems*

Jason Baron, Duke University and NBER
Will Dobbie, Harvard University and NBER
Richard Lombardo, Harvard University
Ashesh Rambachan, MIT
Joseph Ryan, University of Michigan

March 2026

Abstract

Many high-stakes decisions unfold through multistage systems, where cases pass through successive layers of review before a final decision is made. Yet predictive algorithms typically target a single stage in the decision-making process, ignoring how interventions at one stage reshape the downstream flow of cases. We develop new empirical tools to evaluate predictive algorithms within these multistage systems. We decompose the gap between observed outcomes and algorithm benchmarks into stage-specific contributions, revealing where predictive algorithms can add the most value. We apply these tools to the child protection system, where incoming allegations are reviewed by a hotline screener before being assigned to an investigator who gathers additional information and decides whether to open a case. Leveraging the quasi-random assignment of both screeners and investigators, we find that a system-wide algorithm would modestly reduce child maltreatment while maintaining the same case opening rate and completely eliminating invasive investigations. The algorithm outperforms all screeners at the initial screening stage but underperforms the top 15 percent of investigators at the final stage, suggesting that the most skilled investigators uncover valuable information during field visits that algorithms cannot observe. Before the summer institute, we will apply language models to the detailed case notes recorded by investigators to discover the features that are most predictive of child safety outcomes and understand the sources of private information that distinguish high- from low-performing investigators.

JEL Codes: C01, C45, D80, I31, I38.

*Preliminary and Incomplete. Please do not cite or distribute.

1 Introduction

Many high-stakes decisions are organized into multistage systems: a case advances through successive layers of review before a final decision is made, with later stages collecting information unavailable at the outset. In hiring, a recruiter decides whether to advance an applicant for an interview, where the interviewer collects additional information before extending an offer. In child protection, a hotline screener decides whether to refer an incoming allegation for investigation, where an investigator visits the family, gathers additional information, and decides whether to open a case. The final decision emerges from the sequence of choices made by individuals across all stages.

Human decision-makers, however, are often biased and noisy (Tversky and Kahneman, 1974; Kahneman, 2011; Kahneman, Sibony and Sunstein, 2021), and their resulting mistakes are particularly consequential in these high-stakes settings. This has spurred the widespread adoption of predictive algorithms across criminal justice, medicine, hiring, and child protection. A growing body of research suggests that these tools can improve decisions by predicting more accurately than existing decision-makers (e.g., Kleinberg et al., 2018; Mullainathan and Obermeyer, 2022; Rambachan, 2024; Angelova, Dobbie and Yang, 2025). Yet existing algorithms target a single decision, ignoring the multistage structure within which many of these decisions sit. The pool of cases reaching later stages — and the information available there — is shaped by choices made earlier. A mistake at one stage can be corrected or compounded by the next, with the cost of human errors depending on what happens downstream. These interactions shape the comparison between humans and algorithms and determine where in a multistage system algorithms add the most value.

Measuring how predictive algorithms can improve human decision-making in multistage systems is challenging for several reasons. Outcomes are selectively observed, a familiar missing data problem that is compounded here because selection occurs at every stage and each stage inherits a selected population from the last. Datasets linking all stages are rare, and rarer still are settings with the exogenous variation needed to evaluate human decisions and algorithms at each stage. Even with such data, interpreting stage-specific comparisons between humans and algorithms and integrating them at the system-level requires new analytic tools.

In this paper, we develop empirical tools for measuring the value of predictive algorithms in multistage systems and apply them to study case-opening decisions in child protective services (CPS). Contact with CPS is surprisingly common in the U.S.: 37% of children are the subject of a maltreatment investigation by age 18 (Kim et al., 2017). CPS aims to prevent child maltreatment by investigating reported allegations of abuse or neglect, making it

a canonical prediction policy problem (Kleinberg et al., 2015; Mullainathan, 2025) in which costly interventions must be targeted toward the highest-risk cases. Consequently, predictive algorithms have attracted substantial interest as a tool in this setting. For example, recent federal guidance encourages states to integrate predictive risk models into child welfare decision-making to support more consistent and data-driven screening and investigation decisions (Administration for Children, Youth and Families, 2026).

The CPS system operates in two stages. In an initial screening stage, incoming allegations of child maltreatment are routed to a centralized state-level hotline, where call screeners decide whether to "screen-in" the call for investigation. In a subsequent investigation stage, screened-in calls are allocated to a regional investigator who visits the family, gathers additional information, and decides whether to open a case. Errors by call screeners can be corrected or compounded by investigators. Predictive algorithms could target the screening stage, the investigation stage, or both. Existing work, however, like the broader literature on humans and algorithms, targets one decision at a time.[1]

We develop three comparisons of human-algorithm ranking quality that capture how well humans identify the highest-risk cases relative to an equally selective algorithm. The system-wide measure asks: if an algorithm replaced the entire CPS system, using only information available at the hotline call and opening as many cases as the status quo, how would subsequent maltreatment change among unopened cases? The screening-stage and investigation-stage measures ask the same question for each stage alone, holding fixed the screen-in rate and the case-opening rate among investigated cases. By holding action rates fixed, each measure isolates the comparison of ranking quality by existing decision-makers relative to the algorithm. These three measures build directly on Kleinberg et al. (2018) and Angelova, Dobbie and Yang (2025), who compare misconduct rates among released defendants in pretrial settings, holding the release rate fixed.

Our main conceptual result is a decomposition of the system-wide comparison into stage-wise ranking quality gaps attributable to screeners and investigators, plus two residual terms that depend on the structure of the multistage system. The first residual term is a selection effect: because screeners determine who investigators see, the system-wide and investigation-stage benchmarks evaluate different populations. When screeners pass lower-risk cases to investigators, the system may partially correct this mistake downstream, which would lead the system-wide gap to be smaller than the stage-wise screening gap. The second is a threshold adjustment: stage-specific benchmarks evaluate performance at the action rates of each stage, which are relatively high (e.g., the screen-in rate or the case-opening rate

[1]See, for example, Fitzpatrick, Sadowski and Wildeman (2025); Grimon and Mills (2025); Putnam-Hornstein et al. (2022); Rittenhouse, Putnam-Hornstein and Vaithianathan (2025).

among investigated cases). In contrast, the system-wide benchmark must match the much lower overall case-opening rate. This forces the system-wide algorithmic benchmark to a stricter threshold, leaving behind a higher-risk non-opened pool — which closes the gap relative to humans. We show that these residuals imply that the system-wide comparison can substantially understate how much humans underperform the algorithm at each individual stage — and can even reverse sign.

The core identification challenge is the selective labels problem (Lakkaraju et al., 2017; Kleinberg et al., 2018): subsequent maltreatment is observed only for children whose cases are not opened. Evaluating the algorithm requires knowing what would have happened to children in opened cases had no case been opened — precisely because the algorithm may wish to not open some of the cases that the existing system did. Estimating the ranking quality comparisons therefore requires recovering these missing counterfactuals. Following Baron et al. (2024), we exploit the quasi-random assignment of screeners and investigators within each decision stage. Because observationally similar calls are handled by screeners and investigators with different action rates, we can trace out how subsequent maltreatment varies with the probability of case opening and extrapolate to the counterfactual in which no case is opened — the identification-at-infinity approach of Arnold, Dobbie and Hull (2022); Angelova, Dobbie and Yang (2025). We also complement these extrapolation estimates with the worst-case bounds of Rambachan (2024); Rambachan, Coston and Kennedy (2025). Together, these approaches allow us to estimate all components of the decomposition — the system-wide and stage-wise ranking quality comparisons and the two system-wide residuals.

We apply this framework to Michigan's child protective services system using administrative records covering the universe of hotline calls and investigations from 2017 to 2022. Our analysis sample includes 524,657 unique calls involving 392,691 children, handled by 292 call screeners and 8,753 investigators. We train a predictive algorithm using the subset of calls where no case was opened, predicting subsequent maltreatment – proxied by a new allegation within 30 to 180 days – based only on information available at the time of the hotline call. The algorithm demonstrates meaningful predictive power in a held-out test sample (Appendix Figure A.3) and serves as the counterfactual benchmark throughout our analysis.

We exploit the quasi-random assignment of screeners and investigators in Michigan CPS. Screeners are quasi-randomly assigned to calls within day-by-shift cells; investigators within month-by-ZIP-code-by-shift cells. Balance checks confirm that observable case characteristics are uncorrelated with decision-maker assignment within these cells (Appendix Figures A.1–A.2). We use this variation to extrapolate subsequent maltreatment rates to the counterfactual threshold at which no case is opened. Figure 2 shows the extrapolation fits for the

screener and investigator samples: the implied intercepts recover the base rate of subsequent maltreatment among all incoming calls (Panel A) and among screened-in calls (Panel B) — the direct inputs to the system-wide and stage-wise ranking quality comparisons. Figure 3 then plots the algorithm's estimated subsequent maltreatment rate across decision thresholds, showing how its performance varies depending on whether it is applied to all incoming calls or only to those that reach the investigation stage.

We report three main empirical results, each corresponding to one of our human-algorithm comparisons. The first is the system-wide comparison, Δ^{CPS}. This counterfactual is deliberately extreme: the algorithm replaces the entire two-stage system, eliminating all investigations and the information gathered during them, opening cases using only call-time characteristics at the same overall rate. We estimate that this replacement reduces subsequent maltreatment among unopened cases by 1.2 percentage points against a base rate of approximately 33.8% (Table 2). These findings suggest that the investigation stage — the most resource-intensive component of the child protective system — could be entirely replaced by an algorithm using only call-time information without negatively impacting child safety.

Our second result decomposes this system-wide gain into stage-specific contributions (Table 2). The investigation-stage comparison asks how child safety would change if investigators were replaced by the algorithm among screened-in cases, opening the same number of cases. We estimate a gain of 2.9 percentage points. The screening-stage comparison asks how child safety would change if screeners were replaced by the algorithm at the same screen-in rate. We estimate a gain of 6.7 percentage points. Together, the stage-wise gains sum to 9.6 percentage points — eight times the system-wide gain. The multistage residuals account for this gap: they reflect features of the multistage system itself — how selection by screeners shapes the pool reaching later stages, and how variation in decision rates across stages compresses the system-wide comparison. While the algorithm offers large gains at each individual stage, the system-wide comparison substantially understates them. The system-wide counterfactual — replacing the entire pipeline with an algorithm — is therefore an incomplete guide to policy: it can obscure where performance gaps are largest. Our decomposition makes these margins explicit, revealing which stages contribute most to inefficiency and where algorithms are best positioned to substitute for human judgment or complement it.

Our final result explores heterogeneity across decision-makers within each stage. We further decompose each stage-wise comparison into a weighted average of individual decision-maker ranking quality — how well each screener or investigator ranks cases relative to the algorithm at their own action rate — plus average leniency and composition residual terms. We summarize the distribution of individual ranking quality within each stage by its mean and standard deviation, estimating these objects using empirical Bayes techniques (Efron,

2016; Bonhomme and Weidner, 2022).

At the screening stage, we find that hotline screeners are uniformly worse than the algorithm: no screener outperforms it (Table 3). Furthermore, there is very little heterogeneity across screeners (Figure 4, Panel A). Hotline screeners systematically screen in lower-risk cases than the algorithm would select. Before the summer institute, we will report additional results to understand the sources of these systematic errors (e.g., Mullainathan and Obermeyer, 2022; Rambachan, 2024; Angelova, Dobbie and Yang, 2025).

At the investigation stage, the picture is different. While investigators are on average also worse than the algorithm (Figure 4, Panel B), there is far greater heterogeneity: the standard deviation of individual ranking quality is roughly the same magnitude as its mean, and approximately 15% of investigators perform significantly better than the algorithm (Table 4). These higher-skilled investigators may be gathering information during visits — through observation, interviews, and field judgment — that meaningfully improves on call-level characteristics. To understand what distinguishes high- from low-quality investigator decisions, we will apply language models and hypothesis generation techniques to the unstructured case notes collected during investigations (Ludwig and Mullainathan, 2024; Movva et al., 2025), discovering the sources of private information that distinguish high- from low-performing investigators. This approach — using AI to identify the sources of private information in human-algorithmic comparisons — may prove valuable more broadly in settings where decision-makers gather unstructured information that may be relevant to the decision at hand. We anticipate that these results will be ready to present before the summer institute.

These empirical findings speak to a broader methodological challenge addressed by our analysis. Most existing work evaluates algorithms at a single stage in isolation, treating it as independent of the decisions that precede and follow it. Our results show that in multistage systems, this can be misleading: changing one stage reshapes the pool and context facing the next, so stage-level improvements do not translate one-for-one into system-wide gains. Evaluating algorithms using a counterfactual that replaces the entire system may mask substantial inefficiencies at individual stages. Our framework provides a unified approach for evaluating algorithms in multistage systems that links stage-specific performance to system-wide outcomes. This allows researchers and policymakers to move beyond both single-stage and all-or-nothing system comparisons and instead evaluate targeted interventions within multistage systems.

This paper is a work-in-progress. The remainder proceeds as follows. Section 2 describes the conceptual framework underlying our analysis. Section 3 describes the setting and data. Section 4 outlines our empirical approach. We present the preliminary tables and figures for our empirical results.

2 Predictive Algorithms in a Multistage System

We develop a framework for evaluating predictive algorithms in multistage decision systems — settings in which cases pass through successive stages of review and outcomes are observed only for cases that do not receive the final action. The framework applies broadly: in hiring, a resume screen selects candidates for interviews that inform a final offer; in medicine, triage determines whether a patient receives further evaluation that guides treatment. We describe the framework in the context of child protective services for exposition.

We consider a population of calls referred to child protective services (CPS). Each call i is associated with initial characteristics X_i observed at the time of the hotline call and a potential for subsequent maltreatment $Y_i^* \in \{0, 1\}$. Calls proceed through two decision stages. In the *screening stage*, a hotline screener observes X_i and decides whether to screen the call in for investigation, $S_i \in \{0, 1\}$. In the *investigation stage*, an investigator visits the family, gathers additional information, and decides whether to open a case. Let $D_i^* \in \{0, 1\}$ denote the case-opening decision the investigator would make for call i; we observe it only when $S_i = 1$. The system's case-opening decision is $C_i = S_i \times D_i^*$: a case is opened only if the call is both screened in and then opened by an investigator. Subsequent maltreatment is observed only when no case is opened, so the observed outcome is $Y_i = (1 - C_i)Y_i^*$.

Let $m(X_i) \in [0, 1]$ denote a fixed algorithm trained to predict maltreatment risk using only the initial characteristics X_i — the information available at the time of the hotline call. We treat $m(\cdot)$ as given, corresponding to evaluation on a hold-out sample, and use it to define counterfactual decision rules of the form $\mathbf{1}\{m(X_i) \geq \tau\}$: the algorithm takes action whenever predicted risk exceeds the threshold τ.

The screener and investigator may each observe information beyond X_i that is relevant to maltreatment risk. The investigator, in particular, gathers additional information during the field visit that the algorithm does not access. Whether this additional information improves decisions is an empirical question: decision-makers may also respond to factors that are uncorrelated with outcomes, or may fail to use their private information effectively. The ranking quality measures we define below answer this question. To develop these comparisons, we initially treat each stage as having a single representative decision-maker. Heterogeneity across decision-makers in each stage will play an important role in both our decompositions within each stage (Section 2.3) and in our identification strategy (Section 4.2).

Any policymaker who prefers lower investigation rates, lower case-opening rates, and lower subsequent maltreatment rates can be represented by a payoff function of the form

$$\pi(y^*, S, D) = -\kappa_S S - \kappa_C C - \lambda \cdot (1 - C)y^*, \tag{1}$$

for costs $\kappa_S, \kappa_C, \lambda > 0$. The average payoff of any decision system therefore depends on three quantities: the investigation rate $E[S_i]$, the case-opening rate $E[C_i]$, and the subsequent maltreatment rate among children without an opened case, $E[(1-C_i)Y_i^*]$. A counterfactual decision system dominates the status quo if it weakly reduces all three quantities and strictly reduces at least one. This ranking holds for every choice of cost parameters, so no tradeoff need be specified. When two quantities are held fixed by construction, the comparison reduces to the effect on subsequent maltreatment alone. We exploit this dominance logic by choosing counterfactuals that hold specific action rates fixed — the total case-opening rate, the investigation rate, or the case-opening rate conditional on investigation — and compare the resulting subsequent maltreatment rates.

2.1 Human-Algorithmic Comparisons

We develop three comparisons that together summarize the ranking quality of the multi-stage system: how well does the system identify the highest-risk cases relative to an equally selective algorithm? Each comparison corresponds to a distinct counterfactual in which the algorithm replaces a different part of the system.

We first compare the full system against the algorithm. Define the algorithmic case-opening rule $C_i^m = \mathbf{1}\{m(X_i) \geq \tau^*\}$, where τ^* is chosen so that $\Pr(C_i^m = 1) = \Pr(C_i = 1)$. The algorithm bypasses both stages entirely: it ranks calls by predicted risk using only X_i and opens cases for the top-ranked calls at the same overall case-opening rate as the status quo. The *system-wide ranking quality*

$$\Delta^{CPS} := \mathbb{E}[Y_i^* \mid C_i = 0] - \mathbb{E}[Y_i^* \mid C_i^m = 0] \tag{2}$$

evaluates the total multistage system, comparing how the combined decisions of the screening and investigation stages compare to the algorithm. When $\Delta^{CPS} > 0$, the system leaves higher-risk children without an opened case than the algorithm would, misallocating its limited interventions. When $\Delta^{CPS} < 0$, the system outperforms the algorithm, indicating that information gathered across both stages is, on net, valuable for targeting. Under this counterfactual the investigation rate falls to zero — the algorithm eliminates all investigations — while the case-opening rate is unchanged by construction, so the comparison reduces to the sign of Δ^{CPS}.

We next isolate the investigation stage. Define the algorithmic case-opening rule among screened-in calls $D_i^m = \mathbf{1}\{m(X_i) \geq \tau^D\}$, where τ^D is chosen so that $\Pr(D_i^m = 1 \mid S_i = 1) = \Pr(D_i = 1 \mid S_i = 1)$. The screening stage is left untouched: we take screening decisions as given and ask what would happen if the algorithm re-ranked screened-in calls by predicted

risk and opened cases at the same rate as the status quo. The *investigation-stage ranking quality* evaluates how investigator case-opening decisions among screened-in calls compare to the algorithm:

$$\Delta^D := \mathbb{E}[Y_i^* \mid S_i = 1, D_i = 0] - \mathbb{E}[Y_i^* \mid S_i = 1, D_i^m = 0]. \tag{3}$$

When $\Delta^D > 0$, the investigation stage leaves higher-risk children without an opened case than the algorithm would, misallocating interventions among screened-in calls. When $\Delta^D < 0$, the investigation stage outperforms the algorithm, indicating that information gathered during the field visit is valuable for targeting. Under this counterfactual, the investigation rate falls to zero while the case-opening rate among screened-in calls is unchanged by construction. Define $\omega_D = \Pr(S_i = 1, D_i = 0 \mid C_i = 0)$ as the share of children without an opened case who exit at the investigation stage, and define $C_i^{D,m} = S_i \cdot D_i^m$ as the case-opening rule that replaces only the investigation stage. The total effect on subsequent maltreatment is

$$\mathbb{E}[Y_i^* \mid C_i = 0] - \mathbb{E}[Y_i^* \mid C_i^{D,m} = 0] = \omega_D \, \Delta^D, \tag{4}$$

so the comparison reduces to the sign of Δ^D.

Finally, we isolate the screening stage. Define the algorithmic screen-in rule $S_i^m = \mathbf{1}\{m(X_i) \geq \tau^S\}$, where τ^S is chosen so that $\Pr(S_i^m = 1) = \Pr(S_i = 1)$. Unlike the previous two comparisons, this counterfactual preserves the investigation stage: the algorithm replaces only the screener's decisions at the same screen-in rate, and investigators respond to whatever pool of calls the algorithm screens in. The *screening-stage ranking quality* evaluates how screener decisions compare to the algorithm at the screened-out margin:

$$\Delta^S := \mathbb{E}[Y_i^* \mid S_i = 0] - \mathbb{E}[Y_i^* \mid S_i^m = 0]. \tag{5}$$

When $\Delta^S > 0$, screeners leave higher-risk calls screened out than the algorithm would at the same screen-in rate. Δ^S therefore measures the direct effect at the screened-out margin. But because this is a multistage system, Δ^S does not fully determine the total effect of replacing the screener. The algorithm screens in a different pool of calls, which changes both the case-opening rate and the composition of children who ultimately exit without an opened case. The total effect on the system depends on the downstream response of the investigation stage.

To state this precisely, let $\omega_S = \Pr(S_i = 0 \mid C_i = 0)$ denote the share of children without an opened case who exit at the screening stage, so that $\omega_S + \omega_D = 1$. Define $C_i^{S,m} = S_i^m \cdot D_i^*$ as the case-opening rule that replaces only the screening stage. The total effect of replacing

only the screening stage is

$$\begin{aligned}
&\mathbb{E}[Y_i^* \mid C_i = 0] - \mathbb{E}[Y_i^* \mid C_i^{S,m} = 0] \\
&\quad = \omega_S \, \Delta^S + \omega_D \Big(\mathbb{E}[Y_i^* \mid S_i = 1, D_i^* = 0] - \mathbb{E}[Y_i^* \mid S_i^m = 1, D_i^* = 0] \Big) \\
&\qquad + (\omega_S - \lambda_S) \Big(\mathbb{E}[Y_i^* \mid S_i^m = 0] - \mathbb{E}[Y_i^* \mid S_i^m = 1, D_i^* = 0] \Big),
\end{aligned} \tag{6}$$

where $\lambda_S = \Pr(S_i^m = 0 \mid C_i^{S,m} = 0)$. The first term, $\omega_S \, \Delta^S$, captures the direct change at the screened-out margin. Two additional terms appear that have no counterpart in the investigation-stage comparison. The *downstream selection effect* reflects that the algorithm screens in a different pool, changing risk among calls that are screened in but not opened by the investigator. The *composition effect* captures the change in the fraction of non-opened calls exiting at each stage. Both depend on D_i^* — the investigator's response to the algorithmically screened-in pool — which Δ^S alone does not determine.

2.2 Decomposing the System-Wide Comparison

The three comparisons evaluate different configurations of the multistage system. How do they relate to one another? We next decompose the system-wide comparison into the contribution of eliminating the screening stage and the contribution of eliminating the investigation stage.

Proposition 1. *The system-wide ranking quality decomposes as*

$$\Delta^{CPS} = \omega_S \, \Delta^S + \omega_D \, \Delta^D - \omega_D \, \Psi^{CPS} + \Lambda^{CPS}, \tag{7}$$

where $\Psi^{CPS} := \mathbb{E}[Y_i^* \mid C_i^m = 0] - \mathbb{E}[Y_i^* \mid C_i^m = 0, S_i = 1]$ *is a* selection term*, and*

$$\begin{aligned}
\Lambda^{CPS} := \omega_S &\Big(\mathbb{E}[Y_i^* \mid S_i^m = 0] - \mathbb{E}[Y_i^* \mid C_i^m = 0] \Big) \\
&+ \omega_D \Big(\mathbb{E}[Y_i^* \mid D_i^m = 0, S_i = 1] - \mathbb{E}[Y_i^* \mid C_i^m = 0, S_i = 1] \Big)
\end{aligned} \tag{8}$$

is a threshold adjustment*. If* $\mathbb{E}[Y_i^* \mid m(X_i) < \tau]$ *and* $\mathbb{E}[Y_i^* \mid m(X_i) < \tau, S_i = 1]$ *are non-decreasing in* τ*, then* $\Lambda^{CPS} \leq 0$*.*

The direct stage-specific contributions are $\omega_S \, \Delta^S$ and $\omega_D \, \Delta^D$ — the stage-specific ranking quality gaps weighted by exit shares. Two additional terms, $-\omega_D \, \Psi^{CPS}$ and Λ^{CPS}, create wedges between Δ^{CPS} and this naive sum. Both arise from the structure of the multistage system.

The first wedge, $-\omega_D \Psi^{CPS}$, arises from the screener's selection. The investigation-stage comparison Δ^D evaluates the algorithm only on screened-in calls, but the system-wide comparison evaluates it on all calls. Ψ^{CPS} measures the resulting gap: the difference in risk between the algorithm's full non-opened pool and its non-opened pool among screened-in calls. When screeners route lower-risk calls to investigators ($\Psi^{CPS} > 0$), the algorithm faces an easier pool at the investigation stage, and Δ^D overstates how much the investigation stage contributes to the system-wide comparison. The correction $\omega_D \Psi^{CPS} < 0$ depresses Δ^{CPS} below the naive sum. It is therefore natural to group $\omega_S \Delta^S$ and $-\omega_D \Psi^{CPS}$ together as the full contribution of the screening stage: the first term captures direct ranking quality at the screened-out margin; the second captures how the screener's selection shapes the downstream comparison.

The second wedge, Λ^{CPS}, arises because stage-specific benchmarks are calibrated to each stage's own action rate, while the system-wide benchmark must match the overall case-opening rate: $\Pr(C_i = 1) = \Pr(S_i = 1) \cdot \Pr(D_i^* = 1 \mid S_i = 1)$. Each stage's action rate exceeds the overall rate, so the multistage structure implies $\tau^S \leq \tau^*$ and $\tau^D \leq \tau^*$.[2] Stage-specific benchmarks therefore evaluate the algorithm at more permissive thresholds, leaving behind lower-risk pools than the system-wide benchmark. This term is weakly negative provided that the algorithm sorts cases correctly on average — that $\mathbb{E}[Y_i^* \mid m(X_i) < \tau]$ is non-decreasing in τ, unconditionally and among screened-in calls.

Because $\Lambda^{CPS} \leq 0$ always and $-\omega_D \Psi^{CPS} \leq 0$ when screeners route lower-risk cases to investigators, both wedges can simultaneously depress Δ^{CPS} below the naive sum of the stage-specific comparison. Consequently, a researcher who evaluates the multistage system only through Δ^{CPS} can systematically understate the scope for stage-level algorithmic improvement.

2.3 Decomposing the Stage-Specific Comparisons

The stage-specific comparisons Δ^S and Δ^D each aggregate across many individual screeners and investigators. The same stage-level comparison can arise from two different sources: decision-makers may rank cases poorly, or they may rank them well but act at heterogeneous rates. We now decompose each stage-specific comparison to separate the two channels.

Because the decomposition takes the same form at both stages, we develop it in general notation. Consider a single stage with action $A_i \in \{0, 1\}$, decision-maker assignment $J_i \in \{1, \ldots, N^A\}$, and indicator R_i for whether call i reaches the stage (for example, $R_i = 1$

[2]Formally, $\Pr(S_i = 1) > \Pr(C_i = 1)$ implies $\tau^S < \tau^*$. For the investigation stage, $\Pr(m(X_i) \geq \tau^* \mid S_i = 1) \leq \Pr(m(X_i) \geq \tau^*)/\Pr(S_i = 1) = \Pr(D_i^* = 1 \mid S_i = 1)$, so the algorithm at τ^* acts at most as often as investigators among screened-in calls, hence $\tau^D \leq \tau^*$.

for screening; $R_i = S_i$ for investigation). The stage-wide algorithmic benchmark $A_i^m = \mathbf{1}\{m(X_i) \geq \tau^A\}$ is calibrated to match the stage-wide action rate: $\Pr(A_i^m = 1 \mid R_i = 1) = \Pr(A_i = 1 \mid R_i = 1)$.

For each decision-maker j, define a decision-maker-specific algorithmic rule $A_{j,i}^m(X_i) = \mathbf{1}\{m(X_i) \geq \tau_j^A\}$ calibrated to match j's own action rate:

$$\Pr(A_{j,i}^m = 1 \mid R_i = 1, J_i = j) = \Pr(A_i = 1 \mid R_i = 1, J_i = j).$$

The *ranking quality* of decision-maker j is

$$\Delta_j^A := E[Y_i^* \mid R_i = 1, A_i = 0, J_i = j] - E[Y_i^* \mid R_i = 1, A_{j,i}^m = 0, J_i = j],$$

which compares j's decisions against an equally selective algorithm, isolating how well j identifies the highest-risk cases at their own action rate (**?**Angelova, Dobbie and Yang, 2025). The *threshold adjustment* for decision-maker j is

$$\Lambda_j^A := E[Y_i^* \mid R_i = 1, A_{j,i}^m = 0, J_i = j] - E[Y_i^* \mid R_i = 1, A_i^m = 0, J_i = j],$$

which captures the cost or benefit of j acting at a different rate than the stage as a whole, holding the algorithm's ranking fixed. A more lenient decision-maker has $\tau_j^A < \tau^A$, leaving behind a lower-risk pool than the stage-wide benchmark, so $\Lambda_j^A < 0$; a more stringent one has $\Lambda_j^A > 0$.

Let $\alpha_j = \Pr(J_i = j \mid R_i = 1, A_i = 0)$ be the share of non-acted-upon cases at the stage handled by decision-maker j, and define weighted averages $\bar{\Delta}^A = \sum_j \alpha_j \Delta_j^A$ and $\bar{\Lambda}^A = \sum_j \alpha_j \Lambda_j^A$.

Proposition 2. *The stage-level ranking quality decomposes as*

$$\Delta^A = \bar{\Delta}^A + \bar{\Lambda}^A + R^A, \tag{9}$$

where $R^A = \sum_j \alpha_j \Big(E[Y_i^* \mid R_i = 1, A_i^m = 0, J_i = j] - E[Y_i^* \mid R_i = 1, A_i^m = 0] \Big)$ *is a composition residual.*

The three terms have distinct interpretations. $\bar{\Delta}^A$ is the weighted-average ranking quality of individual decision-makers: how well they identify the highest-risk cases at their own action rates. $\bar{\Lambda}^A$ is the weighted-average threshold adjustment: the aggregate cost or benefit of decision-makers acting at heterogeneous rates rather than the stage-wide rate. The residual R^A reflects differences in the composition of cases assigned across decision-makers; it captures

no feature of their decisions and vanishes in expectation when cases are randomly assigned within strata.

Applying Proposition 2 to the screening stage ($A_i = S_i$, $J_i = K_i$, $R_i = 1$) and the investigation stage ($A_i = D_i^*$, $J_i = G_i$, $R_i = S_i$) yields

$$\Delta^S = \bar{\Delta}^S + \bar{\Lambda}^S + R^S, \qquad \Delta^D = \bar{\Delta}^D + \bar{\Lambda}^D + R^D. \tag{10}$$

Beyond the means, the distribution of individual ranking quality and threshold adjustments across decision-makers is of direct policy interest. We summarize this distribution by the weighted standard deviations Σ_Δ^A and Σ_Λ^A, and by the share of decision-makers individually outperformed by the algorithm, $\Pr(\Delta_j^A > 0)$.

3 Setting and Data

3.1 Michigan Child Protective Services

The child protection system in the United States is designed to protect children from abuse and neglect. The process typically begins when a concerned individual—a mandated reporter such as a teacher, doctor, or police officer, or a non-mandated reporter such as a neighbor or family member—contacts a CPS hotline to report concerns about a child's safety. In Michigan, the vast majority of reports (97 percent) are made by phone, with the remaining share submitted through an online reporting system. For ease of exposition, we refer to all reports as "calls" throughout the paper.

Figure 1 illustrates the CPS process in Michigan, which functions similarly to many other states. The state operates two centralized intake centers, located in Detroit and Grand Rapids, that receive and process reports from across the state (and share the same phone number). When a call is received, a hotline screener gathers information about the child and the alleged maltreatment and determines whether the report meets the statutory criteria for investigation. Calls last approximately 15 minutes on average, at the end of which the screener makes a binary decision: whether to "screen in" the report for a formal investigation or "screen out" the report. Reports that are screened out receive no further CPS action. Roughly 37 percent of calls in our sample are screened in.

At the time of the call, screeners gather detailed information provided by the reporter, including the nature of the allegation, the identity and type of reporter (e.g., teacher, medical professional, family member), and information about the child and household. In addition, screeners have access to administrative records through the state's case management system, including prior CPS contact for the child and family (e.g., prior screened-in and screened-out

reports), as well as information on the findings of past investigations or foster care involvement. Screeners may also access other administrative systems to verify addresses, identify household members, and observe participation in public assistance programs. Together, this information provides a snapshot of both the current allegation and the family's prior system involvement. At the same time, screeners must make decisions quickly—typically within a single call—and not all available information is necessarily consulted in every case. Screening decisions are guided by relatively broad statutory criteria and agency guidelines that emphasize assessing both the validity of the allegation and the risk of harm to the child. In practice, this framework leaves substantial room for discretion, as screeners must weigh incomplete information and make judgment calls about whether a report warrants investigation.

Importantly for our analysis, incoming reports—whether submitted by phone or through the online portal—are assigned to screeners through a centralized, queue-based system. Reports enter a common queue and are routed to the available screener who has been waiting the longest since their previous assignment. This assignment process generates quasi-random variation in which screener handles a given report, conditional on the exact day and shift in which the report occurs.

If a report is screened in, it is transferred from the centralized intake center to the local child welfare office in the child's county of residence, where an investigator is assigned to the case. Investigations typically begin within 24 hours for higher-risk cases and within several days for lower-risk cases, and must generally be completed within 30 days. The investigation process involves direct interaction with the family and is designed to assess both the credibility of the allegation and the child's safety. In most cases, investigators conduct one or more home visits, during which they observe the home environment, assess the child's immediate safety, and interview caregivers and children. Investigators typically speak separately with children, parents, and other household members, and may also contact individuals outside the household such as teachers or medical providers. In addition, investigators may review administrative records, including prior CPS reports, school records, or medical information, to inform their assessment.

Screened-in reports are assigned to investigators on a rotational basis, rather than based on investigator characteristics or expertise. During the investigation, caseworkers complete standardized, paper-based safety and risk assessment tools that combine objective information (e.g., child age, prior system involvement) with subjective evaluations—for example, whether the caregiver "views the incident less seriously than the department." Prior work shows that investigators may adjust responses based on their priors about a child's risk level (Gillingham and Humphreys, 2010; Bosk, 2015). Consequently, investigators retain substantial discretion over the final disposition of the case. We observe both the resulting safety and

risk assessment scores and detailed narrative summaries describing the investigation and the rationale for the final disposition in the administrative data.

At the conclusion of the investigation, each case is assigned to one of five disposition categories defined by Michigan's Child Protection Law. These categories summarize both the evidence of maltreatment and the assessed risk of future harm, and jointly determine the agency's response. Categories III and above correspond to substantiated cases—meaning the investigator determines there is sufficient evidence of abuse or neglect. Categories II and I, by contrast, require the agency to open a protective services case and provide ongoing services to the family. Category I represents the most severe cases and may involve court action or removal, while Categories III–V do not trigger ongoing supervision.

We focus on the decision to open a case (Categories I and II) as the central intervention margin in the child welfare system. Case opening determines which families receive sustained monitoring and services aimed at reducing future maltreatment, including home visits, service coordination, and referrals to support programs. In addition, substantiated findings that result in Categories I and II lead to caregivers being listed in the state's central registry of child abuse and neglect, which may have lasting consequences. In our data, approximately 11% of screened-in calls result in a case opening, making this a relatively common but targeted intervention.

Although foster care placement is the most intensive and high-stakes intervention, it is rare in practice—occurring in only about 2% of screened-in calls in our sample. The vast majority of agency activity instead involves identifying families who would benefit from in-home services and support. As a result, the key policy problem may not be primarily which children to remove, but rather how to target limited resources toward families at highest risk of future harm.

Under CPS guidelines, investigators are instructed to intervene when the risk of future maltreatment is sufficiently high, while otherwise minimizing unnecessary intrusion into the family. The investigation stage therefore plays a central role in determining which families receive services, making the quality of case-opening decisions critical for the effectiveness of the system.

3.2 Data and Analysis Sample

Our analysis draws on child welfare records from the Michigan Department of Health and Human Services (MDHHS), which include the universe of reports made to Michigan's centralized CPS intake hotline between 2017 and 2022. These records capture both reports that are screened in for investigation and those that are screened out. For each call, we

observe the identity of the assigned screener, the screening decision, the call's exact date and time, and information about the children and allegations involved as coded by the screener. For screened-in calls, we additionally observe the identity of the assigned investigator, the relationship of the child to the alleged perpetrator (as coded by the investigator), the child's ZIP code, as well as investigation outcomes including substantiation decisions, case openings, and foster care placements.

In addition to the structured administrative data described above, we observe two sources of information collected during the investigation stage that are not available at the time of the hotline call. First, beginning in 2018, we observe detailed narrative case notes written by investigators that document the evidence gathered during the investigation and the reasoning underlying the final disposition decision. These notes provide rich, unstructured descriptions of the home environment, caregiver behavior, and interactions with the child and family. Second, we observe the full set of structured safety and risk assessment instruments completed during the investigation. The safety assessment is completed near the end of the investigation and records the presence of a comprehensive set of immediate harm factors - covering domains such as abuse, supervision, living conditions, and caregiver behavior. The risk assessment is also completed prior to disposition and aggregates information across a large number of items capturing prior maltreatment history, caregiver characteristics, household composition, and behavioral risk factors into a summary risk score (e.g., low, moderate, high, or intensive). While these instruments provide structured measures of risk, they are completed by investigators and incorporate subjective judgments, leaving scope for discretion in how underlying information is recorded. We incorporate the structured assessment data into our analysis, and in ongoing work we use the narrative case notes to better understand the sources of investigator heterogeneity and the dimensions of risk not captured by the structured administrative data.

A given hotline call may contain multiple children, but screen-in and case opening decisions are consistent across all children involved in the call, so our screener and investigator analysis samples are at the call level. We begin with $937,457$ unique hotline calls in Michigan between January 2017 and December 2022. We first drop calls with a missing screener identifier or screened-in calls with a missing investigator identifier ($N = 137,180$). We then exclude calls involving children younger than one year old, as state policy requires that substance-exposed infants be screened in for investigation, leaving little or no discretion to screeners ($N = 96{,}431$). To ensure that calls in our sample are quasi-randomly assigned to both screeners and investigators, we drop: (i) calls involving sexual abuse, which are handled by a non-random subset of investigators ($N = 36{,}106$); (ii) calls involving trafficking allegations reported by law enforcement, which tend to be handled by a non-random subset

of screeners ($N = 4$); and (iii) screened-in calls in which a child was involved in a previous investigation within the past 60 days or a case opening within the past 365 days, as these cases are often assigned to the same investigator ($N = 42{,}407$). We also drop screened-in calls with missing child ZIP code information, since quasi-random assignment of investigators is conditional on geography ($N = 34{,}529$), as well as duplicate calls for a given child on the same day ($N = 23{,}174$).[3] To minimize noise in our measures of screener and investigator tendencies, we drop calls assigned to screeners with fewer than 100 calls in the sample ($N = 1{,}000$), as well as screened-in calls in month-by-year-by-ZIP-code-by-shift cells that contain only a single investigator ($N = 41{,}969$).

The resulting screener analysis sample, summarized in Columns (1)–(3) of Table 1, consists of 524,657 unique calls involving 838,650 call-by-child records assigned to 292 screeners. Seventy-six percent of calls involve White children, and 62% involve a female child; the average age of the youngest child in the call is 8.7 years. Twenty-two percent of calls include a physical abuse allegation, and 68% include an improper supervision allegation. The overall screen-in rate is 37.2%. Columns (4)–(6) of Table 1 describe the investigator analysis sample, which conditions on screened-in calls (so Columns (2) and (4) are identical). The 195,355 screened-in calls are assigned to 8,753 investigators. Conditional on being screened in, 10.8% of calls result in a case opening.

The primary outcome of the paper is a proxy for subsequent maltreatment potential, Y_i^*, reported in Panel D. This measure is an indicator for whether any child involved in the call is the subject of a new hotline report between 31 and 180 days after the initial call. We exclude the first 30 days after the initial hotline call because investigations typically take up to 30 days, ensuring that the outcome is defined consistently across screened-in and screened-out calls. The key identification challenge is that Y_i^* is only observed for cases in which no case is opened: for cases that result in a case opening, we do not observe whether the child would have been re-reported in the absence of intervention. This outcome is natural for our analysis because it is based on subsequent reports to the hotline - an external measure of concern - and is therefore not mechanically driven by downstream screening or case-opening decisions. In contrast, alternative measures such as subsequent investigations or case openings would depend on endogenous screening decisions within the CPS system. We also examine alternative outcome definitions, including more intensive forms of subsequent CPS contact and different time horizons. Conditional on a case not being opened, 34% and 31% of hotline calls in the screener and investigator samples, respectively, are followed by a

[3]Approximately 5% of children have multiple calls on the same day in the initial data. When removing duplicate calls within a given day, we prioritize calls that resulted in a case opening, followed by calls that were screened in, and then randomly select among the remaining calls when ties remain.

subsequent report within six months.

4 Empirical Approach

4.1 Building the Prediction Algorithm

We first construct a predictive algorithm using the structured information available at the time of the hotline call to estimate subsequent maltreatment risk. We train the algorithm on a training sample and evaluate its performance - relative to screener and investigator decisions - on a separate evaluation sample. The training sample consists of a random 50% sample of screened-out calls (256,053 call-by-child observations), together with all screened-in observations assigned to investigators with fewer than 50 screened-in calls (176,616 observations). The latter restriction is designed to reduce noise from investigator-specific tendencies in the evaluation sample.

The algorithm is estimated in two stages. The first stage uses a randomly selected 90% of the training sample to estimate a gradient-boosted decision tree (XGBoost) at the call-by-child level. The model uses the full set of structured information available to the screener at the time of the call, including complaint timing, allegation type, reporter and contact-method characteristics, child age, and prior CPS history. The outcome is an indicator for whether the child is the subject of a subsequent hotline report following the initial call. We restrict estimation to observations that were either screened out or screened in without an opened case, so that subsequent maltreatment risk is observed. For the first-stage XGBoost model, we fix the number of boosting rounds at 100 and choose hyperparameters by five-fold cross-validation, grouped at the child level. We tuned hyperparameters for maximum tree depth, subsample rate, column subsample rate, minimum child weight, ℓ_2 regularization parameter, and learning rate; we iterate over a grid of hyperparameter values for each. Hyperparameters are selected to maximize out-of-fold AUC.

In the second stage, using the remaining 10% of the training sample, we aggregate the data to the call level and estimate whether any child associated with the call is the subject of a subsequent report. The model includes the number of children associated with the call and summary measures of the first-step predicted risks within the call (mean, maximum, and minimum). This second-stage model is also estimated using a gradient-boosted decision tree and is restricted to calls for which the outcome is observed. The resulting predictions provide a call-level measure of subsequent maltreatment risk.

Appendix Figure A.3 reports the performance of the risk-prediction algorithm in the evaluation sample. In the screener sample, the algorithm achieves an AUC of 0.688. Restricting

to screened-in calls yields similar performance in the investigation sample, with an AUC of 0.661. Panel B also reports a specification that incorporates risk and safety assessments into the first stage of the algorithm; this has virtually no effect on predictive performance, with an AUC of 0.659.

4.2 Identification and Empirical Strategy

We next discuss how we identify and estimate the key components of the system decomposition. Most terms in the decomposition framework are point identified; the main challenge is identifying terms involving subsequent maltreatment under algorithmic thresholds. For example, the term $\mathbb{E}[Y_i^* \mid S_i^m = 0]$ represents the counterfactual maltreatment rate among cases with algorithmic risk below τ^S. This quantity is not directly observed because maltreatment potential is unobserved for screened-in calls in which a case is opened. We address this challenge by leveraging the quasi-random assignment of screeners and investigators to identify or bound these algorithmic counterfactuals.

Introducing notation, let $K_i \in \{1, \ldots, K\}$ denote the screener assigned to call i, and let $G_i \in \{1, \ldots, G\}$ denote the investigator assigned to call i. Define indicators for agent assignment: $Z_{ik}^S = 1\{K_i = k\}$ and $Z_{ig}^D = 1\{G_i = g\}$.

As discussed above, screeners and investigators are quasi-randomly assigned conditional on shift-by-day strata for screeners (denoted T_i^S) and office-by-time strata for investigators (denoted T_i^D). Our first key identifying assumption is that assignment is conditionally independent of case characteristics and potential outcomes:[4]

$$\begin{aligned}
&\left(Z_{i1}^S, \ldots, Z_{iK}^S\right) \perp\!\!\!\perp (X_i, Y_i^*) \mid T_i^S, \\
&\left(Z_{i1}^D, \ldots, Z_{iG}^D\right) \perp\!\!\!\perp (X_i, Y_i^*) \mid T_i^D,\ S_i = 1.
\end{aligned}$$

We further assume that screener and investigator assignment affect Y_i^* only through the screening and case-opening decisions, S_i and D_i. This corresponds to an exclusion restriction, as in Baron et al. (2024). We discuss supporting evidence for this assumption below.

Denote $m(X_i)$ as our fixed algorithmic prediction, trained on a separate sample. For any threshold t, denote $\mu(t) = E[Y_i^* \mid m(X_i) \leq t]$ as the counterfactual subsequent maltreatment among calls with algorithmic risk less than or equal to t. Similarly, denote $\mu_D(t) = E[Y_i^* \mid m(X_i) \leq t, S_i = 1]$ as the counterfactual subsequent maltreatment among screened-in calls

[4]In our decompositions of the screening and investigation stages, the residual terms R^S and R^D capture compositional differences in the cases assigned to screeners and investigators. Random assignment within strata does not eliminate these residuals, as decision-makers appear in different strata with different frequencies and algorithmic counterfactual outcomes may vary across strata.

with algorithmic risk less than or equal to t. For example, the key algorithmic counterfactuals that enter into the decomposition are:

$$E[Y_i^* \mid C^m(X_i) = 0] = \mu(\tau^*)$$
$$E[Y_i^* \mid S^m(X_i) = 0] = \mu(\tau^S)$$
$$E[Y_i^* \mid D^m(X_i) = 0, S_i = 1] = \mu_D(\tau^D)$$

Hence, the identification challenge is to recover $\mu(t), \mu_D(t)$. This is the same challenge faced by Angelova, Dobbie and Yang (2025), which we similarly resolve using an identification-at-infinity approach. As a robustness check before Summer Institute, we will also report analogous results by calculating worst-case bounds on algorithmic performance (Rambachan, 2024; Rambachan, Coston and Kennedy, 2025).

To build intuition, fix a threshold t and consider a setting with random investigator assignment and a hypothetical investigator g^* who never opens cases for calls with predicted risk below t. For screened-in calls assigned to g^* with $m(X_i) \leq t$, subsequent maltreatment is always observed, since no cases are opened. As a result, the average observed maltreatment rate in this group identifies $\mathbb{E}[Y_i^* \mid m(X_i) \leq t, S_i = 1, Z_{ig^*}^D = 1]$. By random assignment, this equals

$$\mathbb{E}[Y_i^* \mid m(X_i) \leq t, S_i = 1, Z_{ig^*}^D = 1] = \mathbb{E}[Y_i^* \mid m(X_i) \leq t, S_i = 1] = \mu_D(t).$$

Such a perfectly lenient investigator does not exist. Instead, we estimate these counterfactuals by extrapolating observed variation in maltreatment rates across quasi-randomly assigned decision-makers with different case-opening rates. Intuitively, investigators (or screeners) with lower case-opening rates reveal more information about underlying maltreatment risk, since outcomes are observed for a larger share of cases. For example, we estimate $\mu(t)$ by extrapolating the relationship between screener-specific maltreatment rates and case-opening rates for calls with $m(X_i) \leq t$. In particular, we regress $\mathbb{E}[Y_i \mid S_i = 0, Z_{ik}^S = 1, m(X_i) \leq t]$ on $\mathbb{E}[C_i \mid Z_{ik}^S = 1, m(X_i) \leq t]$, and take the intercept at a zero case-opening rate as an estimate of $\mu(t)$.

In practice, random assignment of screeners to calls is conditional on shift-by-day strata, T_i^S, and random assignment of investigators to screened-in calls is conditional on office-by-time strata, T_i^D. Following Angelova, Dobbie and Yang (2025), we account for these strata using a linear regression adjustment.

For the screener analysis, we estimate:

$$S_i = \sum_{k=1}^{K} \phi_k^S Z_{ik}^S + T_i^S \gamma + \epsilon_i^S,$$
$$C_i = \sum_{k=1}^{K} \eta_k^S Z_{ik}^S + T_i^S \gamma + \epsilon_i^C,$$
$$Y_i = \sum_{k=1}^{K} \mu_k^S Z_{ik}^S + T_i^S \gamma + \epsilon_i^Y \quad \text{for } C_i = 0.$$

From these regressions, we construct screener-specific screen-in rates as $\widehat{\mathbb{E}}[S_i \mid K_i = k] = \mathbb{E}_n\left[\widehat{\phi}_k^S + T_i^S \widehat{\gamma}\right]$, where $\mathbb{E}_n[\cdot]$ denotes the sample average over the relevant subsample. We similarly construct screener-specific case-opening rates and conditional maltreatment rates. We then estimate $\mu(t)$ via linear extrapolation, taking the intercept from a regression of $\widehat{\mathbb{E}}[Y_i \mid K_i = k, C_i = 0, m(X_i) \leq t]$ on $\widehat{\mathbb{E}}[C_i \mid K_i = k, m(X_i) \leq t]$.

For the investigator analysis, we estimate:

$$C_i = \sum_{g=1}^{G} \phi_g^D Z_{ig}^D + T_i^D \gamma + \epsilon_i^C \quad \text{for } S_i = 1,$$
$$Y_i = \sum_{g=1}^{G} \mu_g^D Z_{ig}^D + T_i^D \gamma + \epsilon_i^Y \quad \text{for } D_i = 0,\ S_i = 1.$$

We construct investigator-specific case-opening and conditional maltreatment rates analogously. We then estimate $\mu_D(t)$ via linear extrapolation, taking the intercept from a regression of $\widehat{\mathbb{E}}[Y_i \mid G_i = g, D_i = 0, S_i = 1, m(X_i) \leq t]$ on $\widehat{\mathbb{E}}[C_i \mid G_i = g, S_i = 1, m(X_i) \leq t]$.

This approach identifies the ranking quality terms, such as $\Delta^{CPS} = \mathbb{E}[Y_i^* \mid C_i = 0] - \mathbb{E}[Y_i^* \mid C_i^m = 0]$, where the first term is observed and the second term is estimated via linear extrapolation. The remaining decomposition estimands ($\overline{\Delta}^S, \overline{\Lambda}^S, \overline{\Delta}^D, \overline{\Lambda}^D$) are weighted averages of screener- or investigator-specific quantities.

These agent-specific quantities are estimated using the same identification strategy as above. For example, for $\Delta_k^S = \mathbb{E}[Y_i^* \mid S_i = 0, K_i = k] - \mathbb{E}[Y_i^* \mid S_{k,i}^m = 0, K_i = k]$, the first term is given by the screener-specific conditional maltreatment rate (adjusted for T_i^S), while the second term is obtained from $\mu(t)$ evaluated at the threshold t that matches the screen-in rate of screener k. We can therefore identify and estimate all components of Proposition 1, as well as the within-stage decompositions in Proposition 2.

For decision-maker heterogeneity, plug-in estimates conflate true heterogeneity across screeners and investigators with sampling noise. To address this, we use empirical Bayes

shrinkage. We describe the procedure for the screening stage; the investigation stage is analogous. Using the agent-specific estimates Δ_k^S and Λ_k^S, along with weights α_k^S, we apply the posterior average effects approach of Bonhomme and Weidner (2022) to recover the posterior standard deviations Σ_Δ^S and Σ_Λ^S, as well as the share of decision-makers individually outperformed by the algorithm, $\Pr(\Delta_k^S > 0)$.

4.3 Identifying Assumptions

Our identification strategy relies on the assumption that, conditional on the relevant time and location strata, screeners and investigators are quasi-randomly assigned to cases. This implies that observable case characteristics are uncorrelated with the assigned decision-maker's tendencies within strata. This assumption has been examined extensively in prior work using the same institutional setting (Baron, Jacob and Lombardo, 2026; Baron and Gross, 2025; Baron et al., 2024). We also assess this implication in our sample using balance tests reported in Appendix Figures A.1 and A.2.

For the screening stage, Appendix Figure A.1 reports coefficients from regressions of the screening decision (Panel A) and the assigned screener's leave-out screening propensity (Panel B) on child and call characteristics, controlling for day-by-shift fixed effects. As expected, observable characteristics are highly predictive of the screening decision, but are not jointly predictive of the assigned screener's tendency, consistent with quasi-random assignment within strata.

Appendix Figure A.2 presents analogous results for the investigation stage. Panel A shows that observable characteristics strongly predict case opening decisions. In contrast, Panel B shows that most observable characteristics are not statistically related to the assigned investigator's leave-out case opening propensity. The only exception is the average number of prior investigations, which is statistically significant, although its magnitude is small. The joint F-statistic is correspondingly larger, consistent with the presence of this imbalance.

To assess the robustness of this pattern, Panel C restricts the sample to screened-in calls in which none of the children had a prior substantiation. In this subsample, all coefficients are small and statistically insignificant, and the joint F-statistic declines substantially. Our main results use the full sample, but we show in robustness checks that the results are similar in this more balanced subsample. Taken together, these results support the identifying assumption that variation in decision-maker tendencies is orthogonal to observable case characteristics within the assignment strata.

The second key assumption is an exclusion restriction embedded in our definition of Y_i^*: subsequent maltreatment is not directly affected by screener identity, investigator iden-

tity, or the screening decision itself. That is, we treat Y_i^* as the underlying potential for maltreatment absent intervention, rather than an outcome that depends on assignment or intermediate decisions.

For the screening stage, this assumption is plausible for at least two reasons. First, hotline screeners play a narrow and short-lived role in the CPS process: calls last approximately 15 minutes, and the screener's sole responsibility is to determine whether a report meets the criteria for investigation. Second, once a report is screened in, it is transferred to a local child welfare office and assigned to a different investigator. Screeners have no role in the subsequent investigation, case management, or service provision. As a result, screener assignment is unlikely to affect child outcomes except through the decision to initiate an investigation.

For the investigation stage, the exclusion restriction requires that differences across investigators affect outcomes only through the case-opening decision. This assumption is more demanding, as investigators interact directly with families during the investigation and have discretion over other decisions beyond case opening. We consider two potential channels.

First, investigators may affect outcomes through their interactions with families during the investigation itself, even when no case is opened. While such effects cannot be ruled out, investigations do not involve ongoing monitoring or supervision, limiting the scope for sustained impacts on subsequent maltreatment.

Second, investigators retain discretion over other decisions, such as substantiation and referrals to services, that may occur even when a case is not opened. In practice, these interventions are typically light-touch - for example, providing referrals to community resources - and do not involve follow-up or enforcement. As a result, they are unlikely to have large effects on subsequent outcomes. Taken together, these institutional features suggest that investigator assignment primarily affects outcomes through the case-opening decision.

A related and important implication of this assumption is that the screening decision itself does not directly affect Y_i^*. In particular, screening a call in for investigation could affect subsequent outcomes by changing family behavior or home conditions, even in the absence of a case opening. This concern is central for our screening-stage analysis, as direct effects of investigation would confound comparisons between cases that are screened out and those that are screened in but not opened.

We assess the plausibility of this condition using evidence from prior work. Baron et al. (2024) exploit the quasi-random assignment of screeners in Michigan and show that the effect of being screened in for investigation on subsequent CPS contact is small for compliers, with bounds that are close to zero and statistically insignificant. Similarly, Baron, Jacob and Lombardo (2026) use a related research design and find that the impact of investigations on

subsequent maltreatment and educational outcomes is modest and generally not statistically significant. Taken together, this evidence suggests that any direct effects of investigation on Y_i^* are likely limited in our context. Before Summer Institute, we will report analogous robustness checks to evaluate violations of the exclusion restriction for the specific algorithmic counterfactuals we study.

4.4 Counterfactual Maltreatment Rates under the Algorithm

Figure 2 illustrates the extrapolation-based strategy at three thresholds for the screener and investigator samples. In Panel A, the top set of points reports results for the full sample of hotline calls, corresponding to an algorithmic case-opening threshold of 0%. The lower two sets of points show results for thresholds of 20% and 40%.

The horizontal axis of Panel A reports binned estimates of screener-specific case-opening rates, $\widehat{\mathbb{E}}[C_i \mid K_i = k, m(X_i) \leq t]$, adjusting for shift-by-day fixed effects. For each threshold, many screeners have very low case-opening rates, implying that the extrapolation to a zero case-opening rate is relatively short. The vertical axis reports binned estimates of conditional subsequent maltreatment rates, $\widehat{\mathbb{E}}[Y_i \mid K_i = k, C_i = 0, m(X_i) \leq t]$, again adjusting for shift-by-day fixed effects. Panel B presents analogous results for the investigator sample, which conditions on screened-in calls, and reports thresholds of 0%, 5%, and 10%.

In each panel, the vertical intercept of the fitted line provides the linear extrapolation estimate of the threshold-specific average maltreatment rate. The lines of best fit are weighted by the inverse variance of the estimated maltreatment rates. Standard errors are computed via a bootstrap procedure that resamples the estimated screener- and investigator-specific moments and recomputes the extrapolations. For example, the estimated mean maltreatment rate among all calls is 34.9% (SE: 0.5), while the corresponding estimate among screened-in calls is 25.3% (SE: 0.4). Estimated maltreatment rates decline as the algorithmic case-opening threshold increases. Consistent with the short extrapolation distance, the estimates are similar when using quadratic or local linear specifications.

We repeat this procedure across algorithmic case-opening rates between 0% and 50%, spanning the range of observed screener and investigator decision rates. The results are reported in Figure 3. In both samples, estimated maltreatment rates decline as the case-opening threshold increases. The figure highlights three policy-relevant thresholds: C^m (the overall case-opening rate, $\mathbb{E}[C_i]$), D^m (the case-opening rate conditional on being screened in, $\mathbb{E}[C_i \mid S_i = 1]$), and S^m (the screen-in rate, $\mathbb{E}[S_i]$). The extrapolated values at these thresholds are used in the system-wide and stage-specific decompositions.

For agent-specific comparisons, such as Δ_j^S, these extrapolation estimates allow us to

recover the counterfactual maltreatment rate at each decision-maker's own action rate - for example, the screener-specific screen-in rate or the investigator-specific case-opening rate.

References

Administration for Children, Youth and Families. 2026. "Modernizing Child Welfare Technologies and Tools: Opportunities for Predictive Risk Modeling to Improve Child Safety and Outcomes." Administration for Children, Youth and Families, U.S. Department of Health and Human Services Issue Brief.

Angelova, Victoria, Will S. Dobbie, and Crystal Yang. 2025. "Algorithmic Recommendations and Human Discretion." *Review of Economic Studies*.

Arnold, David, Will S. Dobbie, and Peter Hull. 2022. "Measuring Racial Discrimination in Bail Decisions." *American Economic Review*, 112(9): 2992–3038.

Baron, E. Jason, and Max Gross. 2025. "Is There a Foster Care-to-Prison Pipeline? Evidence from Quasi-Randomly Assigned Investigators." *Review of Economics and Statistics*.

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

Baron, Jason, Brian A. Jacob, and Richard Lombardo. 2026. "The Impact of Maltreatment Investigations on Child Wellbeing." Working paper.

Bonhomme, Stéphane, and Martin Weidner. 2022. "Minimizing Sensitivity to Model Misspecification." *Quantitative Economics*, 13(3): 907–954.

Bosk, Emily Adlin. 2015. "All Unhappy Families: Standardization and Child Welfare Decision-Making." PhD diss. University of Michigan.

Efron, Bradley. 2016. "Empirical Bayes Deconvolution Estimates." *Biometrika*, 103(1): 1–20.

Fitzpatrick, Maria D, Katharine Sadowski, and Christopher Wildeman. 2025. "Algorithms and Decision-making: Evidence from Child Maltreatment Reports." *Journal of Human Resources*.

Gillingham, Philip, and Cathy Humphreys. 2010. "Child Protection Practitioners and Decision-Making Tools: Observations and Reflections from the Front Line." *British Journal of Social Work*, 40(8): 2598–2616.

Grimon, Marie-Pascale, and Christopher Mills. 2025. "Better Together? A Field Experiment on Human-Algorithm Interaction in Child Protection." *Working Paper*.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow.* Macmillan.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment.* New York:Little, Brown Spark.

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.

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

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review*, 105(5): 491–95.

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables." *KDD '17*, 275–284. New York, NY, USA:Association for Computing Machinery.

Ludwig, Jens, and Sendhil Mullainathan. 2024. "Machine Learning as a Tool for Hypothesis Generation." *Quarterly Journal of Economics*, 139(2): 751–827.

Movva, Rajiv, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. 2025. "Sparse Autoencoders for Hypothesis Generation." Vol. 267 of *Proceedings of Machine Learning Research*, 44997–45023.

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

Mullainathan, Sendhil, and Ziad Obermeyer. 2022. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care." *Quarterly Journal of Economics*, 137(2): 679–727.

Putnam-Hornstein, Emily, Rhema Vaithianathan, Jacquelyn McCroskey, and Daniel Webster. 2022. "Los Angeles County Department of Children and Family Services: Risk Stratification Methodology Report."

Rambachan, Ashesh. 2024. "Identifying Prediction Mistakes in Observational Data." *Quarterly Journal of Economics*, 139(3): 1665–1711.

Rambachan, Ashesh, Amanda Coston, and Edward Kennedy. 2025. "Robust Design and Evaluation of Predictive Algorithms under Unobserved Confounding." Working paper.

Rittenhouse, Katherine, Emily Putnam-Hornstein, and Rhema Vaithianathan. 2025. "Algorithms, Humans, and Racial Disparities in Child Protective Services: Evidence from the Allegheny Family Screening Tool." Working paper.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science*, 185(4157): 1124–1131.

A Omitted Proofs

A.1 Proof of Equation (4)

We show that $E[Y_i^* \mid C_i = 0] - E[Y_i^* \mid C_i^{D,m} = 0] = \omega_D \Delta^D$. Since τ^D is chosen so that $\Pr(D_i^m = 1 \mid S_i = 1) = \Pr(D_i^* = 1 \mid S_i = 1)$, it follows that $\Pr(S_i = 1, D_i^m = 0) = \Pr(S_i = 1, D_i^* = 0)$, and hence $\Pr(C_i^{D,m} = 1) = \Pr(C_i = 1)$, so $\Pr(C_i^{D,m} = 0) = \Pr(C_i = 0)$.

Under the status quo, by iterated expectations,

$$E[Y_i^* \mid C_i = 0] = \omega_S \, E[Y_i^* \mid S_i = 0] + \omega_D \, E[Y_i^* \mid S_i = 1, D_i^* = 0].$$

Under the investigation-stage counterfactual, the screened-out group is identical (the screening stage is unchanged), and $\Pr(C_i^{D,m} = 0) = \Pr(C_i = 0)$, so:

$$E[Y_i^* \mid C_i^{D,m} = 0] = \omega_S \, E[Y_i^* \mid S_i = 0] + \omega_D \, E[Y_i^* \mid S_i = 1, D_i^m = 0].$$

Subtracting then yields:

$$E[Y_i^* \mid C_i = 0] - E[Y_i^* \mid C_i^{D,m} = 0] = \omega_D \Big(E[Y_i^* \mid S_i = 1, D_i^* = 0] - E[Y_i^* \mid S_i = 1, D_i^m = 0] \Big) = \omega_D \, \Delta^D. \quad \square$$

A.2 Proof of Equation (6)

We show that

$$\begin{aligned} E[Y_i^* \mid C_i = 0] - E[Y_i^* \mid C_i^{S,m} = 0] = \omega_S \, \Delta^S + \omega_D \Big(E[Y_i^* \mid S_i = 1, D_i^* = 0] - E[Y_i^* \mid S_i^m = 1, D_i^* = 0] \Big) \\ + (\omega_S - \lambda_S) \Big(E[Y_i^* \mid S_i^m = 0] - E[Y_i^* \mid S_i^m = 1, D_i^* = 0] \Big). \end{aligned}$$

We decompose each conditional expectation over the two groups that exit without a case opening. Under the status quo:

$$E[Y_i^* \mid C_i = 0] = \omega_S \, E[Y_i^* \mid S_i = 0] + \omega_D \, E[Y_i^* \mid S_i = 1, D_i^* = 0],$$

where $\omega_S + \omega_D = 1$ by definition. Under the screening-stage counterfactual, the two exit groups are $\{S_i^m = 0\}$ and $\{S_i^m = 1, D_i^* = 0\}$:

$$E[Y_i^* \mid C_i^{S,m} = 0] = \lambda_S \, E[Y_i^* \mid S_i^m = 0] + (1 - \lambda_S) \, E[Y_i^* \mid S_i^m = 1, D_i^* = 0],$$

where $\lambda_S = \Pr(S_i^m = 0 \mid C_i^{S,m} = 0)$. Subtracting and adding and subtracting $\omega_S \, E[Y_i^* \mid$

$S_i^m = 0]$:

$$\begin{aligned}
&E[Y_i^* \mid C_i = 0] - E[Y_i^* \mid C_i^{S,m} = 0] \\
&\quad = \omega_S \Big(E[Y_i^* \mid S_i = 0] - E[Y_i^* \mid S_i^m = 0] \Big) \\
&\qquad + \omega_D \, E[Y_i^* \mid S_i = 1, D_i^* = 0] \\
&\qquad + (\omega_S - \lambda_S) \, E[Y_i^* \mid S_i^m = 0] \\
&\qquad - (1 - \lambda_S) \, E[Y_i^* \mid S_i^m = 1, D_i^* = 0].
\end{aligned}$$

The first line is $\omega_S \, \Delta^S$. For the remaining terms, we use $\omega_S + \omega_D = 1$ to write $1 - \lambda_S = \omega_D + (\omega_S - \lambda_S)$:

$$\begin{aligned}
&= \omega_S \, \Delta^S + \omega_D \Big(E[Y_i^* \mid S_i = 1, D_i^* = 0] - E[Y_i^* \mid S_i^m = 1, D_i^* = 0] \Big) \\
&\quad + (\omega_S - \lambda_S) \Big(E[Y_i^* \mid S_i^m = 0] - E[Y_i^* \mid S_i^m = 1, D_i^* = 0] \Big). \quad \square
\end{aligned}$$

A.3 Proof of Proposition 1

We first establish the decomposition (7), then verify the sign of Λ^{CPS}.

By iterated expectations,

$$E[Y_i^* \mid C_i = 0] = \omega_S \, E[Y_i^* \mid S_i = 0] + \omega_D \, E[Y_i^* \mid S_i = 1, D_i^* = 0].$$

Subtract $E[Y_i^* \mid C_i^m = 0]$ from both sides:

$$\begin{aligned}
\Delta^{CPS} = \omega_S &\Big(E[Y_i^* \mid S_i = 0] - E[Y_i^* \mid C_i^m = 0] \Big) \\
&+ \omega_D \Big(E[Y_i^* \mid S_i = 1, D_i^* = 0] - E[Y_i^* \mid C_i^m = 0] \Big).
\end{aligned}$$

For the first term, add and subtract $E[Y_i^* \mid S_i^m = 0]$:

$$\omega_S \Big(E[Y_i^* \mid S_i = 0] - E[Y_i^* \mid C_i^m = 0] \Big) = \omega_S \, \Delta^S + \omega_S \Big(E[Y_i^* \mid S_i^m = 0] - E[Y_i^* \mid C_i^m = 0] \Big).$$

For the second term, add and subtract $E[Y_i^* \mid S_i = 1, D_i^m = 0]$:

$$\omega_D \Big(E[Y_i^* \mid S_i = 1, D_i^* = 0] - E[Y_i^* \mid C_i^m = 0] \Big) = \omega_D \, \Delta^D + \omega_D \Big(E[Y_i^* \mid D_i^m = 0, S_i = 1] - E[Y_i^* \mid C_i^m = 0] \Big)$$

For the residual in the second term, add and subtract $E[Y_i^* \mid C_i^m = 0, S_i = 1]$:

$$\begin{aligned}
&E[Y_i^* \mid D_i^m = 0, S_i = 1] - E[Y_i^* \mid C_i^m = 0] \\
&\quad = \Big(E[Y_i^* \mid D_i^m = 0, S_i = 1] - E[Y_i^* \mid C_i^m = 0, S_i = 1] \Big) - \Big(E[Y_i^* \mid C_i^m = 0] - E[Y_i^* \mid C_i^m = 0, S_i = 1] \Big).
\end{aligned}$$

The first bracket is the second component of Λ^{CPS}; the second bracket is Ψ^{CPS}. Collecting all terms:

$$\Delta^{CPS} = \omega_S \, \Delta^S + \omega_D \, \Delta^D + \Lambda^{CPS} - \omega_D \, \Psi^{CPS}.$$

Under the stated condition on the algorithm and the threshold ordering $\tau^S \leq \tau^*$ (which follows from the multistage structure), we have $\{i : m(X_i) < \tau^S\} \subseteq \{i : m(X_i) < \tau^*\}$. The cases in the larger set but not the smaller have predicted risk in $[\tau^S, \tau^*)$, which is weakly higher than the predicted risk of cases below τ^S. Therefore $E[Y_i^* \mid S_i^m = 0] \leq E[Y_i^* \mid C_i^m = 0]$, and the first component of Λ^{CPS} is weakly negative. The same argument applied among screened-in calls gives $E[Y_i^* \mid D_i^m = 0, S_i = 1] \leq E[Y_i^* \mid C_i^m = 0, S_i = 1]$, so the second component is also weakly negative. Hence $\Lambda^{CPS} \leq 0$. □

A.4 Proof of Proposition 2

By iterated expectations,

$$E[Y_i^* \mid R_i = 1, A_i = 0] = \sum_j \alpha_j \, E[Y_i^* \mid R_i = 1, A_i = 0, J_i = j].$$

Add and subtract $E[Y_i^* \mid R_i = 1, A_{j,i}^m = 0, J_i = j]$ and $E[Y_i^* \mid R_i = 1, A_i^m = 0, J_i = j]$ within each term:

$$\begin{aligned} E[Y_i^* \mid R_i = 1, A_i = 0] = \sum_j \alpha_j \Big[& \underbrace{E[Y_i^* \mid R_i = 1, A_i = 0, J_i = j] - E[Y_i^* \mid R_i = 1, A_{j,i}^m = 0, J_i = j]}_{\Delta_j^A} \\ & + \underbrace{E[Y_i^* \mid R_i = 1, A_{j,i}^m = 0, J_i = j] - E[Y_i^* \mid R_i = 1, A_i^m = 0, J_i = j]}_{\Lambda_j^A} \\ & + E[Y_i^* \mid R_i = 1, A_i^m = 0, J_i = j] \Big]. \end{aligned}$$

Subtracting $E[Y_i^* \mid R_i = 1, A_i^m = 0]$ from both sides and using $\Delta^A = E[Y_i^* \mid R_i = 1, A_i = 0] - E[Y_i^* \mid R_i = 1, A_i^m = 0]$:

$$\Delta^A = \bar{\Delta}^A + \bar{\Lambda}^A + \sum_j \alpha_j \Big(E[Y_i^* \mid R_i = 1, A_i^m = 0, J_i = j] - E[Y_i^* \mid R_i = 1, A_i^m = 0] \Big),$$

which is the result. □

Table 1: Summary Statistics

	Screener Sample			Investigator Sample		
	All Calls	Screen In	Screen Out	All Calls	Case Opened	Case Not Opened
Panel A. Child Characteristics	(1)	(2)	(3)	(4)	(5)	(6)
White Child Present	0.761	0.728	0.784	0.728	0.754	0.725
Black Child Present	0.347	0.376	0.326	0.376	0.378	0.376
Female Child Present	0.617	0.620	0.615	0.620	0.646	0.617
Age of Youngest Child	8.656	8.319	8.856	8.319	7.681	8.396
Age of Oldest Child	10.368	10.291	10.413	10.291	10.088	10.316
Prior Investigation	0.778	0.781	0.777	0.781	0.883	0.768
Panel B. Call Characteristics						
Number of Children	1.598	1.679	1.551	1.679	1.833	1.660
Physical Abuse Allegation	0.219	0.239	0.208	0.239	0.212	0.243
Physical Neglect Allegation	0.171	0.155	0.180	0.155	0.151	0.156
Improper Supervision Allegation	0.683	0.581	0.743	0.581	0.638	0.574
Panel C. Call Outcomes						
Screening Rate	0.372	1.000	0.000	1.000	1.000	1.000
Substantiation Rate	0.085	0.229	0.000	0.229	1.000	0.135
Case Opening Rate	0.040	0.108	0.000	0.108	1.000	0.000
Child Removal Rate	0.009	0.024	0.000	0.024	0.219	0.000
Panel D. Call Outcomes if Case Not Opened						
New Call in 31-180 Days	0.338	0.305	0.356	0.305	–	0.305
Number of Calls	524,657	195,355	329,302	195,355	21,169	174,186
Number of Children	838,650	328,022	510,628	328,022	38,799	289,223

Notes. This table summarizes the screener and investigator analysis samples. The screener sample includes 524,657 hotline calls in Michigan from 2017 to 2022, assigned to 292 hotline screeners. The investigator sample includes 195,355 investigations (screened-in calls) from 2017 to 2022, assigned to 8,753 investigators. Rows describing the nature of the allegation and the alleged perpetrator can sum to more than one because the categories are not mutually exclusive. See the text for additional details on the sample and variable definitions.

Table 2: System Decomposition Results

	Estimate	Description
Δ^{CPS}	1.2%	System-wide excess maltreatment risk relative to algorithm.
	(0.5)	
$\omega_S \times \Delta^S$	6.7%	Contribution of screening stage.
	(0.4)	
$\omega_D \times \Delta^D$	2.9%	Contribution of investigation stage.
	(0.2)	
$\omega_D \times \Psi^{CPS}$	2.9%	Selection effect between screening and investigation stages.
	(0.2)	
$\overline{\Lambda}^{CPS}$	-5.6%	Threshold adjustment between screening and investigation stages.
	(0.5)	
ω_S	67.8%	Share of non-opened cases due to screening stage.
	(0.1)	
ω_D	32.2%	Share of non-opened cases due to investigation stage.
	(0.1)	

Notes. This table reports the system quality decomposition from Proposition 1. Δ^{CPS} is the difference in subsequent maltreatment rates among children without an opened case under CPS versus under an algorithm that opens cases at the same rate but ranks by predicted risk. Rows 2–5 decompose Δ^{CPS} into four components. Standard errors in parentheses are estimated by bootstrap.

Table 3: Screening Stage Quality Estimates

Estimand	Estimate	Description
Δ^S	9.9%	Screening stage excess maltreatment risk relative to algorithm.
	(0.5)	
$\overline{\Delta}^S$	9.8%	Contribution of screener ranking quality.
	(0.2)	
$\overline{\Lambda}^S$	0.1%	Contribution of screener threshold adjustments.
	(0.5)	
Σ^S_Δ	0.5%	Std. dev. of screener ranking quality.
	(0.2)	
Σ^S_Λ	0.0%	Std. dev. of screener threshold adjustment.
	(0.1)	
Share worse than algorithm	100.0%	Share of screeners outperformed by algorithm.
	(0.0)	

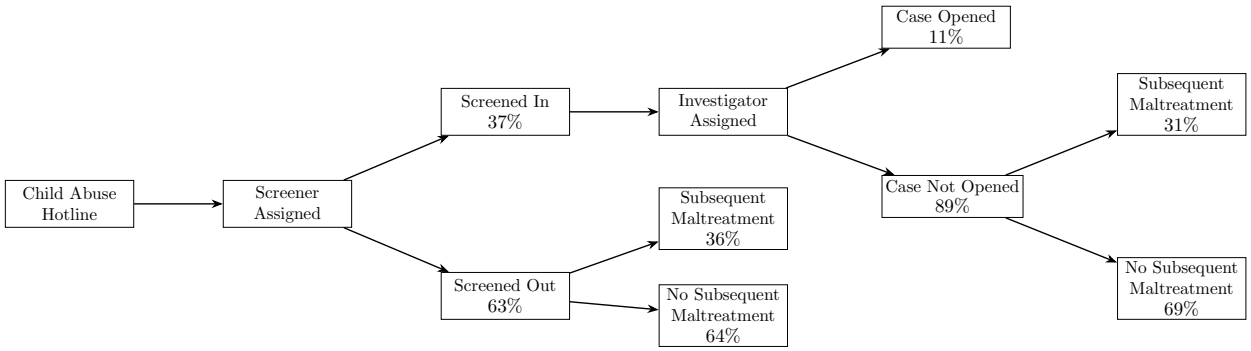
Notes. This table reports the screening stage quality decomposition from Proposition 2. Δ^S is the difference in subsequent maltreatment rates among screened-out children under screeners versus under an algorithm that screens in at the same rate but ranks by predicted risk. $\overline{\Delta}^S$ and $\overline{\Lambda}^S$ decompose Δ^S into ranking quality and leniency components; a composition residual is omitted. Heterogeneity estimates (Σ^S_Δ, Σ^S_Λ, share worse than algorithm) are based on empirical Bayes. Standard errors in parentheses are estimated by bootstrap.

Table 4: Investigation Stage Quality Estimates

Estimand	Estimate	Description
Δ^D	9.0%	Investigation stage excess maltreatment risk relative to algorithm.
	(0.5)	
$\overline{\Delta}^D$	8.9%	Contribution of investigator ranking quality.
	(0.3)	
$\overline{\Lambda}^D$	0.1%	Contribution of threshold adjustments.
	(0.4)	
Σ_Δ^D	8.1%	Std. dev. of investigator ranking quality.
	(0.3)	
Σ_Λ^D	1.1%	Std. dev. of investigator threshold adjustments.
	(0.0)	
Share worse than algorithm	84.7%	Share of investigators outperformed by algorithm.
	(0.9)	

Notes. This table reports the investigation stage quality decomposition from Proposition 2. Δ^D is the difference in subsequent maltreatment rates among screened-in children without an opened case under investigators versus under an algorithm that opens cases at the same rate but ranks by predicted risk. $\overline{\Delta}^D$ and $\overline{\Lambda}^D$ decompose Δ^D into ranking quality and leniency components; a composition residual is omitted. Heterogeneity estimates (Σ_Δ^D, Σ_Λ^D, share worse than algorithm) are based on empirical Bayes. Standard errors in parentheses are estimated by bootstrap.

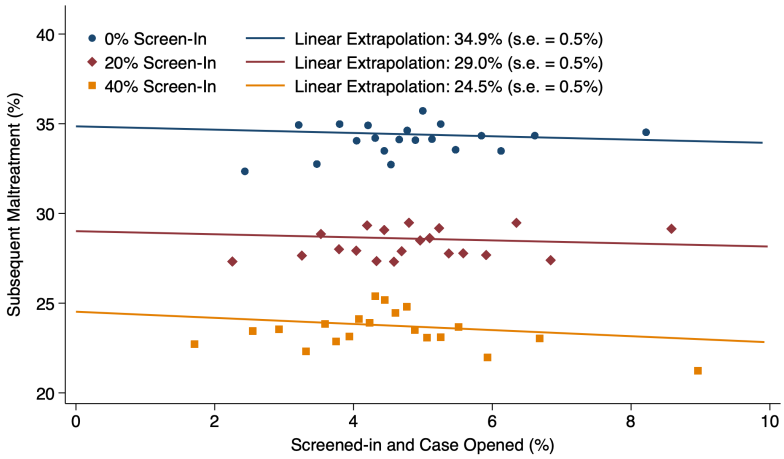
Figure 1: Child Protective Services Process in Michigan



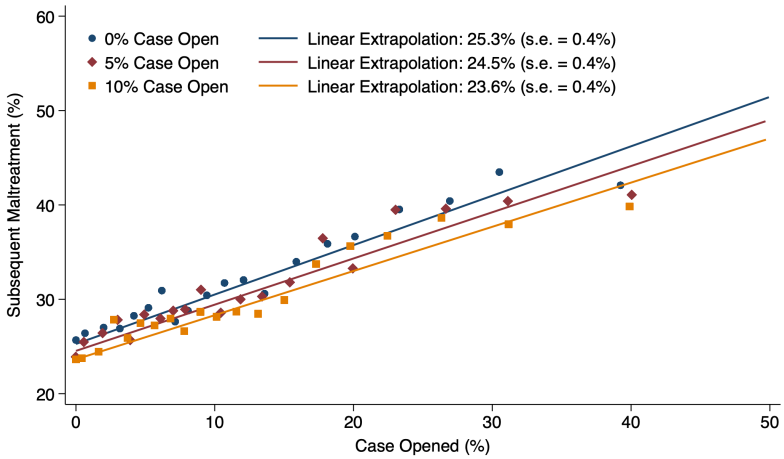
Notes. The figure describes the child protection process in Michigan. Both screeners and investigators are quasi-randomly assigned, as described in the text. The percentages on screening-in and screening-out refer to all calls received; percentages thereafter refer to investigated cases.

Figure 2: Extrapolation Estimates

A. Screener Sample

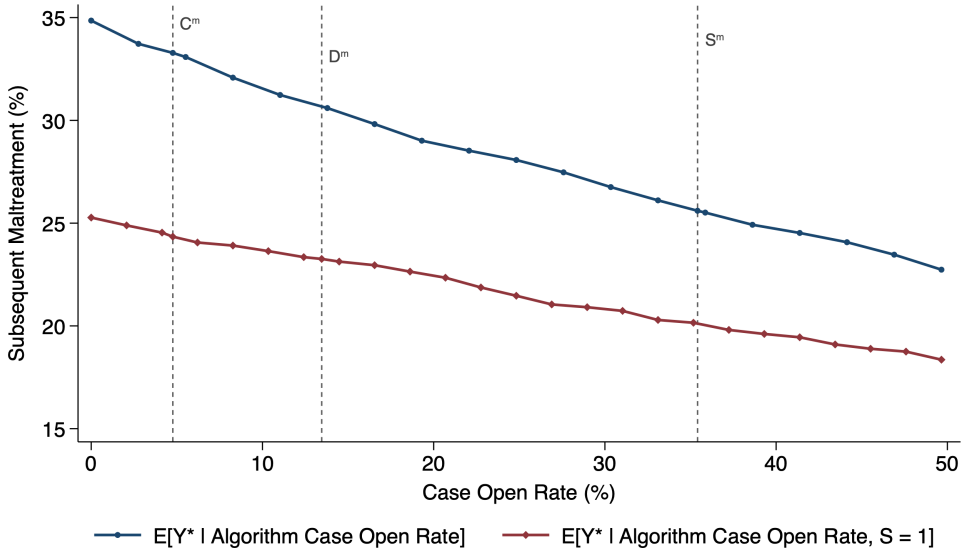


B. Investigator Sample



Notes. This figure shows binscatters (with 20 equal-sized bins) of screener- and investigator-level conditional subsequent maltreatment rates against corresponding decision rates and overlays extrapolation fits used to estimate counterfactual maltreatment risk at policy-relevant thresholds. Subsequent maltreatment is measured as a new hotline call between 30 and 180 days after the initial hotline call among calls where a case was not opened. The solid and dashed lines are linear and local linear extrapolation curves fit to the underlying unit-level data, with inverse-weight by the variance of the estimated subsequent maltreatment rate among calls where a case was not opened. The local linear regression uses a Gaussian kernel with a fixed bandwidth. Screener-level regressions adjust for day-by-shift fixed effects; investigator-level regressions adjust for month-by-zipcode-by-shift fixed effects. Panel A reports results for the algorithm threshold matching the overall case-open rate in the screener sample. Panel B reports results for the algorithm threshold matching the case-open rate in the investigator sample, conditional on being screened-in. We also report the estimated intercept and standard error at a cutoff-specific decision rate under each extrapolation, which equals the estimated average subsequent maltreatment rate for the relevant sample of calls. Standard errors are estimated by taking random draws from the distributions of the estimated screener- and investigator-specific decision rates and subsequent maltreatment rates and then recalculating the threshold-specific extrapolations.

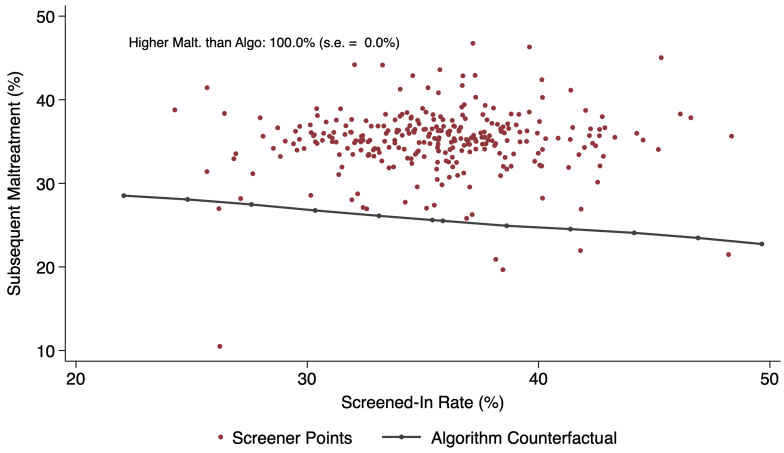
Figure 3: Conditional Maltreatment Rates Under the Algorithm



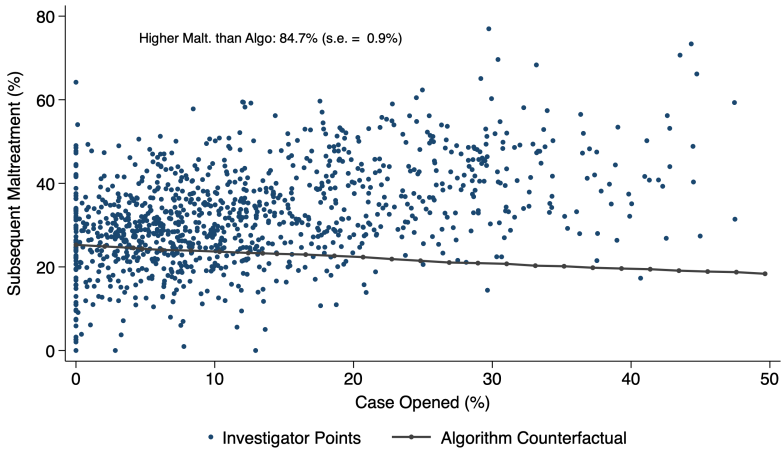
Notes. This figure plots extrapolated counterfactual subsequent maltreatment rates across algorithmic thresholds. The top blue line plots the algorithm's subsequent maltreatment as if it were to open cases directly on calls from the screener sample. The bottom red line plots the algorithm's subsequent maltreatment as if it were to open cases only on referrals that made it to the investigation stage. Subsequent maltreatment rates under the algorithmic case open rule are estimated using linear extrapolations of mean risk at different risk score cutoffs. Screener-level regressions adjust for day-by-shift fixed effects; investigator-level regressions adjust for month-by-zipcode-by-shift fixed effects. Vertical dashed lines mark the overall case open rate on screened-in sample (C^m), screen-in rate (S^m), and case open rate conditional on being screened-in (D^m).

Figure 4: Conditional Maltreatment Rates Relative to the Algorithm

A. Screener Sample



B. Investigator Sample



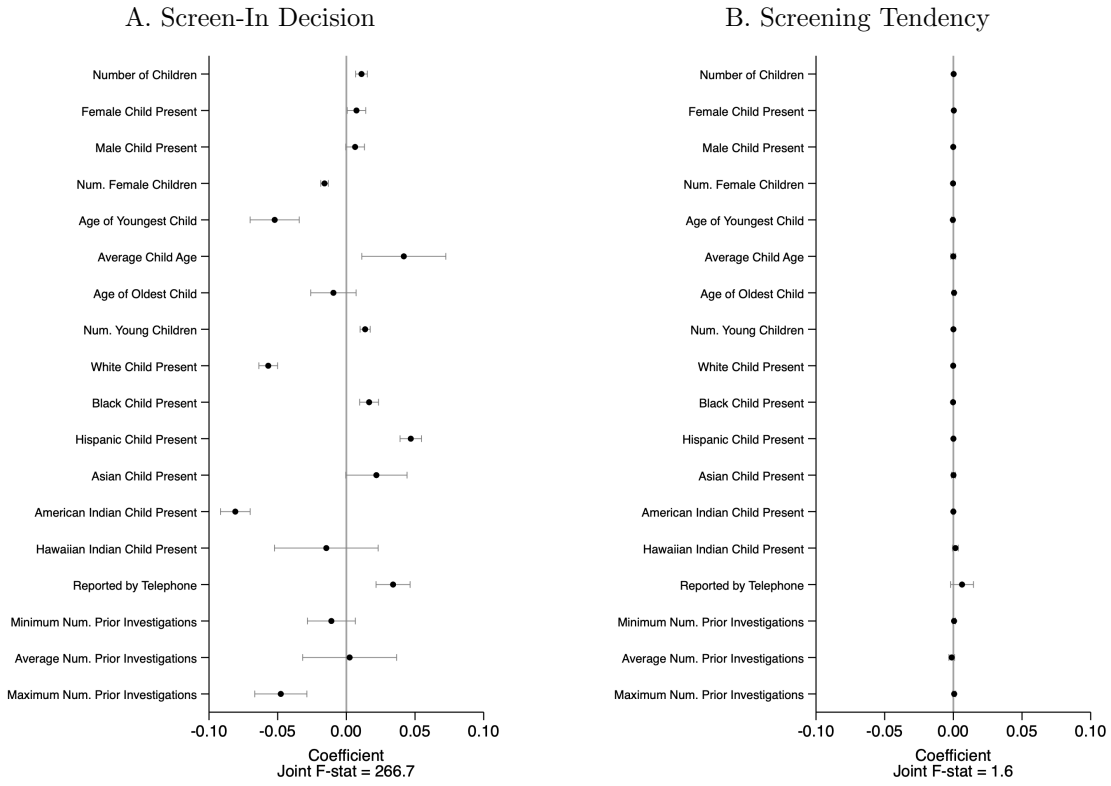
Notes. This figure plots observed subsequent maltreatment rates among screened-out calls for screeners (Panel A) or calls where a case was not opened for investigators (Panel B), along with counterfactual subsequent maltreatment rates among screened-out calls (Panel A) or calls where a case was not opened (Panel B) for algorithmic decisions. Subsequent maltreatment rates under the algorithmic rule are estimated using linear extrapolations of mean risk at different risk score cutoffs. Screener-level regressions adjust for day-by-shift fixed effects; investigator-level regressions adjust for month-by-zipcode-by-shift fixed effects. This figure also reports the fraction of screeners and investigators with higher subsequent maltreatment rates compared to the algorithmic rule using the posterior average effects approach described in the main text. Standard errors are estimated by taking random draws from the distributions of the estimated screener- and investigator-specific distance estimates and then recalculating the statistics of interest.

Appendix Table A.1: Supplementary Decomposition Measures

	(1)
Panel A: Observed Measures	
$\mathbb{E}[S_i]$	35.4%
	(0.1)
$\mathbb{E}[D_i \mid S_i = 1]$	13.6%
	(0.2)
$\mathbb{E}[C_i]$	4.8%
	(0.0)
$\mathbb{E}[Y_i^* \mid S_i = 0]$	35.5%
	(0.1)
$\mathbb{E}[Y_i^* \mid S_i = 1, D_i = 0]$	32.3%
	(0.3)
$\mathbb{E}[Y_i^* \mid C_i = 0]$	34.5%
	(0.1)
Panel B: Decomposition Extrapolation Measures	
$\mathbb{E}[Y_i^* \mid m(X) \leq \tau^*]$	33.3%
	(0.5)
$\mathbb{E}[Y_i^* \mid m(X) \leq \tau^S]$	25.6%
	(0.5)
$\mathbb{E}[Y_i^* \mid m(X) \leq \tau^D, S = 1]$	23.3%
	(0.4)
$\mathbb{E}[Y_i^* \mid m(X) \leq \tau^*, S = 1]$	24.3%
	(0.4)

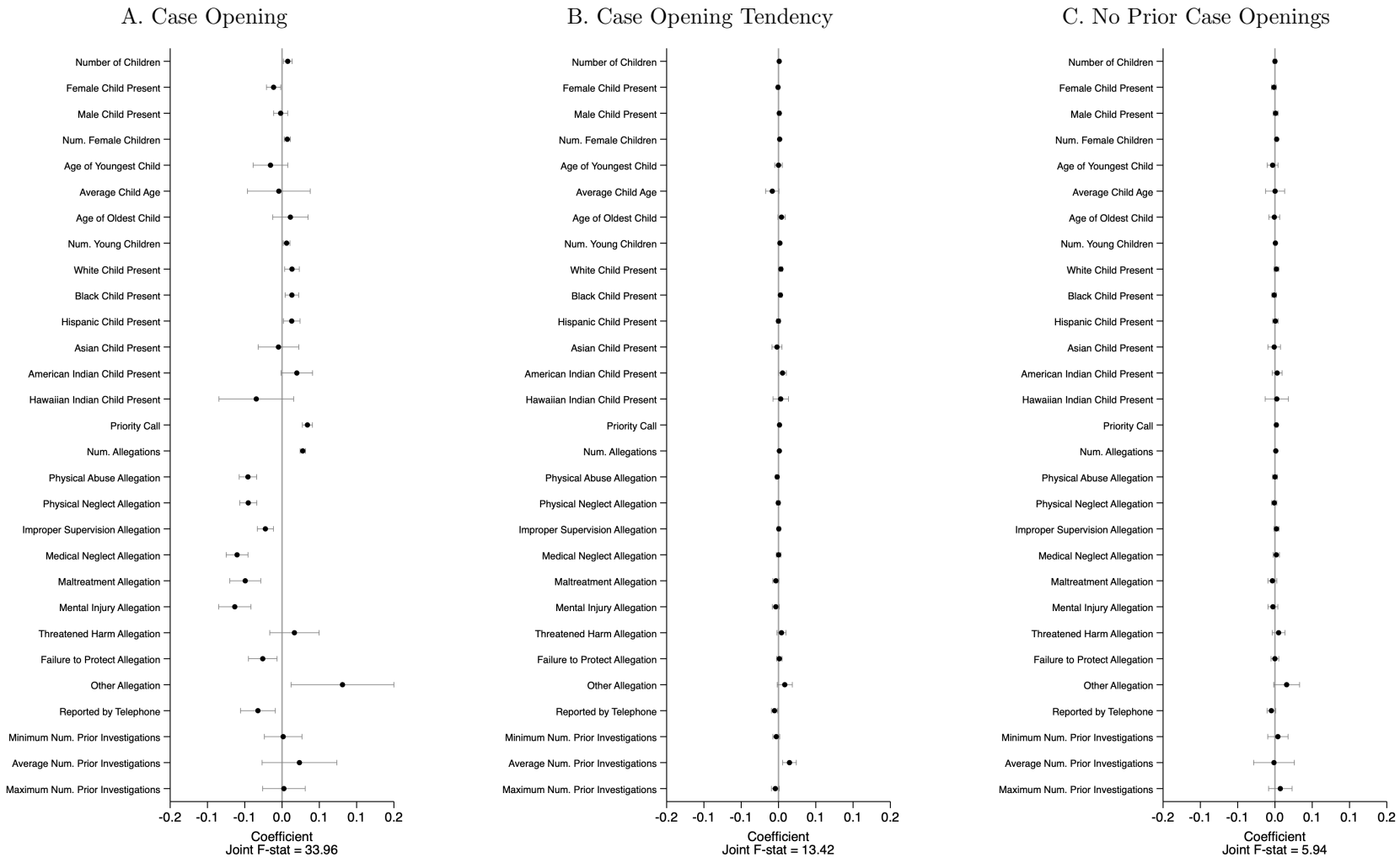
Notes. This table reports supplementary measures underlying the system decomposition in Table 2. Panel A reports observed rates from the analysis sample. $\mathbb{E}[S_i]$ is the screen-in rate; $\mathbb{E}[D_i \mid S_i = 1]$ is the case-opening rate conditional on investigation; $\mathbb{E}[C_i]$ is the overall case-opening rate. $\mathbb{E}[Y_i^* \mid S_i = 0]$, $\mathbb{E}[Y_i^* \mid S_i = 1, D_i = 0]$, and $\mathbb{E}[Y_i^* \mid C_i = 0]$ are subsequent maltreatment rates among screened-out calls, screened-in calls without an opened case, and all calls without an opened case, respectively. Panel B reports counterfactual algorithmic maltreatment rates estimated by linear extrapolation. $\mathbb{E}[Y_i^* \mid m(X) \leq \tau^*]$ is the estimated subsequent maltreatment rate among calls the algorithm would not open at the system-wide threshold τ^*, calibrated to match $\mathbb{E}[C_i]$. $\mathbb{E}[Y_i^* \mid m(X) \leq \tau^S]$ is the analogous benchmark at the screening-stage threshold τ^S, calibrated to match $\mathbb{E}[S_i]$, and $\mathbb{E}[Y_i^* \mid m(X) \leq \tau^D, S = 1]$ is the benchmark at the investigation-stage threshold τ^D, calibrated to match $\mathbb{E}[D_i \mid S_i = 1]$ among screened-in calls. $\mathbb{E}[Y_i^* \mid m(X) \leq \tau^*, S = 1]$ is the estimated maltreatment rate among screened-in calls the algorithm would not open at the system-wide threshold. Standard errors in parentheses are estimated by bootstrap.

Appendix Figure A.1: Screener Balance Test



Notes. This figure shows coefficients and 95% confidence intervals from regressions of screen-in decision (Panel A) and the assigned screener's leave-out screen-in propensity (Panel B) on covariates, controlling for day-by-shift fixed effects. Allegation type variables are not included as these are encoded by the assigned screener. Both panels use all of the test sample calls. Continuous covariates are standardized so that they have standard deviations equal to 1. At the bottom of each panel, we report the F-statistic from the joint F-test of all covariates.

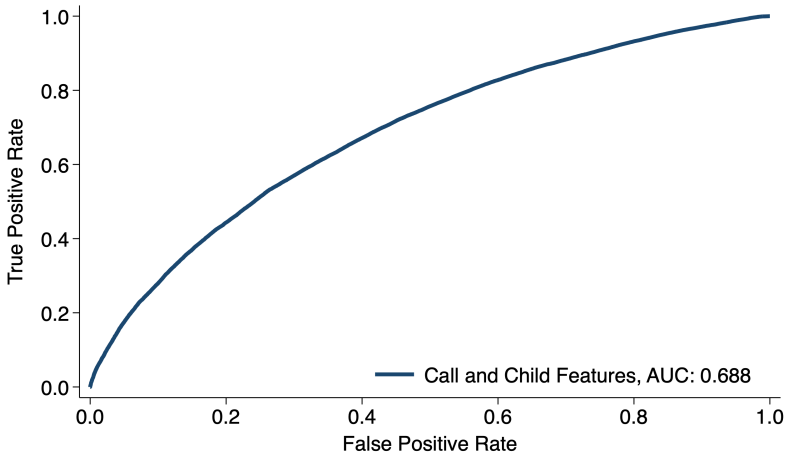
Appendix Figure A.2: Investigator Balance Test



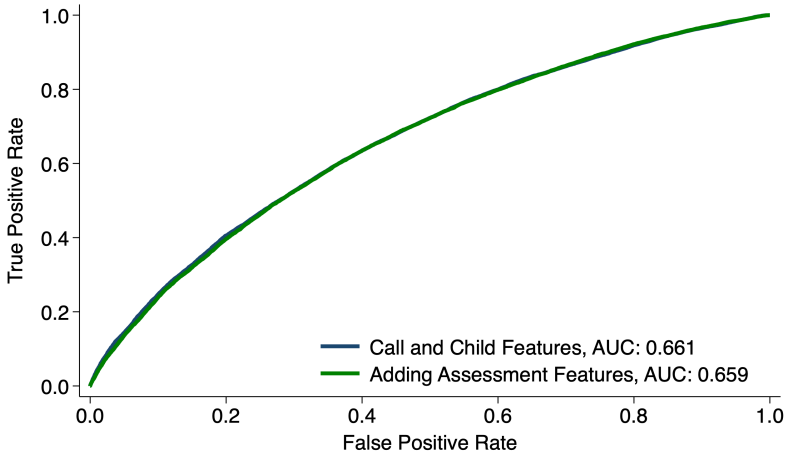
Notes. This figure shows coefficients and 95% confidence intervals from regressions of case opening decision (Panel A) and the assigned investigator's leave-out case open propensity (Panels B and C) on covariates, controlling for month-by-zipcode-by-shift. Panels A and B use all of the test sample screened-in calls. Panel C restricts to test sample screened-in calls of which none of the children involved in call had a prior substantiation. Continuous covariates are standardized so that they have standard deviations equal to 1. At the bottom of each panel, we report the F-statistic from the joint F-test of all covariates.

Appendix Figure A.3: Algorithmic Performance

A. Screener Sample



B. Investigator Sample



Notes. This figure reports the ROC curve and AUC for prediction of a subsequent hotline call between 30 and 180 days after the focal call. Algorithms are trained in the training sample and then evaluated in the held-out test sample. The figure is limited to referrals where a case was not opened so the outcome is observed. The initial algorithm for both samples uses child- and call-level features, such as allegation type, prior child welfare history, and reporter type. We add risk and safety assessment categories, including item-specific risk responses, from the investigator to the algorithm for the second algorithm in the investigation sample. All algorithms are trained using XGBoost.