DEVELOPING PREDICTIVE RISK MODELS to Support Child Maltreatment Hotline Screening Decisions

UPDATED APRIL 2019

In August 2016, the Allegheny County Department of Human Services (DHS) implemented the *Allegheny Family Screening Tool* (AFST), a predictive risk modeling tool designed to improve child welfare call screening decisions. The AFST was the result of a two-year process of exploration about how existing data could be used more effectively to improve decision-making at the time of a child welfare referral. More information can be found here about the AFST.

The process began in 2014 with a Request for Proposals and selection of a team from Auckland University of Technology led by Rhema Vaithianathan and including Emily Putnam-Hornstein from University of Southern California, Irene de Haan from the University of Auckland, Marianne Bitler from University of California – Irvine and Tim Maloney and Nan Jiang from Auckland University of Technology. Input was solicited throughout the exploration and development process and used to inform the final product. Prior to implementation, the model was subjected to an ethical review by Tim Dare of the University of Auckland and Eileen Gambrill of the University of California-Berkeley. Upon the conclusion of this review, to which DHS prepared a response, the developers proceeded with implementation.

Concurrent with this process was the issuance of a second Request for Proposals, at the end of 2015, for an impact and process evaluation of the model. Awarded the contracts were Stanford University (impact evaluation) and Hornby Zeller Associates (process evaluation). Both the process evaluation and the impact evaluation have been completed.

The first version of the AFST (V1) was utilized from August 2016 through November 2018. In December 2018, an improved model, V2, was implemented and included changes to the target outcome, data sources, and visualization of the tool.

Development, implementation and evaluation of the AFST were made possible by a public/private funding partnership that included generous support from the Richard King Mellon Foundation, Casey Family Programs and the Human Services Integration Fund, a collaborative funding pool of local foundations under the administrative direction of The Pittsburgh Foundation.

This publication includes eight documents:

SECTION 1 (APRIL 2017)
Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation
prepared by Rhema Vaithianathan, PhD; Nan Jiang, PhD; Tim Maloney, PhD; Parma Nand, PhD; and Emily Putnam-Hornstein, PhD

SECTION 2 (APRIL 2017)
Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County
by Tim Dare and Eileen Gambrill

SECTION 3 (APRIL 2017)
Response to Ethical Analysis
by the Allegheny County Department of Human Services

SECTION 4 (JANUARY 2018)
Allegheny County Predictive Risk Modeling Tool Implementation: Process Evaluation
by Hornby Zeller Associates, Inc.

SECTION 5 (APRIL 2019)
Impact Evaluation Summary
by the Allegheny County Department of Human Services

SECTION 6 (APRIL 2019)
Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office
by Jeremy D. Goldhaber-Fiebert, PhD and Lea Prince, PhD

SECTION 7 (APRIL 2019)
Allegheny Family Screening Tool: Methodology, Version 2
prepared by Rhema Vaithianathan, PhD; Emily Kulick; Emily Putnam-Hornstein, PhD; and Diana Benavides Prado

SECTION 8 (APRIL 2019)
Frequently-Asked Questions
by the Allegheny County Department of Human Services

Each document may be viewed independently, but together they provide an overview of the process and thinking that went into the development and implementation of the AFST and the conclusions and recommendations of the independent evaluators.

CENTRE FOR SOCIAL DATA ANALYTICS

Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation

April 2017

Rhema Vaithianathan, PhD

Emily Putnam-Hornstein, PhD

Nan Jiang, PhD

Parma Nand, PhD

Tim Maloney, PhD

AUT UNIVERSITY NEW ZEALAND

CENTRE FOR SOCIAL DATA ANALYTICS

Contents

Background....4
Current Practice....5
Calls to the Child Protection Hotline....5
County Screening of Maltreatment Allegations....6
Re-referrals and Placements of Children and Victims....7
Latent Risk vs. Observed Risk....7
Determining the Target Outcome of A PRM....8
Data....11
Methodology for Placement and Re-Referral Model....13
Alternative Methods Considered....14
Race....15
Model Performance....15
Placement Model....15
Re-referral Model....17
Concerns over Policy Changes in 2015....18
External Validation of the Model....19
External Validation: Hospitalisation....19
External Validation: Critical Events....23
Comparison to Structured Decision Making and Rule-based/Threshold Approaches....24
Implementation of the Risk Score....26
Mandatory Screen-In....27
Impact of Race as a Predictor....29
Using the Model in Practice....30
Technical Implementation....31
Training....32
Next Steps: Six Month Rebuild and Adding a Random Forest Model....33
Conclusion....34
Appendix: Variables Used in the Allegheny Child Welfare Predictive Risk Model....36
Appendix: Hospital Injury Classifications....43
References....44

AUT UNIVERSITY NEW ZEALAND

CENTRE FOR SOCIAL DATA ANALYTICS

List of Figures

Figure 1-6: Proportion of Selected Hospital Injury Events for Children Referred to Allegheny County by Maximum Placement Scores 20

Figure 7: Stata output from Estimate of Model 1 23

Figure 8: Screen Shots of the Family Risk Score 27

Figure 9: Referral Progression Process 30

Figure 10: Technical Implementation of the Screening Tool 31

List of Tables

Table 1: GPS Referral Dispositions (Between April 1, 2010 and May 4, 2016) 6

Table 2: Re-referral and Placement Rates Within 2 Years (victims and children in referrals between March 1, 2010 and April 29, 2014) 7

Table 3: Area under ROC curve of Placement PRM (validation sample only, probit and boosted regressions, including race variables) 16

Table 4: Area under ROC curve of Placement PRM (validation sample only, probit regressions, excluding race variables) 16

Table 5: Area under ROC curve of Re-referral PRM (validation sample only) 17

Table 6: Area under ROC curve of Re-referral PRM (validation sample only, probit regressions, excluding race variables) 17

Table 7: Mean-Maximum Referral Score by year (All referrals) 18

Table 8: Placement Score of Admitted Children who were also referred to Child Welfare 22

Table 9: Comparison of SDM with Allegheny County Model 24

Table 10: Threshold Model vs. PRM for identifying “high risk” referral 26

Table 11: Screening Score Groups and Act 33 28

Table 12: Screening Score Groups and Outcomes (all sample of referrals, no race model). 28

Table 13: Screening Score Groups and Outcomes (all sample of referrals, With race model). 29

Table 14: Comparison of those who were placed and flagged as mandatory screen-in risk group 34

CENTRE FOR SOCIAL DATA ANALYTICS

BACKGROUND

Predictive Risk Modelling (PRM) uses routinely collected administrative data to model future adverse outcomes that might be prevented through a more strategic delivery of services. PRM has been used previously in health and hospital settings (Panattoni, Vaithianathan, Ashton, & Lewis, 2011; Billings, Blunt, Steventon, Georghiou, Lewis, & Bardsley, 2012) and has been suggested as a potentially useful tool that could be translated into child protection settings (Vaithianathan, Maloney, Putnam-Hornstein, & Jiang, 2013). In the context of child protective services, PRM tools can be used to help child protection staff make better initial screening and service decisions for children who have been named in reports of alleged abuse or neglect. Specifically, PRM can be deployed at the point that a referral is received by a child protection hotline. These referrals are typically made when someone in the community (e.g., a neighbor or a mandated professional such as a teacher) is concerned that a child has been the victim of abuse or neglect.

In 2014, Allegheny County's Department of Human Services issued a Request for Proposals focused on the development and implementation of tools that would enhance use of the County's integrated data system. Specifically, the County sought proposals that would: (1) improve the ability to make efficient and consistent data-driven service decisions based on County records, (2) ensure public sector resources were being equitably directed to the County's most vulnerable clients, and (3) promote improvements in the overall health, safety and well-being of County residents. A consortium of researchers from Auckland University of Technology (AUT: Vaithianathan, Jiang, Maloney), the University of Southern California (USC: Putnam-Hornstein), the University of California at Berkeley (UCB: Gambrill), and the University of Auckland (UA: Dare) submitted a proposal outlining a scope of work focused on the use of PRM to support decisions made at the time a child has been reported for alleged abuse or neglect. This team was awarded the contract in the Fall of 2014 and commenced work in close concert with the Allegheny County team.

In mid-2015, it was decided that the most promising, ethical, and readily implemented use of PRM within the Allegheny County child protection context was one in which a model would be deployed at the time an allegation of maltreatment was received at the hotline. The objective was to develop a decision aid to support hotline screeners in determining whether a maltreatment referral is of sufficient concern to warrant an in-person investigation. The present report describes the methodology used to develop and implement this model, the Allegheny Screening Tool.

It should be noted that while in some settings machines have been used to *replace* decisions that were previously made by humans, this is not the case for the Allegheny Family Screening Tool. It was never intended or suggested that the algorithm would replace human decision-making. Rather, that the model should help to inform, train and improve the decisions made by the child protection staff.

CENTRE FOR SOCIAL DATA ANALYTICS

CURRENT PRACTICE

Allegheny County's Department of Human Services is unique in the United States: it has an integrated client service record and data management system. This means that the County's child protection hotline staff are already able to access and use historical and cross-sector administrative data (e.g., child protective services, mental health services, drug and alcohol services, homeless services) related to individuals associated with a report of child abuse or neglect. Although this information is critical to assessing child risk and safety concerns, it is challenging for County staff to efficiently access, review, and make meaning of all available records. Beyond the time required to scrutinize data for every individual associated with a given referral (e.g., child victim, siblings, biological parents, alleged perpetrator, other adults living at the address where the incident occurred), the County has no means of ensuring that available information is consistently used or weighted by staff when making hotline screening decisions. As such, for example, recent paternal criminal justice involvement that surfaces in the context of one child's referral may factor into a decision to investigate a report of maltreatment, while for another child that same information could be completely ignored.

To help the reader understand the context in which the new PRM tool will be implemented, a short summary of current screening practice has been provided below.

Calls to the Child Protection Hotline

A referral for suspected child abuse or neglect is received by Allegheny County either via the Pennsylvania State Hotline (i.e., ChildLine) or directly through the County's local hotline. Allegations made to the State Hotline are emailed to the County's local hotline staff. Allegations can be classified as falling under the State's: (1) "child protective service" (CPS) (23 Pa.C.S. § 6303) or (2) "general protective services" (GPS) (23 Pa.C.S. § 6334) statutes. Designation under CPS means that the allegation includes abuse or severe neglect and automatically meets the statutory threshold for it to be screened-in for investigation. For the 2015 year, we find that 17% of all reports in Allegheny County were designated as allegations falling under CPS statutes.

Child maltreatment referrals, whether defined as CPS or GPS, typically identify a variety of individuals. These individuals typically include the alleged child victim(s), the biological mother and father of the alleged victim, the perpetrator (who may or may not be a biological parent), other related and unrelated children in the home, and other adults who may also be residing at the address.

CENTRE FOR SOCIAL DATA ANALYTICS

County Screening of Maltreatment Allegations

If the maltreatment allegation is classified as falling under CPS statutes based on the information reported, then local County screening staff have no further decision-making authority and a child maltreatment investigation must begin within 24 hours. If, however, an allegation is classified as GPS, then County hotline staff (i.e., screener and supervisor) have the joint discretion to respond by: (1) *screening-out* the allegation without any further evaluation or assessment (if there are no children age 6 or younger in the household[1]), (2) *conducting a field screen* of the maltreatment allegation in order to evaluate the safety and well-being of the child and determine whether a full investigation is warranted, or (3) *conducting a formal investigation* of the maltreatment allegation to determine if maltreatment has occurred and there is a potential for future harm to the child. As such, the *screening-in* of a maltreatment allegation is synonymous with conducting a formal "investigation." Meanwhile, following the field screen, a decision is made to screen-in or screen-out the referral.

For GPS reports that are screened in for investigation (either at the outset or after a field screen has been conducted), the report is transferred from the County's hotline office and assigned to one of five regional child welfare offices (typically on the basis of the report's geographic origins) or remains with the intake office so that a formal investigation can be conducted.

To provide a sense of the distribution of maltreatment reports, and the subsequent screening decisions that were made, in Table 1 we present historical data for the period from April 1, 2010 through May 4, 2016 (only for GPS). The table illustrates that a majority of GPS reports (52%) are screened out.

Table 1: GPS Referral Dispositions (Between April 1, 2010 and May 4, 2016)

	Total Numbers in Each Category	% of Total Referrals
Total Screened In	55,513	48%
Total Screened Out (1)	60,923	52%
Total Referrals (with call screening reason given)	116,436	100%

[1] Allegheny County has had a rule that any GPS report involving a child age 6 or younger cannot be screened out without first having a *field screen*. This decision reflects recognition that the vast majority of critical and fatal maltreatment events occur to children in this age group. Upon implementation of this tool, the field screen policy has been modified. Field screens are now conducted when (a) reports involve children age 3 and younger who are impacted by the allegations, (b) when a report is the fourth referral for a family within two years and there has not been a previous investigation, (c) when a report involves children who are in cyber/home school, or (d) whenever call screening staff would like more information about the allegations, children, or family. Notes to table: (1) Screen out reasons include, but are not limited to, information does not meet the legal definition of child maltreatment and no risk of maltreatment or safety concerns noted after a field screen was conducted. Table excludes those that are CPS and therefore automatically screened in.

CENTRE FOR
SOCIAL DATA ANALYTICS

Re-referrals and Placements of Children and Victims

Table 2 shows the re-referral and placement rates of children[2] in a referral, based on their initial disposition. The second row shows that among all children and victims included in a referral (between March 1, 2010 and April 29, 2014) that was opened for investigation, approximately 1 in 2 experienced a follow-up allegation of maltreatment and roughly 1 in 8 were subsequently placed within 2 years of the first referral.

As expected, those children who were screened out had a higher chance of being re-referred than those who were screened in (53% vs. 45%). By contrast, those who were initially screened in have a higher chance of being placed within 2 years than those who were initially screened out (13% vs 5%).

Table 2: Re-referral and Placement Rates Within 2 Years (victims and children in referrals between March 1, 2010 and April 29, 2014)

	Re-referred within 2 years (%)	Placed within 2 years (%)
Screened In	45%	13%
Screened Out (1)	53%	5%
Average	49%	9%

(1) Screen out reasons include, but are not limited to, information does not meet the legal definition of child maltreatment and no risk of maltreatment or safety concerns noted after a field screen was conducted. Table excludes those that are CPS and therefore automatically screened in.

Latent Risk vs. Observed Risk

At hotline screening, a child is assessed for evidence that abuse or neglect has occurred and the probability that the child will experience future harm if no services are provided and/or no action is taken. If the probability of future harm is elevated above a given (admittedly normative and context-specific) threshold, then the County may be justified in acting to serve the family and protect the child in either a voluntary or involuntary manner.

Theoretically, developing a predictive model for this underlying "latent" risk of future harm would require a research data set where no actions (or "interventions") had been taken following the initial maltreatment referral (e.g., investigations, services, placements in foster care). We would then follow these children for two years and see which

[2] Discussions with Allegheny County staff suggest that the role of "victim" does not always identify the only victim in a GPS referral. Often, the victims of GPS referrals include all children (e.g., all children are impacted by parental substance abuse or homelessness), but not all children are called "victim child" in a referral consistently. Call screening staff, however, are making determinations about the risk and safety of all children involved in a call. Therefore, it was determined that the modelling would assess the risk of each child in the referral (whether denoted as victim or child). Therefore, in this document we use the term children to denote those cases where we are discussing anyone in a referral that is denoted a child as well as a victim.

AUT UNIVERSITY NEW ZEALAND

CENTRE FOR SOCIAL DATA ANALYTICS

children went on to experience future abuse, neglect, or other forms of maltreatment and harm. For example, when building a PRM tool for hospital readmission risk, it is typical to use a sample of patients who do not access any kind of post-discharge services so that one can try and identify risk factors that contribute to readmission.

Such a research dataset, however, is never available in the child protection context. At initial hotline screening, decisions are made that influence the child's future trajectory and future risk of harm. Therefore, careful consideration must be given to modelling the outcome that is being predicted in order *not* to predict outcomes that are simply re-producing past decisions made by hotline screening staff.

The challenges related to this should not be understated. In the available historical data for Allegheny County, children are not left alone. Indeed, half of children are screened in for investigation at the time of the initial maltreatment referral used for modelling purposes. Their subsequent course of events is therefore dictated by a series of decisions and actions taken by the child protection system. The risk factors that can then be identified are a combination of the risk factors that reflect latent risk *and* factors that capture hotline screening decisions. To address this, predictions must be developed *conditional* on these historical decisions that influence the outcomes observed.

Determining the Target Outcome of a PRM

While there is not universal agreement on the degree to which the current clinical assessment at point of referral is focused on the longer-term risk of adverse events versus assessing the current crisis of alleged abuse or neglect, the research team and Allegheny County chose to design a model to predict long arc risk. This decision was made because the logic of predictive risk modelling from the health literature is that it is a way of supplementing clinical decision-making. By offering clinicians a risk score that stratifies that the patient is at long term risk of, for example, readmission to hospital, the clinicians could be alerted to looking at the wider context of patient's situation than simply the current medical crisis that brought the patient to the attention of the clinician. Similarly, targeting the PRM on long arc-risk complements the role of the screening staff who are focused on the information about the allegation contained in the referral.

The predictive risk model is designed to support hotline screening staff to determine which reports of maltreatment involve children who are at greatest risk of: (1) future abuse and neglect, (2) future involvement with child protective services, and/or (3) future critical incidents (i.e., near-fatalities and fatalities). Information concerning the statistical probability that a given child will experience one or more of these future events is valuable as these are arguably

CENTRE FOR SOCIAL DATA ANALYTICS

outcomes that all child protection systems seek to prevent.[3] As such, this information can be used to establish statistical thresholds that help prioritize and sort reports of alleged maltreatment into those in which the action of carrying out a full investigation seems particularly warranted and those in which screening out may be justified. Before determining how to operationalize, predict, and condition these future maltreatment and child protection outcomes, however, the inherent trade-offs that are made at the hotline screening decision must be identified. In medical screening parlance, it is important to consider the trade-off between *sensitivity* (the proportion of patients with a disease who are correctly screened positive) and *specificity* (the proportion of patients without the disease who are correctly screened negative) in the specific and nuanced contexts of child protection.

While in the case of clinical diagnosis the ultimate outcome being screened for (i.e., disease or no disease) is clear, in the case of maltreatment allegations screened by child protection hotlines, the concept of "service need" or latent risk is poorly developed. Therefore, we need to take a more nuanced view of what a "good" initial hotline screening decision is.

An ideal system would screen out children who are at low risk of a future event and therefore have less need for early intensive services. One way of assessing lower need is to consider whether children would be re-referred if they are initially screened out. In the context of current screening practices in Allegheny County, over half the children are re-referred.

Another indicator of consistently good screen-out decisions would be that few children amongst those initially screened out would subsequently be substantiated as a victim of abuse or neglect. Unfortunately, GPS referrals (which constitute the majority of all maltreatment allegations) do not have a very meaningful definition of substantiated maltreatment and therefore this outcome was not available for modelling purposes.

Although near-fatalities and fatalities are objective and therefore useful outcomes to predict, Allegheny County is relatively small and the number of these adverse events is (thankfully) too restricted to meaningfully model. For example, in the context of Act 33 events (i.e., events where the child was killed or critically injured because of maltreatment) there were 21 children for whom a referral call was made between April 1, 2010 and February 28, 2015 who went on to have Act 33 events and this call was made more than 50 days prior to the critical incident. Only instances where the Act 33 event occurred more than 50 days following the initial referral call were included to ensure it was a new incident and not associated with the prior referral. Of these, 10 (48%) were screened out. We were able

[3] Using the absence of future involvement with protective services as a desirable goal is only correct if it comes about because addressing safety concerns at the time of the initial contact meant that there was no future need. Absence of contact could also occur for others reasons which does not mean that the child is truly safe.

CENTRE FOR SOCIAL DATA ANALYTICS

calculate a placement risk score for 18 of the referrals where a call was made more than 50 days prior to a critical incident. Of these calls, half of the referrals received a score of 15 or over.

Another proxy for an adverse event is a placement in foster care. Along the spectrum of potential interventions and services that may be offered by the child protection system, a placement falls at one extreme as it indicates that child protection workers were concerned enough about the safety of an individual child that they physically removed him or her from the home. An examination of historical data shows that among those children screened out through current practice, 6% are subsequently placed within 2 years.

Turning now to contemplating a "good screen-in," one would want to consider how many children were placed among those who were initially screened in. Of course, we might argue that if screening in is "preventive" then placement rates among those screened in should be lower than placement rates among those screened out. If we argue, however, that a substantial fraction of placements were inevitable we would like to see a high ratio of placements among those children that were screened in relative to those who were screened out.

We also argue that, all else being equal, society at large should wish to minimize the number of referrals (and therefore children) who are screened in for investigation. The reason is that screening in and a child protection investigation has some potentially deleterious effects on families. If screening in, however, is a prerequisite to being offered higher quality services or being prioritized for a slot in a desired program, one can argue the benefits of an investigation.

Since screening-in for an investigation may be both helpful and harmful to a family, it is critical to minimize the false-positive/negative rate. For instance, children and families misidentified as high risk may be subject to unnecessary involvement with social services and disruption of their home environment. Conversely, families misidentified as low risk may not receive the preventive services they need and may experience subsequent abuse and neglect (Gambrill & Shlonsky, 2000). In addition to minimizing false positives and negatives, it is critical to minimize the adverse effects of identification as at risk, such as possible stigmatization. Any risk of stigmatization is of concern to researchers and the County. For that reason, the County commissioned a full ethical report on the use of the screening tool. Two experts on the ethics of the use of screening scores, Eileen Gambrill (UC, Berkley) and Tim Dare (University of Auckland), provided ethical guidelines that guided the tool development and implementation process.

The discussion above suggests two potential candidates for outcomes to be predicted by the model:

(i) The probability that a child will be re-referred conditional on being *screened out*; and

(ii) The probability that a child will be placed in foster care conditional on being *screened in*.

CENTRE FOR SOCIAL DATA ANALYTICS

The first outcome attempts to capture the objective of screening out children who are at low risk of being re-referred in the future, thus sparing families the intrusion of an initial investigation that may not be needed. The second outcome reflects the goal of screening in children who are at high risk of being placed in foster care, the logic being that these are families where there may be a greater concentration of risk and need.

DATA

We now turn to the procedures we used to build the predictive risk model. The first step was to develop a research data set based on historical referrals for which we could observe the initial decision made at hotline screening and the eventual outcome.

To develop this model, we analysed data for all CPS and GPS referrals[4] made to Allegheny County between September 2008[5] and April 2016. In order to provide a relevant history for each referral, and follow-up time after the referral, we built the PRM using only referrals made between April 2010 and April 2014. This meant that for each referral, we could construct data on the family's history such as the number of referrals within the past 548 days. We also linked referral data to placement data – allowing us to construct a longitudinal view of the child from referral through to possible placement.

We then used this history to model a predicted likelihood of events two years into the future.

Referral and placement data were then merged with the following datasets to establish a set of predictor variables. Please note that the research team used a de-identified version of the linked data set.

County Jail: Dates of past bookings in the Allegheny County Jail.

Juvenile Probation: Dates of past involvement with the Allegheny County Juvenile Probation Office.

Public Welfare: Dates of public welfare receipt and program type (i.e., temporary aid to needy families (TANF), general assistance (GA), supplemental security income (SSI), food stamps (FS), other medical).

Behavioral Health Programs: Dates when behavioral health services were received and diagnoses made (stratified into diagnostic categories).

[4] In conducting these analyses, it was understood that Allegheny County's past CPS referral data have been subject to legally mandated expungement after a certain amount of time has passed since the referral's intake date (with expungement time varying based on the findings of the allegations and whether or not a family is currently active on a child welfare case). This meant that data regarding CPS referrals, which represent between 10-20% of Allegheny County child welfare referrals annually, were more complete for the later years in the sample.

[5] The cut-off date was determined by the fact that Allegheny County transitioned to its current KIDS data system in 2008.

CENTRE FOR SOCIAL DATA ANALYTICS

Census Neighbourhood Poverty Indicators: ZIP code data with Census information on the poverty status of each ZIP code area.

Allegheny County has additional data sets such as birth records, homeless services and educational outcomes from local school districts that were not tested in the first iteration of the model for various reasons. Birth records, for example, were not regularly being integrated into Allegheny County's data warehouse at the time the model was developed. Education data were not included since Allegheny County does not have full coverage of the county; it only partners with a subset of local school districts. The research team will consider adding additional data sets to future iterations of the model but does not expect that they will lead to significant increases in the accuracy of the model.

For each individual named in a referral (i.e., victim, other child, parent, alleged perpetrator, and other adult), we generated history variables from the child protection data and administrative datasets listed above. In total, there were more than 800 variables available for prediction and modelling purposes. These variables were constructed by the research team based on previous experience with building such risk models. In particular, to capture the dynamic nature of risk, history was divided into 90, 180, 365 and 548 day intervals. To capture the effect of the presence and intensity of predictor variables, we constructed categorical variables which reflect the presence of history with a given sector (e.g., ever in County jail) and the duration or intensity of that history (e.g., number of days in jail). Subsequently, some of these variables were aggregated or transformed (e.g., by minimums and maximums).

Since the objective of this modeling effort was to generate a risk score for each child or victim that is involved in a referral separately, records were structured as a flat file where each line of the data reflected a child or victim named in a referral. There were often multiple children named in a single referral; each child could be included in more than one referral. We do not make a distinction between whether a child is recorded in the referral as a "victim" or a "other child." This decision was made in consultation with frontline staff from the County who indicated that recording a victim in the data is somewhat arbitrary and, regardless of whether a child is labeled a victim or not, staff are required to assess all minors named in a referral.

For each observation, we constructed a history based on the date of that referral. For example, consider a referral received on July 1, 2013 and involving two children. This referral is transformed into two observations (or rows of data) in the research data. Each observation constructs the 90, 180, 365 and 548-day history as of July 1, 2013. The outcome period is then July 1, 2012 through to July 1, 2015. Note that a "re-referral" in this period is also another referral in the data set. For conducting causal inference, this might be of concern – for data mining however, it is not.

CENTRE FOR SOCIAL DATA ANALYTICS

Patterns of serial correlation in the data are not important in data mining since such correlation does not bias the estimated coefficients.[6]

METHODOLOGY FOR PLACEMENT AND RE-REFERRAL MODEL

We used non-linear regression methods for generating the final list of predictor variables and their corresponding weights. All estimation was done using Stata version 12. All data were first fully de-identified by the County. The following is a step-by-step description of the method.

1. We used the full sample of referrals (n=76,964) spanning the time period between April 2010 and April 2014 and with each observation corresponding to a unique child or victim in a referral. We estimated a probit regression model on all child-referrals with variables introduced in blocks. These blocks were

    a. Demographics of the Child Victim
    b. Child Protection History of the Child Victim
    c. Child Protection Data for all Individuals Named in the Referral
    d. Maltreatment Referral Source Information
    e. Juvenile Justice History of the Child Victim
    f. Characteristics[7] of Other Child Victims Named in the Referral
    g. Characteristics of Other Children Named in the Referral
    h. Characteristics of all Alleged Perpetrators Named in the Referral
    i. Characteristics of all Parents and Other Adults Named in the Referral
    j. Public Welfare Histories of all Child Victims
    k. Public Welfare Histories of Other Children
    l. Public Welfare Histories of all Alleged Perpetrators
    m. Behavioral Health Histories of all Individuals Named in the Referral

    We dropped all predictors that had a *t*-ratio less than 1.6.[8] We refer to the resultant set as our initial predictor variables.

[6] Serial correlation reduces the efficiency of estimates (i.e., increases their standard error) but not the bias or consistency.

[7] By "Characteristics" we mean Demographics, Welfare History, etc.

[8] Admittedly a t-ratio of 1.6 is rather arbitrary and based on judgement and experimentation with other cut-off levels.

CENTRE FOR SOCIAL DATA ANALYTICS

2. Using these initial predictor variables, we then drew with replacement a random 30% of the sample. We estimated a probit model and recorded the *t*-ratios. We repeated this process 50 times. We then kept those predictor variables with *t*-ratios greater than 2.2.[9] These variables constitute the final list of variables used in our prediction models. Of the more than 800 variables tested, there were 112 variables included in the models. The placement model has 71 weighted variables and the re-referral model has 59 weighted variables. Please see the appendix for the final list of variables. It is important to note that this is a prediction model and not a causal model. Therefore, even researchers cannot interpret the final list of variables and their corresponding weights. Variables that may independently be strong predictors of placement and re-referral may have been omitted if they were highly correlated with other variables included in the model.

3. To assess model performance, we used a randomly chosen 70% of the sample to estimate coefficient weights. Then using the *30% validation sample only,* we calculated the Area Under the Receiver Operator Curve (ROC). By using a validation sample which was separate from the sample with which the weights were established, we avoid "over-fitting" the model. We also tested these results on additional subsets of the original sample including by ethnicity (i.e., Black and White) and by referral year. Area under the ROC is used to measure overall model fit. The results are presented in the *Model Performance* section below.

4. For step 3 above, two methods were tried: ordinary probit and boosted probit.

Alternative Methods Considered

Above we described a maximum likelihood method. Alternative methods exist for constructing the algorithm – which is to use non-parametric methods such as decision-tree methods. These methods have the advantage that they are often more accurate – with higher precision, recall and area under the ROC. However, they have the weakness that they tend to be "black box" in the sense that it is more difficult to understand *why* a family received a high score. The other disadvantage of these methods is that they do not directly translate into a single score.[10] Instead, these alternative methods "flag" a referral call as "at risk" or "not at risk."

Using Weka,[11] which is an open source Data Mining software, we investigated a range of alternative methods: namely, Naïve Bayes, Ada Boost – with Random Forest, Ada Boost with J48 tree, Multilayer Perceptron, J48 Tree, Random

[9] Again, rather arbitrary but based on trial and error with higher and lower cut-off levels.

[10] Although they can be converted to a score

[11] http://www.cs.waikato.ac.nz/ml/weka/

CENTRE FOR SOCIAL DATA ANALYTICS

Tree and Random Forest. Overall, the random forest (tuned) performed the best, and below we compare its output to that of the statistical models.

Race

After an independent ethical review of this project and lengthy discussions between community stakeholders, internal staff, and members of the research team, the County made the decision that race could be included as a predictor variable if it substantively improved the predictive accuracy of the model. Although addressed more fully in the independent ethical report for this project, it should be noted that the inclusion of race in these models did not substantively improve the overall accuracy. Specifically, when we tested the model against how well it identifies Act 33 (or maltreatment fatality and near fatality) cases, we find that there is little difference in the fit between the model which includes race and the model that does not (see discussion below and Table 11).

Model Performance

We use the area under the ROC (AUR) as a general measure of model performance, and also the proportion of children who are observed with that event by the ventile of risk.

Placement Model

In health and human services, there are potentially two uses of predictive screening tools. One is to replace clinical decisions (e.g., through automatically screening in children based on their score) and the other is to augment and standardize clinical decisions (e.g., through a "risk score" or a summary statistic weighting information from the administrative data). Allegheny County was interested in developing the latter type of tool – one in which an empirically derived score could be used in conjunction with clinical judgement (and other sources of data that are not available to the PRM tool) to generate a hotline screening decision (screen in or out). In this context, the AUR is a useful statistic for the purposes of determining goodness of fit or predictive accuracy. While there are multiple interpretations of AUR, one that is helpful to us in such cases is that the AUR can be thought of as the probability that a (randomly chosen) referral that is a true positive (i.e., has a placement or re-referral within 2 years) has a higher risk score than a randomly chosen referral that is a true negative (i.e., does not have a placement or re-referral within 2 years). If the probability is 0.5, then there is no information in the risk score useful to guiding the screening decision. If the probability is 1, then it is a perfectly discriminating score.

Table 3 and Table 4 show the AUR for both the probit and boosted-probit models predicting whether a child will be placed in foster care within 730 days. We report the mean AUR and 95% confidence intervals for the validation

CENTRE FOR SOCIAL DATA ANALYTICS

sample as a whole and for sub-samples. For the overall validation sample, the AUR is 77% with race included as predictors and 76% without race.

Table 3: Area under ROC curve of Placement PRM (validation sample only, probit and boosted regressions, including race variables)

Testing Sample	Area under ROC			Area under ROC (boosted regression)			*N*
	Mean	95% Confidence Interval		Mean	95% Confidence Interval		
All screened in Referrals	**0.7653**	0.75319	0.77734	**0.773**	0.7608	0.78514	13201
Screened in Referrals during 2014	0.7594	0.71604	0.80274	0.7591	0.71456	0.80371	1091[12]
Screened in Referrals during 2013	0.7454	0.72169	0.76912	0.7474	0.72313	0.77173	3200
Screened in Referrals during 2012	0.7770	0.75283	0.80109	0.7769	0.75196	0.80187	3286
Screened in Referrals during 2011	0.7723	0.74738	0.79719	0.7912	0.7668	0.81551	2974
Screened in Referrals during 2010	0.7694	0.7407	0.79816	0.7864	0.75861	0.8141	2650
Screened in Referrals where victim is Black	0.7545	0.73713	0.77178	0.7646	0.74748	0.7817	6026
Screened in Referrals where victim is not Black	0.7686	0.75141	0.78585	0.7736	0.75585	0.7913	7175

Table 4: Area under ROC curve of Placement PRM (validation sample only, probit regressions, excluding race variables)

Testing Sample	Area under ROC			*N*
	Mean	95% Confidence Interval		
All screened in Referrals	**0.7604**	0.74838	0.77244	13031
Screened in Referrals during 2014	0.7536	0.71326	0.79396	1128
Screened in Referrals during 2013	0.7530	0.72882	0.77721	3275
Screened in Referrals during 2012	0.7859	0.76284	0.80901	3204
Screened in Referrals during 2011	0.7566	0.73170	0.78157	2952
Screened in Referrals during 2010	0.7431	0.71355	0.77268	2472
Screened in Referrals where victim is Black	0.7680	0.74908	0.78701	5983
Screened in Referrals where victim is not Black	0.8062	0.78787	0.82457	7048

[12] Note the lower referral counts in 2014 and 2010 due to partial year 2014 (Jan-Apr) and 2010 (Apr-Dec).

CENTRE FOR SOCIAL DATA ANALYTICS

Re-referral Model

Tables 5 and 6 set out the AUR for the re-referral model for all children who were screened out, and for subsamples. In this case, the model predicts re-referral during the 2-year period subsequent to being screened out. The AUR for the validation sample as a whole is 73% -74% when race is included, and 72% without race.

Table 5: Area under ROC curve of Re-referral PRM (validation sample only)

Testing Sample	Area under ROC			Area under ROC (boosted regression)			*N*
	Mean	95% Confidence Interval		Mean	95% Confidence Interval		
All screened out Referrals	**0.7314**	0.72172	0.74117	**0.7447**	0.7352	0.75429	9954
Screened Out Referrals during 2014	**0.684**	0.649	0.71899	**0.6916**	0.65665	0.72658	873
Screened Out Referrals during 2013	**0.7349**	0.71533	0.75447	**0.7429**	0.72349	0.76223	2434
Screened Out Referrals during 2012	**0.7371**	0.71775	0.75652	**0.7433**	0.72407	0.76259	2475
Screened Out Referrals during 2011	**0.7237**	0.70442	0.74303	**0.7451**	0.72647	0.76367	2601
Screened Out Referrals during 2010	**0.7553**	0.73184	0.77876	**0.7776**	0.75508	0.80021	1571
Screened Out Referrals where victim is Black	**0.6920**	0.67471	0.70926	**0.7117**	0.69486	0.72862	3557
Screened Out Referrals where victim is not Black	**0.7485**	0.73673	0.76031	**0.759**	0.74741	0.77059	6397

Table 6: Area under ROC curve of Re-referral PRM (validation sample only, probit regressions, excluding race variables)

Testing Sample	Area under ROC			*N*
	Mean	95% Confidence Interval		
All screened in Referrals	**0.7153**	0.70536	0.72521	10038
Screened in Referrals during 2014	**0.7006**	0.66567	0.73557	853
Screened in Referrals during 2013	**0.7207**	0.70103	0.74045	2509
Screened in Referrals during 2012	**0.7262**	0.70651	0.74581	2498
Screened in Referrals during 2011	**0.7085**	0.68840	0.72854	2493
Screened in Referrals during 2010	**0.7095**	0.68507	0.73389	1685
Screened in Referrals where victim is Black	**0.6719**	0.65439	0.68938	3619
Screened in Referrals where victim is not Black	**0.7339**	0.72180	0.74597	6419

AUT UNIVERSITY NEW ZEALAND

CENTRE FOR SOCIAL DATA ANALYTICS

CONCERNS OVER POLICY CHANGES IN 2015

In late 2014, there were major statutory changes to Pennsylvania's Child Protective Services Law. In particular, there were changes to the definitions of mandated reporters leading to an increase in the number of mandated reporters in Pennsylvania. Additionally, there were changes to the definitions of maltreatment. These changes led to an increase in the volume of maltreatment referrals. Recent media reports[13] have suggested that Pennsylvania's state hotline may have been understaffed to handle the increased volume and as a result there was variability in the screening quality applied to calls and the manner in which they were subsequently triaged.

Our data span this period, and we do find that the re-referral model performs less well for the 2014 referrals (for which the outcomes periods would have been in 2015 and 2016). There is, however, no evidence of similarly poor performance in the placement model. Although speculative, it may be, that for the more extreme outcome of placement in foster care, the policy changes did not have the same impact relative to referrals.

To establish whether there are any related systematic effects, we compared the maximum referral score that would have been assigned by year of the referral. In 2015, the score is lower, a finding that is statistically significant at the 95% confidence level. This suggests that referral dynamics in 2015 might have been affected by the changes in policy.

Table 7: Mean-Maximum Referral Score by year (All referrals)

Year	Mean of Maximum Referral Score of all Referrals
2010	13.2
2011	13.4
2012	13.5
2013	13.5
2014	13.3
2015	13.0
2016	13.2

Note: The year 2016 includes referrals only through April.

We also undertook a Wald test for a structural break in December 2014.

[13] See for example http://www.phillymag.com/news/2016/05/25/audit-42000-unanswered-calls-child-abuse-hotline/.

CENTRE FOR SOCIAL DATA ANALYTICS

EXTERNAL VALIDATION OF THE MODEL

External validation of the model is important to determine if the children identified as high risk for re-referral and placement are congruent to those with more generalized risk of events such as hospitalization and abuse-related fatality or near fatality. True maltreatment is very difficult to determine, and there is evidence that a lot of abuse goes unreported. Additionally, there is concern that this type of modeling is predicting children at risk of institutionalized or system response versus true underlying risk of adverse events. To address these concerns, external validations were conducted using healthcare data.

External Validation: Hospitalisation

This section was co-authored with Rachel P. Berger, MD, MPH and Srinivasan Suresh, MD, MPA, FAAP of the Children's Hospital of Pittsburgh of UPMC

To externally validate the model, we merged the County's GPS referral data with Children's Hospital of Pittsburgh of UPMC data, using a trusted third-party who was able to link the children in the two systems together using first name, last name, date of birth and social security number.

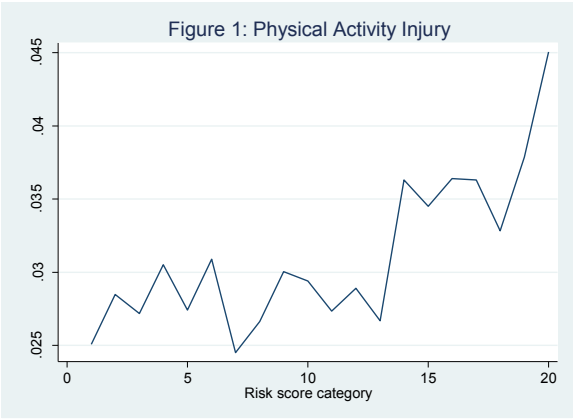
Not all children were able to be linked. Of the 64,371 children who were named in reports of alleged abuse or neglect in the period April 1, 2010 to May 4, 2016, 16,371 (25.23%) children presented at least once to the Children's Hospital of Pittsburgh of UPMC either for evaluation in the Emergency Department (ED) or for an in-patient admission from February 3, 2002 to December 31, 2015[14]. The term 'hospital event' is used in this paper to refer to both ED visits and in-patient hospital admissions.

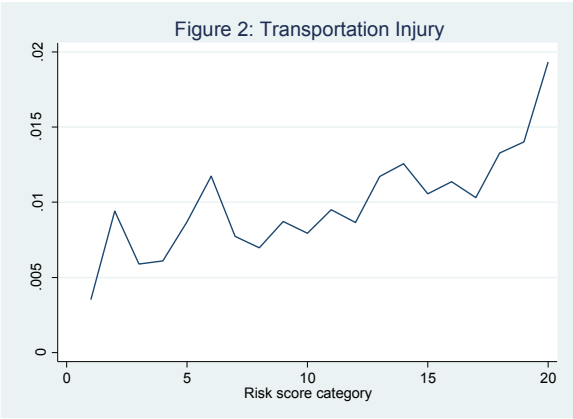
Figures 1 to 6 show hospital events for selected injuries by maximum placement risk scores for those children who were named in reports of alleged abuse or neglect. There may have been multiple referral records for a child during the study period, each having unique risk scores calculated at time of referral. We have used the maximum risk score ever received for each child in the referral data. Figures 1 to 6 demonstrate that over a broad range of injury types there is a positive correlation between the placement scores at call referral and the rate of hospital events. The ICD9 codes used to identify each type of external injury are presented in Table 11. For example, those with a placement risk score in the highest category of 20 have a hospital event rate for self-inflicted injury or suicide of 0.65% compared to 0.03% for risk score category 1. That is a child who scores a 20 at referral is 21 times more likely to be hospitalized

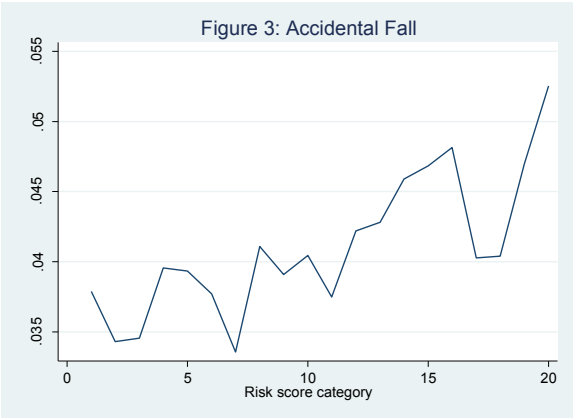
[14] Note that of the Children's Hospital of Pittsburgh of UPMC data obtained there were 33,081 records (18.83% of total) that had no recorded information on diagnosis code or admit time. We excluded these records from the analysis because we cannot analyse injury type or admit time for these records. The percentage of remaining patients that entered hospital and were discharged on the same day is 66.08%, indicating that we are not solely excluding ED visits where less information about patients may have been recorded.

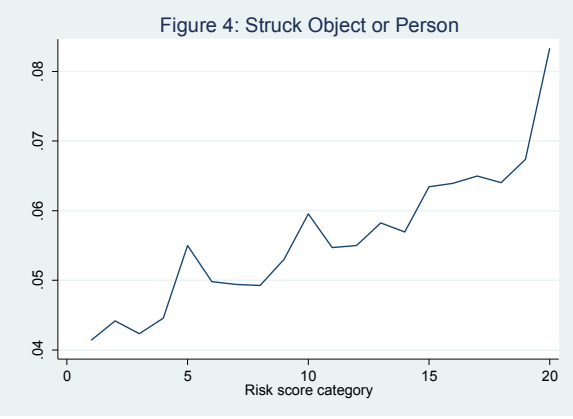
CENTRE FOR SOCIAL DATA ANALYTICS

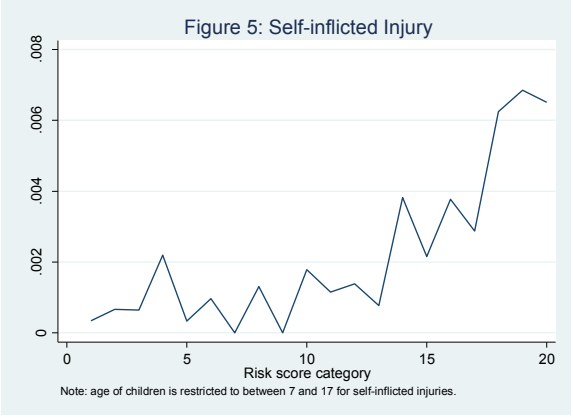
for a self-inflicted injury than a child who scores 1. The rate of hospital events from physical assault is 3.14% for category 20 compared to 0.18% for category 1. This is a factor of 17 times. The hospital event rate for accidental falls is 5.25% for category 20 compared to 3.79% of child referrals with a risk score of category 1 (or 1.4 times).

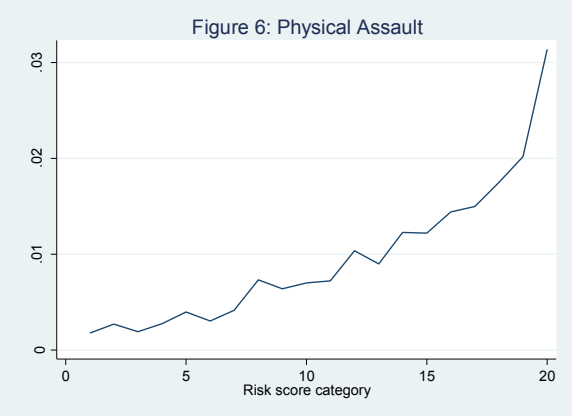












CENTRE FOR
SOCIAL DATA ANALYTICS

Figures 1 to 6: Proportion of Selected Hospital Injury Events for Children Referred to Allegheny County by Maximum Placement Scores.

We also analyzed the placement scores for children who experienced a referral to child welfare (Allegheny County Department of Human Services) within 2 years of a hospital event. Referrals that were recorded in the 30 days after the hospital event were excluded because these referrals may have been as a result of the hospital admission. To assess placement scores for children referred in the 2 years following the hospital event, we analyzed hospital event data from the period between April 01, 2010 and December 15, 2013. To assess placement scores for children referred in the 2 years prior to the hospital event, we analyzed hospital event data from the period between April 1, 2012 and December 15, 2015.

Table 8 shows the mean of the maximum placement score (for each child) in the two years prior to and the two years after the hospital event, by hospital event type. Appendix 2 contains a definition of the injury codes. Note that one admission could appear in multiple categories of hospital event type, as each admission may have multiple coded diagnoses. The highest placement risk scores are for hospital events of Abandonment or Neglect, Suicide and Self-inflicted Injuries, and Physical Assault. For Abandonment or Neglect and Suicide and Self-inflicted Injuries the average placement score in the two years previously is 17.23 and 14.54 respectively, and 18.55 and 16.98 respectively in the following two years. The risk score for Physical Assault hospital events is also among the highest observed with 14.96 for referrals in the previous two years and 15.11 for referrals in the two years following a hospital event.

AUT UNIVERSITY NEW ZEALAND

CENTRE FOR
SOCIAL DATA ANALYTICS

Table 8: **Placement Score of Admitted Children who were also referred to Child Welfare**

	Placement Score Received in 2 Years Prior to Hospital Admission				Placement Score Received 2 Years after Hospital Admission			
Type of Admission	**N**	**Mean Placement Score**	**95% Confidence Interval**		**N**	**Mean Placement Score**	**95% Confidence Interval**	
Accidental fall	1,090	11.97	11.63	12.30	1205	12.02	11.70	12.34
Injury from physical activity	1,319	12.04	11.73	12.34	1549	12.66	12.38	12.95
Accident struck by object/person	1,611	12.22	11.94	12.50	1724	12.45	12.18	12.71
Injury from medical procedure	146	12.27	11.37	13.18	171	12.60	11.84	13.35
Toxic reaction from animal or plant	258	12.51	11.86	13.16	254	12.48	11.77	13.19
Injury from transportation	333	12.53	11.93	13.14	333	12.22	11.60	12.84
Accidental poisoning non-drug/pharm	62	12.65	11.43	13.86	57	13.28	11.94	14.63
Accidental poisoning drugs/pharms	44	12.86	11.16	14.57	60	13.67	12.37	14.96
Injury from smoke/fire	9	12.89	9.06	16.72	7	14.29	8.47	20.10
Injury undetermined accident or on purpose	22	13.86	11.70	16.03	18	15.50	13.18	17.82
Self-inflicted injury	111	14.54	13.50	15.59	91	16.98	16.17	17.78
Adverse effect therapeutic drug use	74	14.82	13.75	15.90	87	12.91	11.78	14.04
Physical assault	433	14.96	14.50	15.42	461	15.11	14.67	15.55
Accident due to abandonment/neglect	13	17.23	15.67	18.79	11	18.55	17.53	19.56

Note: Maximum placement scores are calculated in the two years prior to hospital event, or two years after hospital event for all children who had a referral two years after a hospital event.

AUT UNIVERSITY NEW ZEALAND

CENTRE FOR
SOCIAL DATA ANALYTICS

External Validation: Critical Events

Thankfully, given the rarity of child death, there are too few referrals where the victim/child experienced an abuse-related fatality or near fatality to be useful for prediction purposes. However, these outcomes are useful in providing "external validity" to the model.

Overall, there were 127 referral victims who were at some point involved in an Act 33 event. These include children who were referred only *after* the fatality or near fatality event.

To test the correlation between placement risk score and Act 33, we estimated a probit model where the dependent variable $ACT33_i$ equals 1 if the child was ever involved in a fatality or near fatality and zero otherwise.

$$Pr(ACT33_i = 1|SCORE_i) = \Phi(\alpha + \beta SCORE_i) \qquad \textbf{(Model 1)}$$

We estimate the probability of observing an Act 33 event conditional on the estimated probability from the placement model given to the child $(SCORE_i)$, where $\Phi(\cdot)$ is the Normal cumulative density function. Standard errors were clustered at the child level to account for the fact that children are re-referred and their scores are not independent.

Figure 7: Stata output from Estimate of Model 1

```
Probit regression                                 Number of obs   =      99351
                                                  Wald chi2(1)    =      97.28
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -861.61167                 Pseudo R2       =     0.0272

                              (Std. Err. adjusted for 52379 clusters in MCI_ID)
------------------------------------------------------------------------------
             |               Robust
       ACT33 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       plsm2 |    1.472521   .1492935     9.86   0.000     1.179911    1.765131
       _cons |     -3.2401    .045667   -70.95   0.000    -3.329606   -3.150595
------------------------------------------------------------------------------

.
. mfx

Marginal effects after probit
      y  = Pr(ACT33) (predict)
         =  .00099274
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.      z    P>|z|  [    95% C.I.   ]       X
---------+--------------------------------------------------------------------
   plsm2 |    .0049251      .00089    5.51   0.000   .003174  .006676   .100308
------------------------------------------------------------------------------
```

Figure 7 provides the Stata output from this estimation. The estimated marginal effect is seemingly small in magnitude, but statistically different from zero at better than a 1% level. The model suggests that, on average, a ten-percentage-point increase in the probability of placement leads to an increase in the probability of an Act 33 event by

AUT UNIVERSITY NEW ZEALAND

CENTRE FOR SOCIAL DATA ANALYTICS

0.05 percentage points. This may appear to be a small effect, but this finding needs to be seen in the context of an overall mean probability of 0.1% that an Act 33 event will be observed in our data. Thus, every ten-percentage-point increase in this estimated probability of a placement is associated with a 50% increase in the probability of an Act 33 event.

Comparison to Structured Decision Making and Rule-based/Threshold Approaches

Another way of testing whether the predictions made by the model are accurate "enough" is to compare them to other existing risk scoring tools. Unfortunately, there is very limited information available concerning the performance of other prediction models in the market, such as those developed by Eckerd or SAS.

The Structured Decision Making (SDM) model, however, has been recently validated in California (Dankert and Johnson, 2014). The tool that they tested was one that was introduced in 2007 for predicting the risk that children would go on to experience recurrent maltreatment. Their validation consisted of families that were investigated between July 1, 2010, and June 30, 2011 with an 18-month follow up. In Table ES1 of that report the authors detail the results of the current risk scores and the outcomes for children following the risk scores. Note that for the Dankert and Johnson model the follow-up period was 18 months compared with the 2-year follow up period for the Allegheny County model.

Table 9: Comparison of SDM with Allegheny County Model.

	Dankert and Johnson (2014)			Allegheny County Model		
	N	%	*Removals*	*N*	%	*Placements*
Total Sample	11,444	100%	5%	23,069	100%	9%
Low	2,840	25%	2%	5,448	24%	2%
Moderate	5,130	45%	4%	10,184	44%	6%
High	2,623	23%	9%	5,720	25%	16%
Very High	851	7%	13%	1,717	7%	36%
Lift *			9			23

Note: *Lift is calculated as the ratio of the placement rate for Very High with the placement rates for Low. The Allegheny County data are based on the validation sample only. The follow up period for the Dankert and Johnson model is 18 months and for Allegheny County it is 2 years.

The results reported in Table 9 compare the SDM model applied to the California re-validation sample and reported in Table ES1 in Dankert and Johnson (2014) with the Allegheny County Model. To make the comparison appropriate,

CENTRE FOR SOCIAL DATA ANALYTICS

we generated an SDM equivalent risk score for the Allegheny County model using the Allegheny County placement model. The risk scores were generated so that the distribution of the scores would match the SDM distribution (e.g., only 7% of the sample would receive a score of Very High).

The area under the ROC for Dankert and Johnson was not provided, therefore we use the cumulative lift score calculated at the Very High level as a comparison of the goodness of fit. This ratio should be less affected by the difference in the follow-up periods between these two models. At the Very High level, the Allegheny County Model outperforms the SDM with a lift ratio (Very High to Low risk) of 23 compared to 9. That is, a Very High risk individual in SDM is 9-times more likely to be placed compared to someone in the Low risk group; whereas a Very High risk individual in the PRM model is 23-times more likely to be placed than someone in the lowest risk group.

Since SDM is built on models that use only a restricted number of predictor variables, and also rely on staff entering the values, we might have expected the SDM to perform worse. On the other hand, the SDM has available to it data that are collected for the purposes of risk assessment compared to the PRM which uses administrative data. Therefore, the difference in performance (within this small case study) provides an optimistic view of the potential for PRM to improve call screening decisions.

We also compared PRM to rule-based threshold approaches to identify "high risk" referrals. It is sometimes argued that rather than going through the process of embedding a predictive risk model, we might be able to identify "high risk" referrals simply by employing a series of rules. These are sometimes called "thresholds models" because they assess a call on the basis of a fixed set of thresholds or hurdles. Once referral meets the set of hurdles, it is classified as high risk.

The advantage of such an approach is that it does not need the building of a predictive risk model and is easily applied by frontline caseworkers and screening staff. The disadvantages are that threshold models do not offer a risk score – but rather a single group. The size of this group would vary depending on the nature of the threshold. Table 10 compares the "accuracy" of the threshold approach with a similar proportion of referrals chosen using PRM.

Consider a threshold model which considers all referrals where a child or adult on the referral has had at least 2 referrals in the previous 365 days. Such a threshold model would identify 21% of the sample as "high risk". We find that this criterion identifies referrals where only 15% of the children are placed within the 2 years following the referrals. However, if we identify the same proportion of high risk referrals using the predictive risk model (the top 21% of calculated risk scores from the Allegheny Screening Tool), we find that 27% of these referrals are placed within 2 years.

CENTRE FOR SOCIAL DATA ANALYTICS

Similarly, other criteria we could use based on the source of referrals (mandated vs. non mandated), age of child and combinations can provide smaller sub-groups to identify as high risk. However, in each of these instances choosing a similar size group using a predictive risk score provides a group of referrals with higher baseline risk of placement in the subsequent 2 years.

Table 10: Threshold Model vs. PRM for identifying "high risk" referral

Criteria for Classifying as "High Risk" on a Threshold Model	**Share of Referrals Meeting Threshold**	**Placement Rates in following 2 years for referrals meeting threshold**	**Placement Rates if the same number of referrals are identified by a Predictive Risk Model**
Referral from a mandated referrer (school, medical, court or police)	42%	12%	20%
At least 2 referrals in past 365 days involving any adult or child on the referral	21%	15%	27%
At least 2 referrals in past 365 days **and** a mandatory referring source	15%	14%	30%
Victim or Child age<7 **and** at least 1 referral in past 365 days for any person on the referral	13%	14%	31%

Implementation of the Risk Score

After considerable discussion, the research team and Allegheny County decided that results from this initial modeling effort were promising enough to progress to the implementation stage.

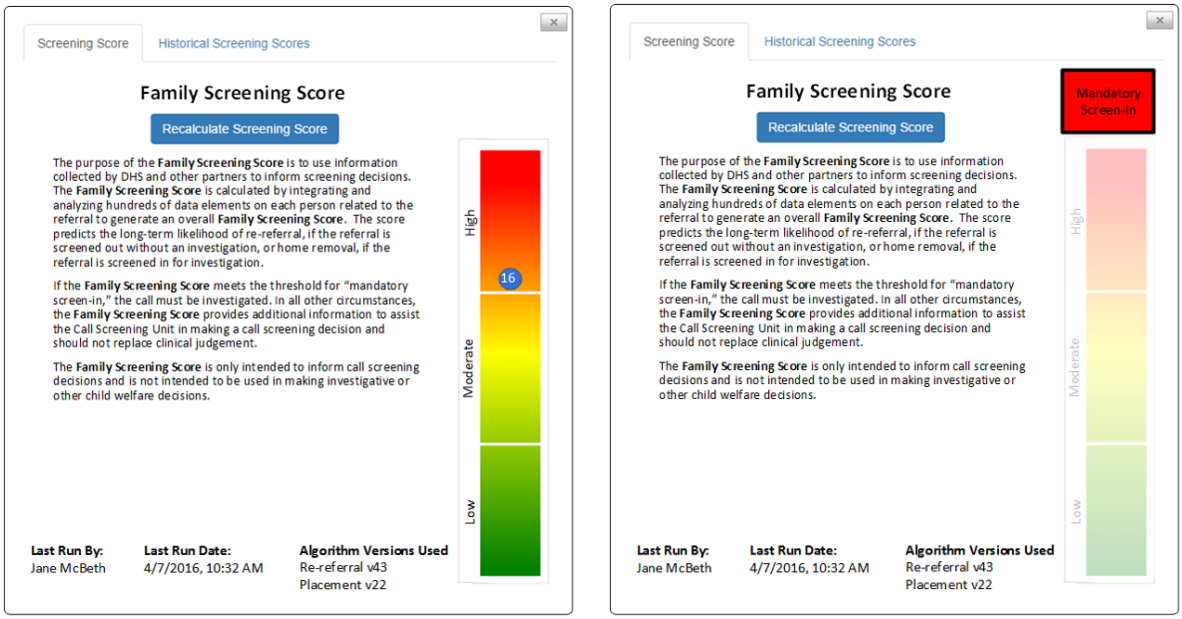
Of considerable debate and discussion were questions surrounding how to present the risk scores to hotline screening staff – and whether workers assigned to investigate a referral should also have access to the score. It was decided that a ventile score would be calculated for each child based on both the placement and re-referral models; that is, a score from 1 to 20 indicating the ventile into which the child's risk score falls. For example, a placement risk score of 20 means that the child is in the top 5% of risk scores from the placement model. The same child might have a re-referral score of 15. It was decided that based on the maximum of the placement risk score, the County would then determine a threshold above which referrals would be required to be screened in. For this group, the call screeners would be required to accept them for an in-person investigation. The model includes functionality that allows call screening supervisors to override this requirement at their discretion; all overrides are documented and reviewed. For the

CENTRE FOR
SOCIAL DATA ANALYTICS

referrals that are not required to be screened in, the referral would be classified into one of three categories (high, medium, and low). This classification would be based on the maximum of the score of any child for either the referral or placement model.

Figure 8 provides screen shots of the model as presented to the call screener. Call screeners are presented with a classification (mandatory screen-in, high, medium or low) and a score based on the maximum score for that referral. This score is the maximum across re-referral and placement score across all children in the referral. Note that there is a different screen presented to the call screener when the referrals is a "mandatory-screen in." The call-screener will be shown an additional alert that says "Mandatory-screen in."

Figure 8: Screen Shots of the Family Risk Score



Mandatory Screen-In

The threshold for the mandatory-screen in was determined solely by placement score and designed to capture as many of the Act 33 children as possible. The high, medium and low categories are based on the maximum of the referrals and placement scores. "

Table 11 outlines the sensitivity of the risk classes with respect to the Act 33 referrals.

The Act 33 referrals were used in this sensitivity analysis because they were the greatest priority for leadership within the County. In this case, we find that 49% of Act 33 events would have been automatically screened-in for investigation. Recall that in the context of Act 33, we include children who might have had an Act 33 in the past or

AUT UNIVERSITY NEW ZEALAND

CENTRE FOR SOCIAL DATA ANALYTICS

concurrently with the referral. The reason we are using Act 33 is that they are good proxies for high risk families – not because these particular Act 33 events would have been preventable in any way. In our Act 33 sample, there were only 18 referrals where the critical incident occurred more than 50 days after the referrals and could therefore have been considered to be in any way "preventable."

Table 11: Screening Score Groups and Act 33

Risk Class	N (No Race Model)	Share
Low	7	0.60
Medium	19	0.15
High	37	0.30
Mandatory Screen-In	60	0.49
Total	123	1.00

Table 12: Screening Score Groups and Outcomes (all sample of referrals, no race model).

	Share of referrals	Placed in 365 days	Placed in 730 days
Low	0.20	0.009	0.018
Med	0.28	0.027	0.044
High	0.27	0.057	0.089
Auto	0.24	0.167	0.223
Total	1.00	0.067	0.097
Ratio		18.26	12.28

	Referred in 365 days	Referred in 730 days	Service Open in 730 days	Currently Screened In	Black Race
Low	0.212	0.297	0.043	0.24	0.193
Med	0.300	0.418	0.090	0.36	0.327
High	0.403	0.548	0.138	0.49	0.410
Auto	0.329	0.468	0.157	0.75	0.514
Total	0.320	0.444	0.111	0.47	0.371
Ratio	1.56	1.58	3.68	2.99	2.66

Table 12 shows a range of outcomes for each of the risk groups and the ratio between those who are classified as auto-screened and those who are classified as low risk. Of all referrals, 24% are classified as auto-screen in, 27% are high risk, 28% are medium and 20% are low risk. Those who are auto-screened in are 18 times more likely to be placed in 1 year and 12 times more likely to be placed in 2 years compared to those classified as low risk. However, 25% of

CENTRE FOR SOCIAL DATA ANALYTICS

those who are in the auto-screen-in category are currently screened out whereas 24% who are in the low risk category are screened in.

Impact of Race as a Predictor

In Tables 11 and 12 we presented the model which does not use any race factors as part of the predictive model. With respect to sensitivity to Act 33 referrals (i.e., the results presented in Table 12), the model which includes race as predictor is identical. It too captures 49% of Act 33 referrals in the auto-screen in group and a similar proportion in the other groups. Table 13 presents the rate-ratios with respect to the other outcomes. As expected, the model performs slightly better (for example, the rate ratio of being placed in 730 days is 14.05 with race included in the model compared with 12.28 when race is excluded). On the other hand, with race included in the model, Black children are 3.76 times as likely to be classified as Auto-screen In vs. Low; when race is excluded from the model, this rate decreases to 2.66.

Table 13: Screening Score Groups and Outcomes (all sample of referrals, With race model).

	Placed in 730 days	Referred in 730	Service Open in 730	Currently Screened In	Race Black
Low	0.016	0.201	0.282	0.24	0.150
Medium	0.046	0.310	0.432	0.38	0.334
High	0.088	0.407	0.552	0.49	0.401
Auto	0.226	0.333	0.474	0.74	0.563
Total	0.097	0.320	0.444	0.47	0.371
Ratio of Low to Auto-Screen In	14.05	1.66	1.68	3.86	3.76

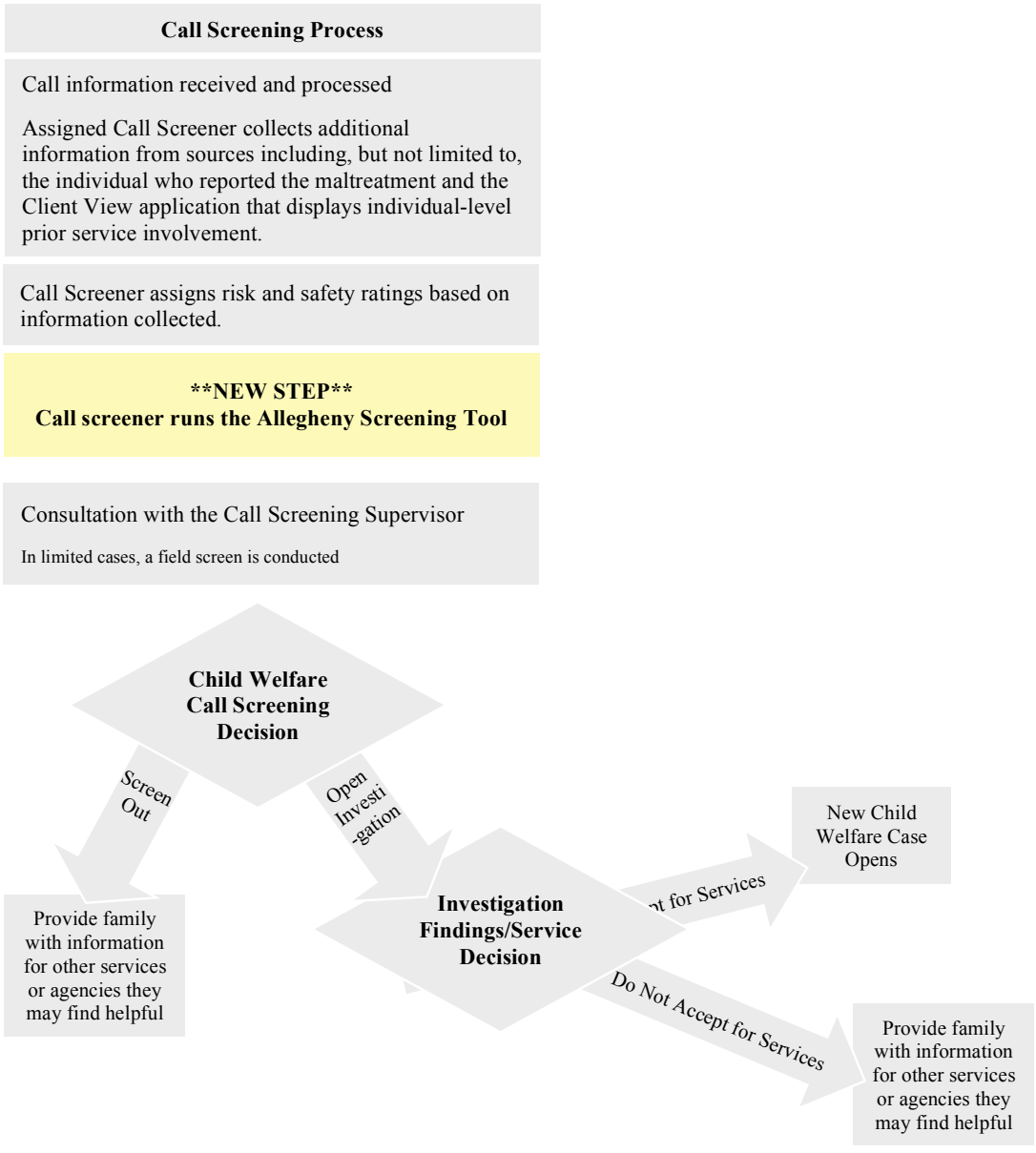
The question of which model to choose depends on the trade-off between any concerns of racial bias in the use of the model, and loss of precision with regard to these outcomes. Overall, given that both models are equally sensitive with regard to Act 33 outcomes, we would recommend that race not be included in the model. Of course, it is important to note that not including race is not to imply that race does not feature into the model because there are other predictors that are highly correlated with race due to potentially institutionalized racial bias (e.g., criminal justice history) that would imply that race is still a factor. It is for this reason that continuing monitoring of the application of the model with regard to racial disparities should be undertaken.

CENTRE FOR
SOCIAL DATA ANALYTICS

Using the Model in Practice

The intent of the model is to inform and improve the decisions made by the child protection staff. As stated in the background, it was never intended that the algorithm would replace human decision-making. To implement the model, a supplemental step in the call screening process was added to generate re-referral and placement risk scores that the call screener and call screening supervisor review when deciding if the referral should be investigated. Beyond this point, the risk scores do not impact the referral progression process.

Figure 9: Referral Progression Process

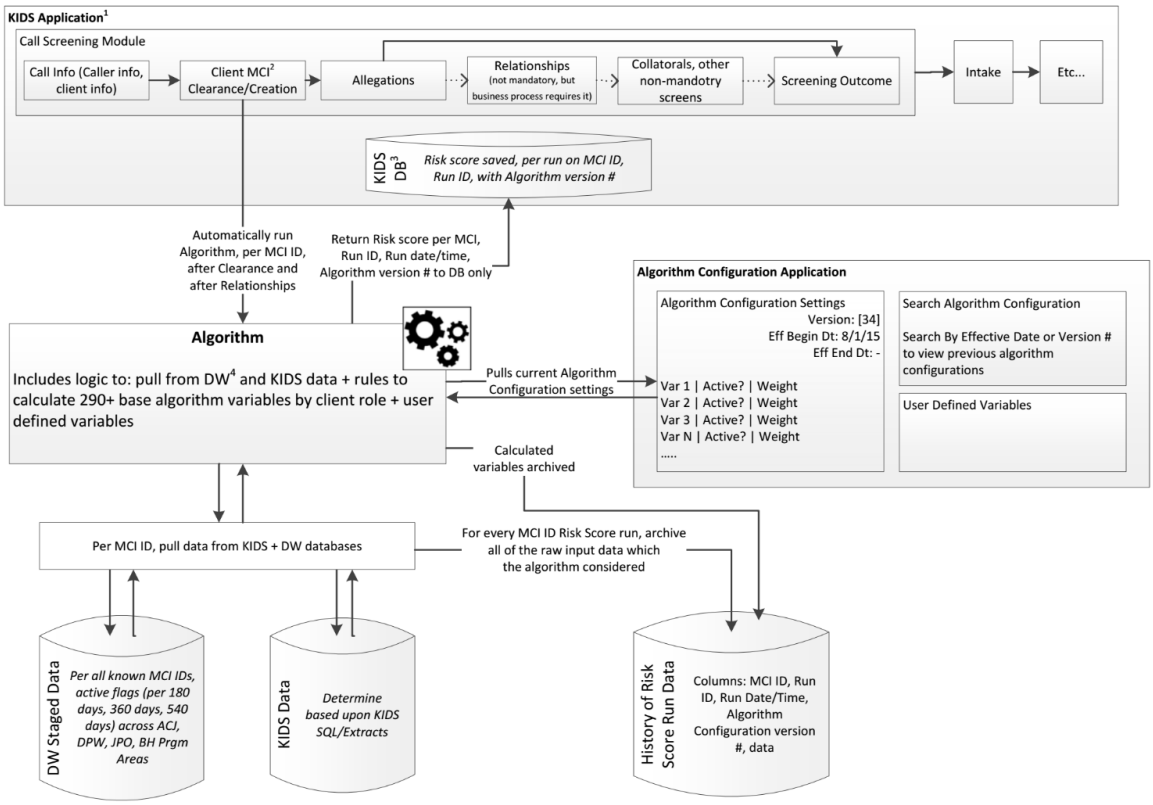


CENTRE FOR
SOCIAL DATA ANALYTICS

Technical Implementation

The front-end of the model was built directly into Allegheny County's child welfare case management system (KIDS). The algorithm is run for every child listed on the referral and includes data on all individuals listed on the referral (child victim, siblings, biological parents, alleged perpetrator, etc.). The algorithm pulls data from KIDS as well as Allegheny County's data warehouse to generate over 800 variables that are each matched with the applicable weight that is stored in the Algorithm Configuration Application. All 800+ variables that were tested in the models are included in the implementation even though only 112 variables have non-zero weights in the current model. The Algorithm Configuration Application was designed to be flexible and transparent. Variables and weights can easily be updated as the model changes. Additionally, records of all versions of the algorithm, as well as a history for every instance the algorithm is run (including the 800+ variables per individual) is maintained to support the team's quality assurance, evaluation and maintenance efforts.

Figure 10: Technical Implementation of the Screening Tool (source: Allegheny County)



Notes to figure: (1) KIDS application is the electronic child welfare case management system in Allegheny County, (2) MCI is the master client index, the unique identifier assigned to clients in Allegheny County's data warehouse, (3) DB refers to the KIDS database, (4) DW refers to Allegheny County's data warehouse

AUT UNIVERSITY NEW ZEALAND

CENTRE FOR
SOCIAL DATA ANALYTICS

Training

A three-hour training was provided to all full-time and occasional call screening staff, intake administrators and key child welfare administrators prior to implementation. The training provided a brief overview of PRM and the application of it within Allegheny County to give participants an understanding of what risk modeling is, how the model was built, and the predictive power of the model. The training also outlined the changes that were made to the child welfare electronic case management system in conjunction with the tool and what different fields or buttons would be available for workers with the implementation of this model.

Much of the training was dedicated to building worker understanding of the policy and practice for using the tool. These discussions were framed using the ethical analysis completed in advance of implementation, with specific emphasis on confirmation bias, stigmatization, and high confidence in the accuracy of scores. Some of the key points emphasized through these discussions included:

- Scores are only available to call screening staff and are not to be shared when discussing referrals with workers who may receive the referral in investigation
- The screening tool is to be used as one of the tools available to screeners when making their recommendations and supervisors when making their decisions
- The tool does not mandate the response the agency will have to any referral (low scores can still be screened in for investigation and high scores can be screened out)
- The scores do not reflect anything about the current allegations of the referral, but rather help to aggregate historical information on the family and what that information means for future risk
- The scores do not reflect anything about whether the allegations presented meet the threshold for case opening, case substantiation or need for involvement of other systems, such as law enforcement or mental health

Discussions of these key points were framed through the use of scenarios. Trainers used de-identified referral information to show screening staff information about a family and to discuss the decision that would be made. Trainers then shared the screening score based on historic modeling and discussed how this may or may not impact the screeners decision.

CENTRE FOR SOCIAL DATA ANALYTICS

NEXT STEPS: SIX MONTH REBUILD AND ADDING A RANDOM FOREST MODEL

In January 2017, we extracted updated data to rebuild the logistic model to test if more updated data might better fit more recent events. We also explored whether additional methods such as Support Vector Machines or Random Forest might offer a more accurate way of flagging those who should be flagged as being "mandated".

For model building, and to be able to predict re-referral and placement within 2 years, we used data spanning the period April 2010 to July 2014. We used 46,503 screened-in child-referrals for placements and 36,585 screened-out referrals for re-referrals, in this period.

We compared the results from the newly weighted regression model that uses more up-to-date data and what scores would have resulted for the existing model. We see no improvement in terms of AUR for the placement nor the re-referral models, so our intention is to continue using the existing weights for the logistic regression of both models.

We also experimented with Support Vector Machine but despite multiple experiments - found little additional predictive power.

However, we have found that a Random Forest with all (approximately 730) variables, has an AUR of **88.1%** for placement and **87.2%** for re-referral. This compares to **77%** and **73%** using logistic regression, respectively.

To understand what this means, rceall that we flag the top 25% as riskiest of placement as "mandatory screen-ins". Using the logistic model, this would have flagged 58% of those who end up being placed within 2 years (i.e: true-positive rate = 0.58). With the random-forest model, we end up flagging 77% of those who are ultimately placed. This represents an improvement of almost 1/3rd with respect to the number of actually placed children that we can identify as "high-risk". We should be aware that the two models do not necessarily flag the same child-referrals (i.e. the 58% is not necessarily fully included into the 77%); we are exploring the characteristics of the predicted population that make a difference between the two models.

It clear that the main advantage of the random forest model is in its ability to capture more of those who end up being placed.

Table 14 shows the correlation between those that were placed and flagged by each of the models as being in the top 25%. Of those who were placed, 54% would have been flagged by both the logistic and random forest. 17% would have been missed by both. However, 24% would have been flagged by the random forest and not the logistic; whereas only 5% would have been flagged by the logistic and not the random forest model.

CENTRE FOR
SOCIAL DATA ANALYTICS

Table 14: Comparison of those who were placed and flagged as mandatory screen-in risk group

	Logistic Flagged	Logistic Not Flagged
Random Forests Flagged	0.53685259	**0.23804781**
Random Forests Not Flagged	0.05179283	0.17330677

This suggests that there is real value in providing the random forest flag in addition to the logistic regression risk score. *Between them, they capture 83% of all those who will end up being placed.*

Despite its advantages, the main challenge with a random forests model using ~730 variables is that it is not transparent for the final users. Though we could draw some conclusions by exploring the importance of each variable for the model, we cannot clearly explain why one person received a higher score than another, because of the complexity of the model representation. Of course, this is not to say that the logistic model is easily interpreted given the number of factors and the high degree of correlation. Nonetheless, the methodology of regressions is more familiar to child welfare workers who have been using actuarial models for some time (albeit not Allegheny County).

Given these results, what we recommend to do is to add a random forest generated flag for the 25% most risky because it provides a higher prediction ability while a logistic regression can provide more explanation in terms of scores that are usable in the front-line.

Conclusion

Overall, a probit model with no race variables was initially implemented. Subsequent exploration in the 6-monthly rebuild suggests that an addition of a Random Forest Model could boost accuracy.

The approach that Allegheny and the research team have taken to the implementation of the Family Screening Score is to see it as a three way evolution between practice, policy and modelling. Because practice and policy is evolving, the best way to build and implement the model will also change. At some point, we would expect this process to settle into a more stable equilibrium.

However, readers should be warned that this report is very much a snapshot of the status of the project as at the date at which it was published.

CENTRE FOR SOCIAL DATA ANALYTICS

There are two independent evaluations of the screening tool in progress. The process evaluation is being conducted by Hornby Zeller Associates, Inc. and will assess how the screening tool is being implemented. The impact evaluation is being conducted by Stanford University and will focus on the accuracy of decisions, reduction in unwarranted variation in decision-making, reduction in disparities and overall referral rates and workload.

We would urge readers to contact Allegheny County or the Research team to learn about the most recent updates.

CENTRE FOR
SOCIAL DATA ANALYTICS

APPENDIX: VARIABLES USED IN THE ALLEGHENY CHILD WELFARE PREDICTIVE RISK MODEL

The weights of the model are available upon request from the Allegheny County Department of Human Services.

Definition of suffixes:

vict_othr	All other victims involved in this referral (other than the victim being risked scored for)
vict_self	The victim being risk scored for
prnt	The parent/guardian
perp	The alleged perpetrator
chld	Other children involved in the referral who are not identified as a victim

Placement Model

Variable	**Description**
adt_vic_null	If the victim is 18 years old or over at the time of the current referral
BH_c_20	Aggregate count of behavioural health events related to neurotic disorders for all individuals in this referral
BH_Substance	Aggregate count of behavioural health events related to inhalants, amphetamines, substance induced disorders, hyp/sed, PCP, cocaine, polysubstance disorder, cannabis, ethanol, and/or opioids for all individuals in this referral
chld_age_pre_null	The number of other children involved in this referral who are $3 \leq \text{age} < 6$
chld_age_sc1_null	The number of other children involved in this referral who are $6 \leq \text{age} < 9$
chld_age_sc2_null	The number of other children involved in this referral who are $9 \leq \text{age} < 13$
chld_age_teen_null	The number of other children involved in this referral who are $13 \leq \text{age} < 18$
PaDHS_fs_1_per_chld	Supplemental Nutrition Assistance Program - % of time seen in Pa DHS last year
PaDHS_fs_2_per_chld	Supplemental Nutrition Assistance Program - % of time seen in Pa DHS last 2 yrs.
PaDHS_fs_2_per_vict_othr	Supplemental Nutrition Assistance Program - % of time seen in Pa DHS last 2 yrs.
PaDHS_fs_3_per_chld	Supplemental Nutrition Assistance Program - % of time seen in Pa DHS last 3 yrs.

AUT UNIVERSITY NEW ZEALAND

CENTRE FOR SOCIAL DATA ANALYTICS

Variable	Description
PaDHS_fs_3_per_vict_othr	Supplemental Nutrition Assistance Program - % of time seen in Pa DHS last 3 yrs.
PaDHS_fs_everin_chld	Supplemental Nutrition Assistance Program - If ever in Pa DHS before
PaDHS_ssi_1_per_perp	Supplemental Security Income - % of time seen in PADHS last year
PaDHS_ssi_now_chld	Supplemental Security Income - if in PADHS at time of referral
PaDHS_ssi_now_oth	Supplemental Security Income - if in PADHS at time of referral
PaDHS_ssi_now_perp	Supplemental Security Income - if in PADHS at time of referral
PaDHS_tanf_1_per_prnt	Temporary Assistance for Needy Families - % of time seen in PADHS last year
PaDHS_tanf_2_per_vict_othr	Temporary Assistance for Needy Families - % of time seen in PADHS in the last 2 years
PaDHS_tanf_3_per_vict_othr	Temporary Assistance for Needy Families - % of time seen in PADHS in the last 3 years
PaDHS_tanf_everin_prnt	Temporary Assistance for Needy Families - if was ever in PADHS before
PaDHS_tanf_now_oth	Temporary Assistance for Needy Families – if was in PADHS at time of referral
PaDHS_tanf_now_prnt	Temporary Assistance for Needy Families – if was in PADHS at time of referral
fndg_past548_count_vict_self	Aggregate number of referral calls with validated findings in past
jpo_1_per_chld	Juvenile Probation Office - % of time seen in JPO in the last year
jpo_2_per_chld	Juvenile Probation Office - % of time seen in JPO last 2 years
jpo_everin_perp	Juvenile Probation Office - If the perpetrator was in JPO before
jpo_everin_vict_self	Juvenile Probation Office - If the victim was in JPO before
jpo_now_vict_self	Juvenile Probation Office - If the victim was in JPO at time of current referral
perp_0_null	If no perpetrator in referral

CENTRE FOR SOCIAL DATA ANALYTICS

Variable	Description
perp_2_null	If 2 perpetrators in referral
perp_age_5564_null	Count of the number of perpetrators that are 55 ≤ age < 65
perp_age_65_null	Count of the number of perpetrators that are over age 65
perp_females_null	Count of the number of perpetrators that were female
plsm_past180_dummy_null	If the victim was in placement in the last 180 days
plsm_past548_count_null	Aggregate count of placement associated with a unique ID in the last 548 days
poverty_30over_null	If poverty rate is greater than 30
poverty_under30_null	If poverty rate is greater than 20 but less than 30
presc_vic_null	If victim is 3 ≤ age < 6
prnt_0_null	If there is no person listed as 'Parent' in the 'Primary Referral Role'
prnt_2_null	If there are 2 people listed as 'Parent' in the 'Primary Referral Role'
prnt_age_2024_null	Count of number of parents in 20 - 24 age group
prnt_age_2534_null	Count of number of parents in 25 - 34 age group
prnt_age_3544_null	Count of number of parents in 35 - 44 age group
prnt_age_4554_null	Count of number of parents in 45 - 54 age group
prnt_age_65_null	Count of number of parents over 65
prnt_over2_null	If there are more than 2 people listed as 'Parent' in the 'Primary Referral Role'
ref_anon_null	If unknown referral source
ref_past365_count_vict_self	Aggregate count of referrals associated with a unique ID which happened within the last 365 days of the current referral
Ref_past548_serv	Aggregate counts of referrals accepted for service in the last 18 months across all individuals involved in the referral, except the victim being risk scored, whose history was accounted for separately by other variables
ref_past90_count_vict_self	Aggregate count of referrals associated with a unique ID which happened within the last 90 days of the current referral
ref_polc_null	If Law Enforcement Referral Source
ref_relt_null	If Relative Referral Source

CENTRE FOR SOCIAL DATA ANALYTICS

Variable	Description
sc1_vic_null	If victim is 6 ≤ age < 9
sc2_vic_null	If victim is 9 ≤ age < 13
tod_vic_null	If victim is 1 ≤ age < 3
vic_2_null	If exactly 2 victims in referral
vic_3_null	If exactly 3 victims in referral
vic_4_null	If exactly 4 victims in referral
vic_5_null	If exactly 5 victims in referral
vic_6_null	If exactly 6 victims in referral
vic_age_adt_null	Number of adult victims in the referral
vic_age_inf_null	Number of infant victims in the referral
vic_age_pre_null	Number of preschool victims in the referral
vic_age_sc1_null	Number of school-aged victims in the referral (6 ≤ age < 9)
vic_age_teen_null	Number of teenaged victims in the referral
vic_age_tod_null	Number of toddler victims in the referral
vic_over6_null	If more than 6 victims in referral

Re-referral model

Variable	Description
chld_2_null	If there are 2 children involved in the referral who are not identified as victims of the referral
BH_c_12	Aggregate count of behavioural health events related to depressive disorder for all individuals in this referral
BH_Substance	Aggregate count of behavioural health events related to inhalants, amphetamines, substance induced disorders, hyp/sed, PCP, cocaine, polysubstance disorder, cannabis, ethanol, and/or Opioids for all individuals in this referral
chld_3_null	If there are 3 children involved in the referral who are not identified as victims of the referral
chld_4_null	If there are 4 children involved in the referral who are not identified as victims of the referral
chld_5_null	If there are 5 children involved in the referral who are not identified as victims of the referral
chld_over5_null	If there are more than 5 children involved in the referral who are not identified as victims of the referral

CENTRE FOR SOCIAL DATA ANALYTICS

Variable	Description
PaDHS_fs_2_per_prnt	Supplemental Nutrition Assistance Program - % of time seen in PADHS in the last 2 years
PaDHS_fs_now_perp	Supplemental Nutrition Assistance Program - if in PADHS at time of referral
PaDHS_om_1_per_chld	Other medical assistance - % of time on other medical assistance in last year
PaDHS_om_1_per_prnt	Other medical assistance - % of time on other medical assistance in last year
PaDHS_om_1_per_vict_othr	Other medical assistance - % of time on other medical assistance in last year
PaDHS_om_2_per_chld	Other medical assistance - % of time on other medical assistance in last 2 years
PaDHS_om_2_per_prnt	Other medical assistance - % of time on other medical assistance in last 2 years
PaDHS_om_2_per_vict_othr	Other medical assistance - % of time on other medical assistance in last 2 years
PaDHS_om_2_per_vict_self	Other medical assistance - % of time on other medical assistance in last 2 years
PaDHS_om_3_per_prnt	Other medical assistance - % of time on other medical assistance in last 3 years
PaDHS_om_3_per_vict_othr	Other medical assistance - % of time on other medical assistance in last 3 years
PaDHS_ssi_2_per_chld	Supplementary Security Income - % of time seen in PADHS in last 2 years
PaDHS_ssi_3_per_chld	Supplementary Security Income - % of time seen in PADHS in last 3 years
PaDHS_ssi_everin_oth	Supplementary Security Income - if ever in received SSI
PaDHS_tanf_3_per_vict_othr	Temporary Assistance for Needy Families - % of time seen in PADHS last 3 yrs.
jpo_1_per_chld	Juvenile Probation Office - % of time seen in JPO last year
jpo_2_per_prnt	Juvenile Probation Office - % of time seen in JPO in the last 2 years
jpo_3_per_chld	Juvenile Probation Office - % of time seen in JPO in the last 3 years
jpo_3_per_perp	Juvenile Probation Office - % of time seen in JPO in the last 3 years

CENTRE FOR SOCIAL DATA ANALYTICS

Variable	Description
jpo_everin_chld	Juvenile Probation Office - If the other child was in JPO before
jpo_everin_perp	Juvenile Probation Office - If the alleged perpetrator was in JPO before
jpo_everin_vict_othr	Juvenile Probation Office - If the other victim was in JPO before
jpo_now_chld	Juvenile Probation Office - If the other child was in JPO at time of current referral
perp_2_null	If there are 2 perpetrators in referral
perp_age_12_null	The number of perpetrators that are younger than age 13
perp_age_2534_null	The number of perpetrators that are between age 25 and 34
perp_females_null	The number of perpetrators that are female
plsm_past548_dummy_null	If the victim was in placement in the last 548 days
presc_vic_null	If victim is $3 \leq$ age < 6
prnt_0_null	If there is no person listed as 'Parent' in the 'Primary Referral Role'
prnt_2_null	If there are 2 people listed as 'Parent' in the 'Primary Referral Role'
prnt_age_5564_null	The number of parents aged 55-64
prnt_age_65_null	The number of parents aged 65 or over
prnt_over2_null	If there are 2 people identified as parents
ref_Unknown_count	Aggregate counts of "Unknown" race in this referral across all victims, children, perpetrators and parents
ref_anon_null	Anonymous/unknown referral source
ref_med_null	Medical Referral Source
ref_other_state_null	If it is an out of state address
ref_past365_count_perp	Aggregate count of referrals associated with a unique ID which happened within the last 365 days of the current referral - perpetrator
ref_past365_count_prnt	Aggregate count of referrals associated with a unique ID which happened within the last 365 days of the current referral – parent

CENTRE FOR SOCIAL DATA ANALYTICS

Variable	Description
ref_past548_count_prnt	Aggregate count of referrals associated with a unique ID which happened within the last 548 days of the current referral - parent
ref_past548_count_vict_self	Aggregate count of referrals associated with a unique ID which happened within the last 548 days of the current referral - victim
ref_prnt_null	Parental referral source
ref_relt_null	Relative referral source
adt_vic_null	If the victim is 18 years old or over at the time of the current referral
ref_schl_null	School referral source
sc1_vic_null	If the victim is $6 \leq age < 9$
sc2_vic_null	If the victim is $9 \leq age < 13$
ser_past548_count_vict_self	Aggregate count of open-for service-referrals associated with a unique ID which happened within the last 548 days of the current referral
tod_vic_null	If victim is $1 \leq age < 3$
vic_age_sc1_null	Number of school-aged victims in each referral (aged 6-8)

CENTRE FOR SOCIAL DATA ANALYTICS

APPENDIX: HOSPITAL INJURY CLASSIFICATIONS

Hospital event Injury Type and ICD9 Codes

Injury type	ICD9 Codes
Injury from physical activity	E0000-E030; E927-E9282
Injury from transportation	E8000-E848; E9290-E9291
Accidental poisoning drugs/pharms	E8500-E8699; E9292
Injury from medical procedure	E8700-E8799
Accidental fall	E8800-E8889; E9293
Injury from smoke/fire	E8900-E899
Accident climatic or natural disaster	E9000-E903; E9294-E9295
Accident due to abandonment/neglect	E9040-E9049
Toxic reaction from animal or plant	E9050-E9069
Accident climatic or natural disaster	E907-E9099
Accidental drowning	E9100-E9109
Accidental obstruction respiratory	E911-E9139
Accident struck by object/person	E914-E9269; E9283-E9289; E9298-E9299
Adverse effect therapeutic drug use	E9300-E9499
Self-inflicted injury	E9500-E959
Physical assault	E9600-E978
Injury on accident or purpose	E9800-E989

CENTRE FOR SOCIAL DATA ANALYTICS

REFERENCES

Billings, J., Blunt, I., Steventon, A., Georghiou, T., Lewis, G., & Bardsley, M. (2012). Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30). BMJ open, 2(4), e001667.

Billings, J., Dixon, J., Mijanovich, T., & Wennberg, D. (2006). Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. BMJ, 333(7563), 327.

Dankert, Erin Wicke and Kristen Johnson (2014) *Risk Assessment Validation: A Prospective Study*. California Department of Social Services. Children and Family Services Division.

Gambrill, E., & Shlonsky, A. (2000). Risk Assessment in Context. *Children and Youth Services Review,* 22(11), 813-37.

Ministry of Social Development (2014) "The feasibility of using predictive risk modelling to identify new-born children who are high priority for preventive services – companion technical report. 4th February 2014, *Ministry of Social Development*. Wellington, New Zealand.

Vaithianathan, R., Maloney, T., Putnam-Hornstein, E., & Jiang, N. (2013). Children in the public benefit system at risk of maltreatment: Identification via predictive modeling. American journal of preventive medicine, 45(3), 354-359.

Panattoni, L. E., Vaithianathan, R., Ashton, T., & Lewis, G. H. (2011). Predictive risk modelling in health: options for New Zealand and Australia. Australian Health Review, 35(1), 45-51.

Wilson, M. L., Tumen, S., Ota, R., & Simmers, A. G. (2015). Predictive Modeling: Potential Application in Prevention Services. American journal of preventive medicine, 48(5), 509-519.

AUT UNIVERSITY NEW ZEALAND

SECTION 2

Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County

by Tim Dare and Eileen Gambrill

INTRODUCTION

This report comments on two linked papers produced by Rhema Vaithianathan, Nan Jiang, Tim Maloney and Emily Putnam-Hornstein as part of the development of a predictive risk modeling tool to improve child protection decisions being made by the Allegheny County Department of Human Services (DHS) (Vaithianathan, et al., 6 Feb, 2016, and Vaithianathan, et al., 23 March, 2016). The details of the predictive risk model are presented in those papers and we do not here attempt to repeat that presentation. We assume those reading this ethical assessment will be familiar with the papers.

Since our assessment depends on the accuracy of our understanding of the tool, however, we begin with a brief summary so that it will be clear what we are taking them to have proposed.

SUMMARY OF THE PROPOSED ALLEGHENY FAMILY SCREENING TOOL

In short, in 2014, DHS sought partners to work with them on using their integrated data systems to make better child protection decisions. The consortium of researchers led by Vaithianathan was awarded the contract and commenced work on building a predictive risk modelling tool. Following discussion and preliminary work, it was decided to develop a tool that would provide a risk assessment when a call about an allegation of maltreatment was received by the DHS call center, rather than at the birth of a child.

The Allegheny Family Screening Tool (AFST) will produce a risk score which will help call screeners decide whether a call warrants a visit and whether there is a justification for screening the child in and carrying out an investigation.

Once the call is established as a referral, call screening staff will be able to search KIDS, the child welfare electronic information system, to determine whether any of the people named in the referral are already in the system. If so, there will be an ID number for those people, which will allow immediate linking of data held about them from various sources including health or court records and previous welfare contacts. (Temporary IDs will be created where none is held or where there is insufficient information to identify a person. Permanent or corrected IDs will be added retrospectively once all the information is established). Once identity and basic relationships are established — typically a few hours after the call arrives — a risk score and data visualization will be generated.

Calls typically refer to multiple people and the risk score will relate to the *call as a whole*. The risk score will present the *maximum* risk score for all children in the referral. While calls will identify a child who is named as a victim and other children living in the house as "other children," the AFST will score every child in the referral regardless of whether they were identified as the victim.

PARTICULAR ETHICAL ISSUES

a. Consent

Predictive risk modeling often generates significant difficulties around obtaining meaningful consent from those whose information is used and for whom risk profiles are generated. Typically, data will be aggregated in ways that make it difficult to trace clear relationships between data-providers and end-users, and data collected for one purpose will typically be used for another. Under those circumstances it is difficult — perhaps impossible — to design effective informed consent procedures. (These difficulties are exacerbated where individuals really have no choice about whether to provide the information at the outset. That will be the case *de jure* with criminal justice and birth data and may be the case *de facto* if individuals cannot, for instance, access essential services or support without providing the data.)

This is one of a number of points at which we think that it is ethically significant that the AFST will provide risk assessment in response to a call to the call center, rather than at the birth of every child. In the latter case there is no independent reason to think there are grounds to override default assumptions around consent. The fact there has been a call, however, provides at least some grounds to think that further inquiry is warranted in a particular case.

In addition, accessing data in response to a call will reduce the numbers of families or individuals whose data is being accessed by the tool and so reduce the overall incidence of access to family or individual information.

Finally, if DHS were *already* entitled to access the data gathered by the tool in response to a call, then it seems legitimate to regard the use of the tool at that point as a new and more effective way of doing something already permitted. The force of this point depends, we think, on the extent to which the AFST delivers information that would have been available, *in principle*, to a diligent call screener.

b. Information about other family members

As noted, calls typically involve multiple people: the victim, other children in the home, the mother, father and other adults. The risk score will be based on information held about all of these people. It may seem that there are significant issues around access to information about those individuals who are not the primary concern of the call. They might wonder about the justification for using information about them as part of an assessment to which they are, perhaps, only peripherally related.

We think that there should be protocols around the use of this information about individuals who are not the primary concern of the call.

Notwithstanding the need for such protocols, we believe the fact that it is at the point of call that risk assessment is carried out again has ethical significance. As above, the fact information about 'other' individuals is accessed in response to a call raising concerns about the welfare of a child provides grounds for access; accessing information only where there has been a call will reduce the numbers of families or individuals whose data is being accessed by the tool; and, while access to such information may have been more haphazard prior to the introduction of the AFST, we assume that the model does not create new rights of access to that information — that a diligent child welfare call screener would already have been entitled to gather the information now to be accessed by the tool.

c. False Positives/False Negatives

All predictive risk models will make some errors at any threshold for referral, and so, in the child protection context, identify as low risk some children who go on to experience abuse or neglect and identify as high risk some children who do not.

When considering the significance of these ineliminable errors for the AFST it is essential to keep in mind that decisions informed by predictive risk modeling tools will in almost every case have been made by some other means prior to the use of the tool and will continue to be made if such tools are not adopted. Consequently, ethical questions about predictive risk modeling tools are essentially and unavoidably *comparative*: they are questions not simply about the costs and benefits of a particular predictive risk modeling tool, but also about how those costs and benefits compare from an ethical perspective with the costs and benefits of plausible alternatives. They must be considered in light of alternatives that carry costs of their own.

And, while it is true that all predictive risk modeling tools will make errors at any threshold, at is also true that they are both more accurate than any alternative — they make fewer errors than manually driven actuarial risk assessment tools and even very good child protection professionals relying on professional judgement and experience — and they are more *transparent* than alternatives, allowing those assessing a tool's performance to accurately identify likely error rates and to accommodate them in responses to the predictions of a particular modeling tool. The greater accuracy and transparency of predictive risk modeling tools also allows them to serve as (inevitably imperfect) checks against well-understood flaws in alternative approaches to risk assessment.

So, while one should of course reduce the false-positive/negative rate as far as possible (by, for example, choosing higher thresholds for intervention, though that will carry its own costs), one can also reduce the ethical significance of false-positives and negatives by, *for instance*:

1. Providing opportunity for experienced child welfare professionals to exercise judgment about appropriate responses to a family's identification as at-risk. *(We note that one possible response to high risk scores under the AFST are mandated home visits, which would provide just this sort of opportunity)*
2. Ensuring that professionals who are using information provided by predictive modeling tools understand the potential of those tools to mis-categorize families
3. Providing training to guard, in so far as possible, against confirmation bias in the professional engagement with families identified as low- or high-risk
4. Ensuring that intervention triggered by identification as at-risk is positive and supportive rather than punitive
5. Ensuring that intervention triggered by identification as at-risk is as non-intrusive as possible consistent with the overall aims of reducing child maltreatment risk
6. Identifying and minimizing the adverse effects of identification as at-risk, such as, for instance, possible stigmatization

d. Stigmatization

There are obvious burdens associated with identification as an at-risk child or family. Those burdens may range from those that are fairly straightforward and transparent, and to some extent at least under the control of social services, to the more complex and diverse burdens of social stigmatization. We should not underestimate the significance of stigmatization:

- The associated burdens may be borne in *anticipation* of conduct that might never come to pass.
- In many cases, the burdens that follow from being identified as a member of a group arise from false beliefs about what that identification means. The burdens associated with identification as an at-risk individual or group may actually increase risk of the adverse outcome.

- The burdens of stigmatization often fall upon those who are already the subject of social disapproval or demarcation, 'appropriating and reinforcing pre-existing stigma'

These are matters for significant ethical concern. Again, however, it must be remembered that that they are not distinctive of predictive risk models. It would be naive to suppose, for instance, that negative conclusions were not already drawn from correlations between child maltreatment and socio-economic position, that existing approaches to child protection did not carry risks of confirmation bias, of unwarranted intrusion on families who were not at risk, of appropriating and reinforcing existing stigma. The point is not to suggest that these costs can be disregarded, but to emphasize the importance of weighing the costs and benefits of implementing the AFST against the costs and benefits of alternatives. Plausibly, for instance, the AFST may reduce some of these potential burdens, allowing child protection professionals to avoid confirmation bias more effectively, and allowing more effective targeting of services that, while not eliminating unwarranted intrusion, may reduce it.

In addition, we believe that there are responses to stigmatization that can at least reduce its impact and which tip the balance in favor of predictive risk modeling. Those responses include:

i. Maintaining careful control over the dissemination of the 'product' of the AFST. Access to risk scores and visualization should be distributed only to those who a) have appropriate training and b) need the information in order to further child protection goals.

ii. Provide appropriate training targeted at reducing stigmatization and its negative effects. Such training might be expected to:

   a. Emphasize the possibility of false positives/negatives.

   b. Emphasize that even given high confidence in risk scores, they are *only* risk scores and predictions. Individuals identified as at high risk must not be treated as though they have already been victims or perpetrators.

   c. Include training against confirmation bias, one of the most obvious dangers of stigmatization.

In addition, many of the responses to false positives/negatives set out above will also be directly relevant to concerns about stigmatization.

e. Racial Disparity

Many of the issues around false positives/ negatives and stigmatization are manifest in problems associated with racial disparities in the data upon which the AFST would rely. The researchers have established that current decisions around referring and placing children who are the subject of calls are affected by race. Overall, black children are almost three times more likely to have some interaction with the child welfare system than white children. Having been referred, black children are also more likely than white children to be screened in and placed. If they are screened out, black children are more likely than white children to be re-referred and placed.

Note that these disparities are to be found in the existing data. They exist independently of predictive risk modeling. The difficulty for the AFST is that such disparities in the data are potentially reinforcing. If the AFST relies upon existing data it will see evidence that black children are at higher risk than white children. If the disparities in the data reflect genuine underlying differences in the need for protection – perhaps because ethnicity tracks socio-economic disadvantage – they may not be of cause for concern: they might reflect underlying need rather than bias. If the disparities do reflect race-based bias, however, they may be ethically problematic.[1]

1 The researchers seem to show that poverty is not sufficient to explain the different referral and placement rates.

A well-known and ethically problematic example of racial disparity and its effects on predictive risk modeling occurs in the criminal justice context. In the U.S., young black men are more likely to be stopped and searched by police than their white counterparts, and having been stopped and searched are more likely to be arrested both because the stop and search provides opportunity to find evidence of offending such as drug possession, and because police are more likely to arrest young black men for offences for which their white counterparts are more likely to receive a warning. It is clear that these contacts and arrests arise to a significant extent because of racial bias. The contacts and arrests appear in the data used by predictive risk modeling tools to predict offending. Since those tools find greater evidence of contact and arrest for young black men, they are likely to place young black men in a higher risk category than their white counterparts, and since the contact and arrests reflect bias and not underlying criminality, that risk classification is unwarranted. The use of predictive risk modeling in such contexts requires at least great care lest it reinforce stigmatization, bias and disadvantage.

Examples such as the stop and search case might lead one to think that predictive risk modeling is inappropriate in contexts where one cannot be sure that data is not affected by racial bias, or at least that one should ensure that race is not taken into account by tools used in those contexts. However, there are important differences between the stop and search case and the modeling proposed in the AFST. A predictive policing tool may well recommend stopping and searching young black men *because* they have been stopped and searched in the past. That intervention is not designed to prevent future stops and searches. We think it matters in the AFST case that while a history of engagement with child protection services may lead the AFST to overstate the actual risk status of a child or family, the intervention which flows from that classification is designed and intended precisely a) to identify that family or individual's actual risk status through home visits and professional judgement, and b) to address in so far as possible any risk factors which are found to exist. It matters, ethically, this is to say, that a high risk score will trigger further investigation and positive intervention rather than merely more intervention and greater vulnerability to punitive response. We believe, that is, that the fact that the AFST will prompt further detailed inquiry into a family's situation and that any intervention is designed to assist gives grounds to think the model is not vulnerable to the legitimate concerns generated by the existence of disparities in data used in punitive contexts.

We note that the research — although not intended to show the effectiveness of field screening — suggests that such screening *reduces* the effects of disparities in the child protection data. Under the current system as we understand it, all children under seven who are the subject to a call must be field screened. Field screens appear to correct for the bias that sees a disproportionate number of black children referred and placed. The researchers write that:

> *We find that when call screeners were forced to field screen, they were more inclined to screen out black children, whereas when they did not have to conduct field screens (age seven and older), they were more inclined to screen in Black children compared to White children. This suggests that the requirement for more information (i.e. via a field screen) reduced the disparities in screening (Vaithianathan et al, 23 March, 2016, 8)*

Note, as an aside, that this appears to be an example of the additional transparency of predictive risk models over alternatives, suggesting that it is possible to track and correct for disparities that may have remained hidden under alternative approaches. More generally, it is important not to understate the burden that engagement with child protection services may place on families, but it is also important not to respond to the disparity issue in ways that worsen or leave unaddressed the position of children who might be helped.

f. Professional Competence/Training

As we have mentioned at a number of points, it is essential — if predictive risk modeling tools are to operate ethically — that staff using and relying upon them are competent with their use and interpretation. The use of such tools must be accompanied by appropriate training to ensure that competence. We set out some specific elements of such training under the stigmatization discussion above where we mentioned training to recognize the possibility of false positives/negatives; to see that even given high confidence in risk scores, they are *only* risk scores and predictions; and to recognize and guard so far as possible against common reasoning flaws and biases.

g. Provision and identification of effective interventions

Predictive risk modeling is a form of screening. So regarded, it is natural to suppose that it is subject to ethical constraints taken to apply to screening programs. One of the current reviewers has discussed the relevance of the standard statement of these constraints, the WHO Screening Principles, for predictive risk modeling in the child maltreatment context. We will not repeat that analysis here, but simply indicate that accurate predictive risk models appear to perform well under the principles (see Dare, 2013, pp. 36-47).

We think, however, that it is worth specifically mentioning one of the WHO principles. Principle 2 specifies that in order for a screening program to be ethical it must be the case that "[t]here should be a treatment for the condition" for which screening is being carried out. Dare argues that that principle is best seen as resting on the idea that screening programs which might

themselves generate harms must be capable of delivering countervailing benefits (Dare, 2013, pp. 43-44) and argues that there is sufficient evidence that interventions prompted by predictive risk models in the context of child protection meet this demand.

Here we wish to make that point in more general terms. One ethical concern about the AFST springs from the question "why pursue better prediction, if services offered will not be evidence-informed; those most likely to result in hoped for outcomes." We view this as an ethical problem. And there is another one. Why predict better if staff are not well trained in the conduct of empirically informed assessments? How well trained are they in common factors related to positive outcomes such as empathy and warmth? Yet another is how well trained staff are in gathering valid outcome measures. This raises questions concerning what will happen after risk scores are acted on. What good does it do for example to diagnose more asthma if nothing is done about it that is effective?

Drawing attention to these concerns may be a potential bonus (and an ethical one) of the use of more accurate risk prediction. Professional decision-making is not a one-shot affair. There is a sequence of decisions, each potentially affected by earlier ones, each of which may or may not be acted on as an opportunity to direct decisions in a more positive direction. It is our hope that the use of a more accurate risk estimation will highlight these other issues that affect quality of care for clients.

h. Ongoing monitoring.

The last point leads naturally to another: *Since* professional decision-making in the child protection area is not a one-shot affair, it is essential, we believe, that the County commit to ongoing monitoring of the AFST to ensure that the tool and staff training in its use is maintained, and that the interventions remain as effective as possible. The tool does generate legitimate ethical concerns and those issues must be monitored, and the justification for the burdens the tool imposes requires DHS to identify and implement reasonably effective counter-balancing responses.

i. Resource allocation.

There is an assumption implicit in the discussion in the last few sections that can usefully be made explicit. Whether the AFST is ethical depends to a large extent on its capacity to deliver benefits sufficient to outweigh its costs. We believe that it has the capacity to meet that standard. However, its doing so will require, in addition to training and monitoring and effective intervention, the provision of adequate resourcing. The AFST must not, on ethical grounds, be seen as an opportunity to reduce child protection resourcing or to reallocate child protection professionals in ways that prevent the tool from delivering the benefits upon which its ethical justification relies.

IN SUM

In our assessment, subject to the recommendations in this report, the implementation of the AFST is ethically appropriate. Indeed, we believe that there are significant ethical issues in not using the most accurate risk prediction measure.

Instruments that are more accurate will result in fewer false positives and false negatives, thus reducing stigmatization (false positives) and more lost opportunities to protect children. It is hard to conceive of an ethical argument against use of the most accurate predictive instrument.

As we have emphasized throughout, decisions are being made right now. It is not a matter of making or not making related decisions. The decisions involved are complex ones made in a context of inevitable uncertainty that contributes to inevitable error. Research on decision-making in the helping professions highlights the play of biases and fallacies. Confirmation biases are common in which we seek information that corresponds to our preferred view (e.g., there is no abuse) and fail to seek evidence that contradicts preferred views. Errors of omission (failing to act) are viewed as less harmful than errors of commission (acting - for example, removing a child from the care of her family). The question is, how can we make the fewest errors in our efforts to protect children and families? AFST seems an ethical and potentially important contribution to that effort.

REFERENCES

Dare, T. (2013) *The Dare Report: Predictive Risk Modelling and Child Maltreatment: An Ethical Review*, Ministry of Social Development, Wellington, New Zealand. http://www.msd.govt.nz/documents/about-msd-and-our-work/publications-resources/research/predictive-modelling/00-predicitve-risk-modelling-and-child-maltreatment-an-ethical-review.pdf

Dare, T (2015) 'Predictive Risk Modeling and Child Protection: An Ethical Analysis' in Challenging *Child Protection: Directions in Safeguarding Children* eds. Janice McGhee and Lorraine Waterhouse (Edinburgh; Jessica Kingsley Press, 2015) pp. 64-76

Gambrill, Eileen, and Aron Shlonsky. 'Risk Assessment in Context', *Children and Youth Services Review* 22, no. 11 (2000): 813-37

Gambrill,E.(2012)Critical thinking in clinical practice: improving the quality of judgements and decisions.Hobokan,N.J.: John Wiley & Sons

Vaithianathan, Rhema, Nan Jiang, Tim Maloney, Emily Putnam-Hornstein, (6 February 16) 'Implementation of Predictive Risk Model at the Call Centre at Allegheny County'

Vaithianathan, Rhema, Nan Jiang, Tim Maloney, Emily Putnam-Hornstein, 23 March, 2016, 'Developing Predictive Risk Models At Call Screening For Allegheny County: Implications for Racial Disparities'.

SECTION 3

Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County

Response by the Allegheny County Department of Human Services

The Allegheny County Department of Human Services (DHS) solicited the feedback of an independent team of ethicists regarding the Allegheny Family Screening Tool (AFST). Tim Dare of the University of Auckland and Eileen Gambrill of University of California - Berkeley reviewed the AFST's planned design and explored general ethical considerations. DHS is in agreement with the reviewers' conclusions, which indicate that the AFST is ethically consistent with DHS's values and principles. Most importantly, DHS agrees with the ethicists' assessment that, given the AFST's demonstrated accuracy above current decisions, "...there [would be] significant ethical issues in not using the most accurate risk prediction measure." The following outlines DHS's response to the analysis, as well as details about how DHS has incorporated ethical findings into the tool's design and implementation.[1]

1 Some of the reviewers' specific ideas are summarized, but will not be repeated with full context; we assume that the reader is also familiar with the original ethical analysis which can be found at www.alleghenycountyanalytics.us

1. Consent and privacy not considered to be areas of concern

The reviewers identified two topic areas that might typically raise questions in predictive risk modeling: (a) client consent and (b) the appropriateness of accessing/utilizing information of individuals only indirectly associated with the maltreatment event. However, after considering the ethical analysis and the following factors, DHS does not consider these to be relevant concerns with the AFST:

a. The tool is accessing no additional data other than that which is *already* accessible by call screening workers.

b. DHS already owns — and maintains the rights to utilize — all data that the tool is accessing for the purpose of protecting and serving children and families.

c. As implemented, the tool's content/output is being strictly limited to the same individuals who would already be using such data in their decision-making.

Additionally, from a legal standpoint, DHS complies with HIPAA's privacy and security rules with regard to client information. It believes that sharing its protected client information is important and, at times, critical for care, and also maintains the right to have and to re-disclose client protected information in its role as a contracting entity and as a government service coordination and oversight entity. All data use within the AFST is consistent with DHS's existing data use policies with regard to HIPAA.

2. The importance of judging the tool in comparison to the status quo

The ethicists acknowledged a number of performance challenges that the tool will inherently face. For example:

- **Error margins:** Even models that are highly accurate on average have error margins, estimating certain referrals as either higher- or lower-risk than their "true" level.
- **Racial disparity:** The data underlying the tool reflect racial disparities.

DHS agrees that these performance issues are meaningful and is in agreement with the key perspectives of the reviewers; i.e., that *decisions are already being made daily by call screeners* that are equally subject to any of these imperfections that the AFST would face, so the AFST should be viewed in comparison to the status quo. Given that the existing decision processes already are subject to errors, assumptions/biases and racial disparities, the AFST's performance at least has the advantages of being (a) more *accurate* than current decision-making strategies and (b) inherently more *transparent* than current decision-making strategies.

Despite the AFST's advantages in regard to accuracy and transparency, these performance challenges should still be monitored and mitigated as much as possible. But DHS agrees with two other ethical perspectives of the reviewers: 1) that the ultimate interventions aim to be protective in nature (rather than punitive) and 2) that the AFST's application at the early screening decision stage still allows for the investigation phase, in which additional information/decision-making will help to confirm or deny the appropriateness of the referral for services.

3. Training, monitoring and implementation efforts

Beyond the actual design, the reviewers' analyses emphasized that the context surrounding the tool — including appropriate training, ongoing monitoring and implementation — are critical from an ethical perspective. The ethical considerations have helped inform these activities.

- **Training**
  DHS developed and delivered three hours of staff training prior to the AFST's implementation. Informed by the reviewers' suggestions, the training emphasized the AFST's specific meaning and limitations, and explored how its content should be appropriately incorporated into decision-making. Call screeners engaged in a group discussion of real-world referral vignettes covering diverse scenarios, viewed the associated screening score, and discussed how the score may or may not influence the screening decisions. Additionally, a thorough job aid document is being developed to help ensure ongoing consistency surrounding the use of the AFST.
- **Tool Evaluation and Ongoing Quality Assurance**
  The ethical analyses found ongoing monitoring to be essential. To that end, DHS has contracted with two separate entities to evaluate the performance of the tool. One organization will be thoroughly assessing the implementation and business process changes, while the other will be analyzing the tool's quantitative impact on system trends and outcomes. DHS will also be carefully monitoring the internal use and impacts of the tool. Automated weekly support reports were developed alongside the AFST, and DHS analysts will be routinely providing on-site support and informal interviews with call screeners in the early weeks of its use. DHS also intends to have the content of the model revisited within the first year to make sure its statistical performance is still strong and to provide any necessary updates to the underlying weights.
- **Design and policy considerations**
  Many design elements were conceived within the context of ethical consideration:
  a. Because the tool is not perfect, the official policy for its use makes clear that the screening score is only an additional piece of information, one that should never override the workers' clinical judgment regarding the appropriateness of investigating a referral.
  b. Consistent with the ethical analysis, the AFST score will only be accessible by workers who have been trained and who have a direct need to access the score.
  c. We share the reviewers' concern that better prediction is just one element in a continuum that must end in better, more evidence-based interventions. Our immediate concern is in identifying the right children for an investigation (i.e., the "intervention" resulting from the prediction is the investigation). Only then are we able to identify those children and families most in need of evidence-based programming. Thus, the AFST is one key element in a child welfare system designed to improve outcomes for families and children.
  d. The launch of the tool is accompanied by an alteration in the child welfare field-screening policy, which includes lowering the age for mandatory field screens while expanding the use of discretionary field screens whenever deemed necessary (regardless of age). The reviewers noted the research team's findings that field screens may reduce disparities in child protection data.

SECTION 4

Allegheny County Predictive Risk Modeling Tool Implementation: Process Evaluation

by Hornby Zeller Associates, Inc.

TABLE OF CONTENTS

Executive Summary *2*

Introduction *4*

Methodology *5*

Findings *7*

Recommendations *19*

APPENDIX A: Initial Survey Results *20*

APPENDIX B: Follow-up Survey Results *23*

EXECUTIVE SUMMARY

Background

Allegheny County Department of Human Services (DHS) is using predictive risk modeling (PRM) to assist child welfare staff decide which General Protective Services (GPS) referrals to investigate and which to screen out.

A contracted research team developed the Allegheny Family Screening Tool by conducting an extensive series of analyses using data from the DHS data warehouse and the child welfare case management system to identify factors that are predictive of a child's re-referral to child welfare or placement into foster care. The research team developed an algorithm that applies weights to a series of factors to assist in determining when a GPS referral should be assigned for investigation or screened out.

Hornby Zeller Associates, Inc. conducted a process evaluation involving stakeholder interviews, surveys, and document review to describe Allegheny County's experience, including perceived barriers and facilitators, with implementing PRM.

Methodology

The timeline for the process evaluation is summarized in **Table E-1**, which includes a description of the strategies employed and the sources used to collect data.

TABLE E-1: **Schedule of Process Evaluation Methodology**

PRE-IMPLEMENTATION	POST-IMPLEMENTATION		
SUMMER 2016	FALL 2016	WINTER 2016	SPRING 2017
Interviews with DHS call screening and other DHS staff	Surveys of call screeners	Interviews with DHS research and practice staff Interviews with external stakeholders	Follow-up surveys of call screeners

Findings

Community stakeholders had positive feedback about the presentations introducing Predictive Risk Modeling and hope for continued transparency as the County continues to implement the Allegheny Family Screening Tool.

Considerable effort went into informing internal and external stakeholders through community meetings about the County's decision to implement PRM. External stakeholders who attended the presentations generally found them to be "encouraging" and "informative." They noted the County and its team of experts know what they are doing and inviting stakeholders to the presentation showed DHS intends to be transparent in its implementation of the tool.

Stakeholders noted the need to continue to inform community stakeholders about PRM progress, outcomes and plans for ongoing maintenance and sustainability. For instance, one provider wanted to know what the "disaster plan" is for the tool, as well as what safeguards are in place to ensure that transparency will continue in the future, regardless of who is overseeing the project.

Following implementation, stakeholders continue to have a positive reaction to implementing the Allegheny Family Screening Tool. Their hope is that the tool will result in increased safety of children and enable the County to be more proactive and less reactive in its case practice.

The Predictive Risk Modeling Tool is facilitating data-driven decision-making with Allegheny County staff, but there is further room for system-level change.

Administrators agreed that the tool will help staff to make an informed decision. During the planning period, administrators were confident the tool would lead to more accurate decision-making. More than half of the call screeners (61%) said they believe the tool is facilitating a shift in the workplace environment to be more data-driven.

Call screening staff report having a good understanding of the Allegheny Family Screening Tool, but are mixed on how confident they are in the resulting scores.

The majority of call screeners voiced some concern about the reliability of the score, with 72% stating they thought a score seemed inaccurate occasionally and an additional 11% noting it was inaccurate a moderate amount of the time. Half (50%) of the call screeners surveyed said they are confident in the tool's ability to accurately assess risk. Full-time call screeners were slightly less likely to express confidence than part-time call screeners. The lack of confidence stemmed from the concern that the tool is unable to take a family's individual circumstances into account; for instance, a family may be receiving services that are improving the family's situation. More than half of the call screeners (61%) said they are confident in the research that went into developing the tool.

Call screen staff generally find the Allegheny Family Screening Tool easy to use, and offered technical suggestions for improving the Tool's user experience.

The majority of call screeners understand how the score works and all surveyed said they are "adequately prepared to use the tool." Nearly two-thirds of the call screeners (60%) said the tool is "easy" or "very easy" to use. Approximately one-third (38%) had no opinion on the visual display of the score on the thermometer. More than two-thirds of those with an opinion (70%) said the thermometer is helpful to a limited degree or not at all.

Suggestions offered by call screeners to improve the Allegheny Family Screening Tool were primarily related to technical issues. One of the staff suggested the "score needs to be more visible." Several screeners remarked that the system is slow, noting it takes a long time for the score to generate and the tool sometimes times out.

Recommendations

Maintain transparent communication with internal and external stakeholders.

Stakeholders overwhelmingly applauded the efforts that DHS has made to be transparent and to keep them informed throughout the implementation process. It will be important that this transparency continue.

Increase user buy-in.

Less than half of the call screeners currently view predictive modeling as benefiting the screening practice, though more than 60% agreed that the tool is creating a data-driven culture within the workplace. An opportunity exists to increase user buy-in.

Continue to resolve technical issues as they arise, documenting solutions.

As changes and enhancements are made to the tool, they should be documented to inform further tool development, increasing the return on the technological investment.

Develop benchmarks for implementing predictive risk modeling.

Benchmarks can be developed to 1) foster buy-in/increase use of the tool for decision-making by call screening staff and 2) promote transparency with stakeholders.

INTRODUCTION

The Predictive Risk Tool: Development and Goals

Development of the Allegheny Family Screening Tool

Allegheny County, Pennsylvania, has a rich source of data to inform its decision making. The Data Warehouse[1] of the Department of Human Services (DHS) stores data from a wide array of sources including, among others, the juvenile and adult correction systems, public welfare and behavioral health agencies and programs. Data from the warehouse are available to aid child welfare caseworkers and their supervisors, including call screeners, in their decision making. Data integration has paved the way for the use of administrative data in predictive risk analytic models to target services to children and families most in need.

Building on a concept first developed in New Zealand[2] to target social services to families at high risk of using multiple service systems for lengthy periods of time, Allegheny County DHS elected to use Predictive Risk Modeling (PRM)[3] to help prioritize cases and target services to children most at risk. Allegheny County chose to implement PRM specifically to assist child welfare call screening staff to decide which General Protective Services (GPS) referrals warrant investigation and which should be screened out.[4]

1 Allegheny County Analytics 2017 DHS Warehouse, accessed May 19, 2017.

2 Vaithianathan, Rhema, Tim Maloney, Nan Jiang, Irene De Haan, Claire Dale, Emily Putnam-Hornstein, and Tim Dare 2012 *Vulnerable Children: Can Administrative Data be Used to Identify Children at Risk of Adverse Outcomes? Centre for Applied Research in Economics, Department of Economics, University of Auckland.*

3 Allegheny County Department of Human Services 2017 Predictive Risk Modeling in Child Welfare in Allegheny County: The Allegheny Family Screening Tool, accessed May 19, 2017.

4 Vaithianathan, Rhema, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney 2017 *Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation. Centre for Social Data Analytics, Auckland University of Technology.*

Through a competitive Request for Proposals[5] DHS contracted with an international team of researchers, led by Rhema Vaithianathan from the Auckland University of Technology in New Zealand and joined by Emily Putnam-Hornstein from the University of Southern California, Irene de Haan from the University of Auckland, Marianne Bitler from the University of California-Irvine, and Tim Maloney and Nan Jiang from the Auckland University of Technology, to develop the Allegheny Family Screening Tool. The research team conducted an extensive series of analyses using data from the DHS data warehouse and from the County's child welfare case management information system, the Key Information and Demographic System (KIDS). The analyses identified factors that are predictive of re-referral to child welfare or placement into foster care, and produced an algorithm that applies weights to those factors to assist in identifying which GPS referrals are more or less at risk of these outcomes.

5 Decision Support Tools and Predictive Analytics in Human Services RFP

Introduction to the Process Evaluation

As part of Allegheny County's effort to document and evaluate the implementation of predictive risk modeling, Hornby Zeller Associates, Inc. (HZA), a management consulting firm specializing in evaluations of public human service programs, was contracted through a competitive bid process to conduct the process evaluation of Allegheny County's implementation of PRM.[6] Casey Family Programs and the Human Service Integration Fund of The Pittsburgh Foundation provided funding for this evaluation (and a separate impact evaluation conducted by Stanford University). This report provides an overall summation of the process evaluation conducted by HZA between mid-2016 and early 2017. The report considers the steps the County took prior to and during the initial implementation, the reactions of internal and external stakeholders to predictive risk modeling in Allegheny, challenges that arose and were addressed, as well as lessons learned during the implementation process.

6 Evaluation of a Predictive Risk Modeling Tool for Improving the Decisions of Child Welfare Workers RFP

METHODOLOGY

Data Collection

Interviews

Prior to implementation of the Allegheny Family Screening Tool, HZA conducted interviews with DHS administrators and staff from the call screening unit to learn about: a) their involvement in the implementation of the tool, b) steps taken to prepare call screening staff to use predictive risk modeling to inform their decision-making, and c) the call screening process as it existed prior to implementation of the tool. In all, 23 staff were interviewed at baseline, including: 3 administrators, 8 call screen supervisors, and 12 call screeners. Both part- and full-time call screening staff were interviewed, with 55% of the call screening staff providing input. These interviews were conducted in July of 2016, just prior to the tool being implemented on August 1, 2016.

Four months following implementation of the tool, HZA conducted interviews with stakeholders internal and external to DHS. Interviews with community partners focused on their awareness of

the Department's efforts to implement PRM, their hopes for what the tool would accomplish and the successes and challenges they expected the County to face. Internal stakeholders were asked about their involvement in implementing the tool, the training they received and how the Allegheny Family Screening Tool informs or impacts their work. In all, a dozen individuals were interviewed post-implementation, half of whom were from the Department's Office of Children, Youth and Families (CYF). Other DHS stakeholders included an administrator and staff from the Office of Data Analysis, Research and Evaluation. Representatives from community service providers, advocacy groups, foundations, and a family court judge made up the group of external stakeholders who were interviewed.

Surveys

In September 2016, approximately 2 months post-implementation, HZA administered a web-based survey to call screeners. A total of 16 of 21 call screeners completed the survey for a response rate of 76%. Three-quarters or 12 of the respondents were full-time call screeners. More than half of the survey respondents (56%) had worked as call screeners for at least three years.

Using a series of Yes/No and Likert scale questions, call screeners were asked about the training they received, the functionality of the tool, visualization of the scores, and the impact of the tool on their decision making. Several open-ended questions were also asked to gather input on what could be done to improve the use of the tool and the training provided to prepare staff to use it.

Following a meeting with project staff in early February 2017, which included staff from DHS, representatives from the research team and the process evaluation and outcome evaluation teams, a decision was made to administer a follow-up survey to call screeners to account for improvements that had been made to the tool.

A total of 18 call screeners responded to the follow-up survey for a response rate of 86%. All full-time screeners responded to the survey while 60% of the part-time screeners participated in the second survey. Just under half (48%) of the second survey respondents reported working as call screeners for at least three years.

Data Analysis

Quantitative Analysis

Quantitative analyses included summary statistics, frequency counts and percentages. Aggregate results of the surveys are provided in the Appendices.

Qualitative Analysis

Data collected during the stakeholder interviews and through open-ended questions on the call screener surveys were carefully reviewed to identify common themes and items of importance. The results of the qualitative analysis describe the implementation process from the perspective of the stakeholders, a grounded theory approach.[7] The qualitative results are also used to support and/or explain the quantitative results, where appropriate.

7 Charmaz, Kathy 2000 Grounded Theory: Objectivist and Constructivist Methods. In *The Handbook of Qualitative Research, edited by N. K. Denzin and Y. Lincoln, pp. 509-535. Sage Publications, Inc., Thousand Oaks, California.*

FINDINGS

Pre-Implementation: Preparing for Change

Call Screening Practice

The primary role of call screeners prior to implementation of the Allegheny Family Screening Tool was to gather the information to inform supervisor decision making. Specifically, call screeners collected data about the alleged victim(s), perpetrator(s) and the allegations of suspected maltreatment. Information was collected from four primary sources: 1) the caller, 2) KIDS, 3) the data warehouse and 4) public databases that contain court and jail information. This information was provided to supervisors, who made the decision to screen the call in or out. The information gathered by call screeners was also provided to caseworkers to aid the assessment process after a call was screened in.

Call screeners reported that it is much easier to collect information from a mandatory reporter than from other callers because mandatory reporters are aware of the information they need to supply. Regardless, screeners said they collect as much information about the alleged victim(s), the child(ren)'s family and the alleged perpetrator (e.g., names, addresses, ages, relationships) from the caller as possible, as well as descriptions of the alleged maltreatment. Screeners reported that they check KIDS for every referral to determine if there is already a case open on the child or family, in which case they provide the information to the responsible caseworker, or if the family had past involvement with the Department.

Beyond the information collected from the reporter and KIDS, cross-sector administrative data are available from other County agencies and community providers through a tool commonly known as ClientView.[8] External databases, such as Prothonotary (the Allegheny County Court screen) and PAC file (the Juvenile Court data system), among others, are also searched. When the call screeners were asked during the pre-implementation interviews how frequently they search the data warehouse for data about the family, some stated they check it for nearly every report, while others report that they use it less than half of the time. Staff did report consistent use of the court information.

While a number of call screeners reported that the data they obtain through ClientView contain information that is not accurate or up-to-date, such as a previous address for a family that has moved, call screeners also reported using ClientView to "fill in gaps" in client information and to gain a better picture of a child or family's situation. Call screeners reported taking 10 to 15 minutes to complete a search in ClientView. Several call screeners, however, reported taking as much as an hour, or even several hours for cases with prior history with the Department. When asked how long it typically takes to collect information on a referral, including gathering the information and completing the call report for the supervisor, staff noted it took between 25 and 35 minutes on average. A major factor in the time it takes to complete the intake process, as described by the call screeners, involves gathering information from the caller; the more information the caller has and is able to provide, the longer it takes to complete the screening.

8 Vaithianathan, Putnam-Hornstein, Jiang, Nand and Maloney, 2017, *Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions, Auckland University of Technology in New Zealand*

Sharing Information with Supervisors and Caseworkers

In addition to describing their data collection processes prior to implementation, call screeners were asked how they document and share the information they collect from the various data sources with their supervisors. All screening staff noted that they discussed the reports with their supervisor and often gave their input, but the ultimate decision to refer the report for investigation was made by the supervisor. Thus, call screeners were primarily tasked with gathering the information that supervisors used to determine whether to screen calls in for an investigation or to screen calls out without further child welfare involvement (referrals may be made to appropriate resources). All staff stated that there are sections within the call report template that facilitate how the information was documented. For example, there was a section labeled "Legal" where information regarding a family's court involvement, if applicable, was recorded. According to the call screeners, much of the information collected in the screening process went in the "Additional Information" section. The information collected from KIDS and through ClientView during the call screening process was also available to investigation caseworkers for planning and conducting investigations.

Activities and Responses to Preparing for Change

Policy and Practice Changes

A shift in policy was made to guide several changes Allegheny County introduced to the call screening practice in conjunction with PRM. These changes are illustrated in **Table 1**. Call screeners, rather than just collecting information, were now being asked to complete a risk and safety rating and to generate a family screening score using the Allegheny Family Screening Tool for each child associated with an allegation of maltreatment. Additionally, call screeners now make the recommendation to screen non-mandatory GPS reports in or out while supervisors became responsible for reviewing and approving the call screeners' recommendations.

TABLE 1. CHANGES IN CALL SCREENING POLICY AND PRACTICE

BEFORE IMPLEMENTING PRM	AFTER IMPLEMENTING PRM
Screening decisions made based on clinical judgment.	Screening decisions made based on systematic analysis of data and clinical judgment.
Call screeners collect information.	Call screeners collect information, complete a risk and safety rating and initiate generation of a family screening score.
Supervisors make decision to screen calls in or out.	Call screeners make recommendations to screen calls in or out and supervisor approves or changes.

Informing External Stakeholders

Considerable effort went into informing external stakeholders about the County's decision to implement predictive risk modeling into its call screening process. Community meetings were held to introduce the project to external stakeholders, including advocacy groups, service

providers, court staff and consumer groups. These presentations highlighted the tool's design and underlying research, as well as how it would be integrated into call screening decisions.

The external stakeholders who attended generally found the presentations to be "encouraging" and "informative." The degree to which the presentations enhanced their understanding of Allegheny's application of predictive risk modeling, however, varied. For instance, one community provider found the information to be very helpful and condensed into pieces that were easy to comprehend. Another noted she had to attend a few of the presentations to understand predictive analytics, because the topic is "very complex."

The community meetings DHS held for community stakeholders discussed ethical issues the County was facing related to implementation of the screening tool. One topic of interest was security and privacy, and whether or not the tool would collect or share any new data regarding families. Presenters explained that the tool only leverages data that are already collected and owned by the County. Other than to use the data in a more structured and consistent manner in making a decision to screen in or out a GPS referral, the data are not intended to be used other than they have in the past.

Discussions with stakeholders also invoked the possibility of the screening tool maintaining or exacerbating racial or socioeconomic disparities. Allegheny County's historic data suggest that racial disparities already exist at many outcome and decision points throughout the child welfare system.[9] Presenters suggested that ideally, the tool increases transparency and consistency in decision-making, as well as reduces the possibility of call screeners needing to draw from their own implicit biases. In the spring of 2016, an analysis of the ethical questions surrounding the tool was conducted to explore race's possible role in the tool. Ultimately, in conjunction with the researchers' finding that including race in the model did not significantly improve its accuracy, administrators, in conjunction with ethics and legal staff, determined that race would be omitted as a factor for determining the risk score.[10]

9 Rauktis, Mary E. and Julie McCrae 2010 *The Role of Race in Child Welfare System Involvement in Allegheny County. Allegheny County Department of Human Services, Pittsburgh, Pennsylvania.*

10 Dare, Tim, and Eileen Gambrill 2017 Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County. In Vaithianathan, Rhema, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney 2017 *Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation. Centre for Social Data Analytics, Auckland University of Technology.*

Anticipated Benefits and Challenges

The goal of implementing predictive risk modeling in Allegheny County, according to DHS administrators, was broadly to improve decision making. Collectively, administrators listed six goals related to determinations made at intake, a "key decision point," as one administrator expressed it (See **Table 2**).

TABLE 2. Goals of Implementing Predictive Risk Modeling in Allegheny County

Change the agency culture to data and research based decision making.
Make better and more efficient use of resources, specifically data resources.
Make decisions based on as much information as possible.
Increase the number of people making call-screening decisions
Create a more uniform screening practice.
Increase the accuracy of screening decisions.

Along with the specific goals, administrators expressed several benefits they hoped would result from using predictive risk modeling. The most frequent response regarding the intended benefits centered on the accuracy of decision making.

Concerns were also expressed by DHS staff prior to implementation that the volume of investigations might increase and that implementation was being done at a time when there had been many recent legislative changes, which might complicate the implementation. One of the DHS administrators interviewed prior to implementation voiced a concern that some calls would be mandatorily screened in based on the resulting risk score, even though the information collected from the caller was not suggestive that the report be assessed. DHS elected to designate referrals with a score of 18 or higher on the placement risk model as mandatory to be screened in for an investigation, although supervisors are able to override the mandate if other factors warrant that decision.

Preparing Call Screeners and Supervisors

Staff Training

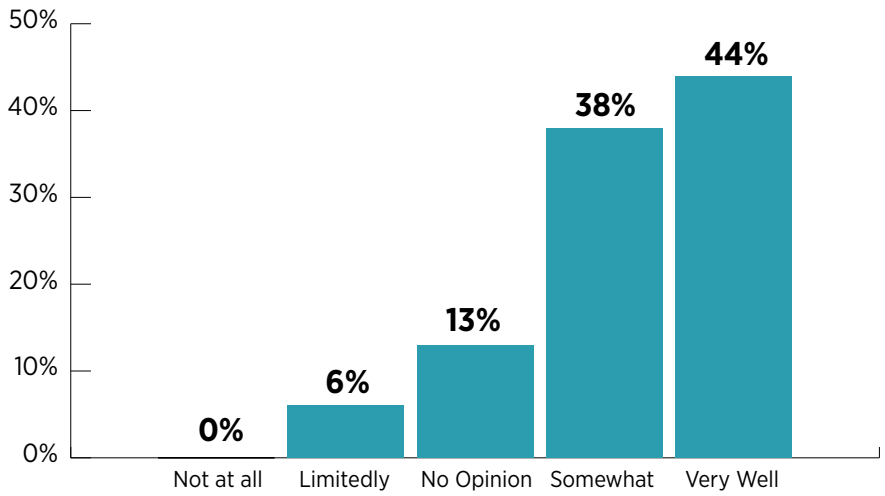
Call screening staff received training in how to use the tool in their decision making. This comprehensive training taught call screeners how to use the tool to generate and then interpret the Family Screening Score. It also provided them with an understanding of how predictive risk modeling works and what the Allegheny Family Screening Tool intends to accomplish. The process which call screeners were to follow upon implementation of PRM was described step-by-step with a workflow presented to illustrate the process. A number of case scenarios were also used to demonstrate use of the tool in knowing when to mine for additional data and apply the results in decision making.

The training also reviewed changes made to the KIDS interface in response to PRM and the added responsibility for call screeners to complete a risk and safety assessment. For instance, the training covered where screeners would record the initial risk and safety decision in the case management system, along with the factors considered in determining the appropriate level of risk and safety. It also included a demonstration of how client service data would be automatically imported into KIDS from the data warehouse, to help the screener document and justify the recommendation to screen in or out the GPS report.

Staff Perspectives on Training

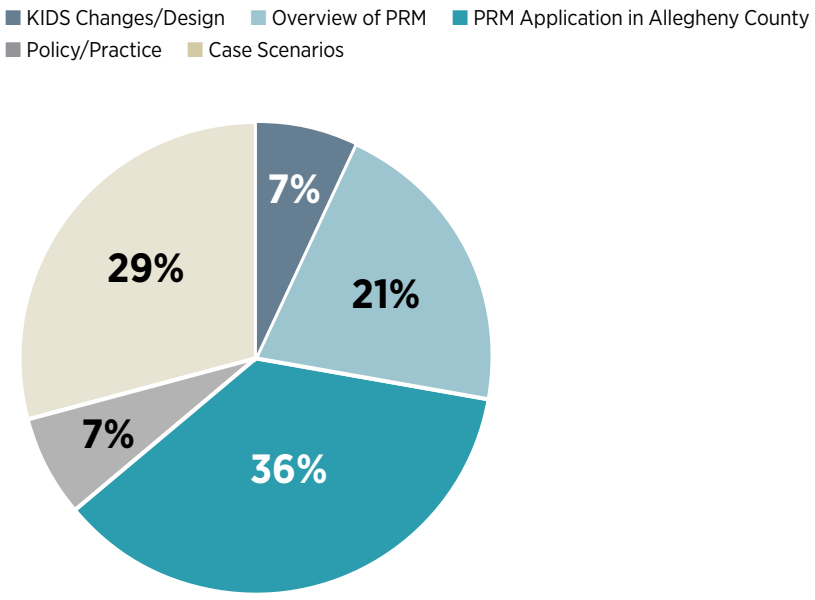
The survey administered to call screeners two months following the implementation of PRM asked if they had received training prior to using the Allegheny Family Screening Tool. All 16 respondents stated that they had completed the training, with the majority of them stating that the training prepared them to use the tool either “somewhat” or “very well,” as shown in **Figure 1**.

FIGURE 1: **How Well Training Prepared Screeners to Use the Allegheny Family Screening Tool**



When asked what aspect of the training was most helpful, as shown in **Figure 2**, over a third of the call screeners (36%) said that information about how predictive analytics was to be applied in Allegheny County was most helpful. Use of case scenarios to review decision-making were rated as the second most helpful part of the training with just under a third (29%) of the call screeners reporting that training activity to be helpful. The overview of predictive risk modeling was rated the third most helpful with 21% of the survey participants indicating that it was helpful. Although the sections on changes to KIDS and those made to policy/practice were rated as helpful by fewer call screeners, they were still thought to be important components of the training.

FIGURE 2: **Training Components Found to be Most Helpful**



Call screeners were asked what could have been done differently to better prepare them to use the Allegheny Family Screening Tool. Over 40% of the screeners offered no comment or said that nothing additional was needed. The remaining respondents gave a wide range of open-ended responses, including that the tool should have been tested by Intake prior to roll-out, or at least call screeners should have been able to provide input into its design. One staff member indicated additional training would be beneficial, while another noted a handout explaining the information would have been sufficient.

Other internal stakeholders were also asked what could have been done differently to better prepare staff to use the Allegheny Family Screening Tool. One staff member from the Quality Assurance, Best Practices and CYF Analytic teams suggested that the concept of the score not being about the current allegations "needs to be said often because people forget." The score takes into account the history of household members, along with the current allegation. Another team member suggested clarification was needed about the differences in the definition of "risk." This stakeholder pointed out that the definition of "risk" according to the tool (future risk of subsequent allegations of maltreatment or placement into foster care) is different from the definition of "risk" of which staff are most familiar (imminent risk of serious harm).

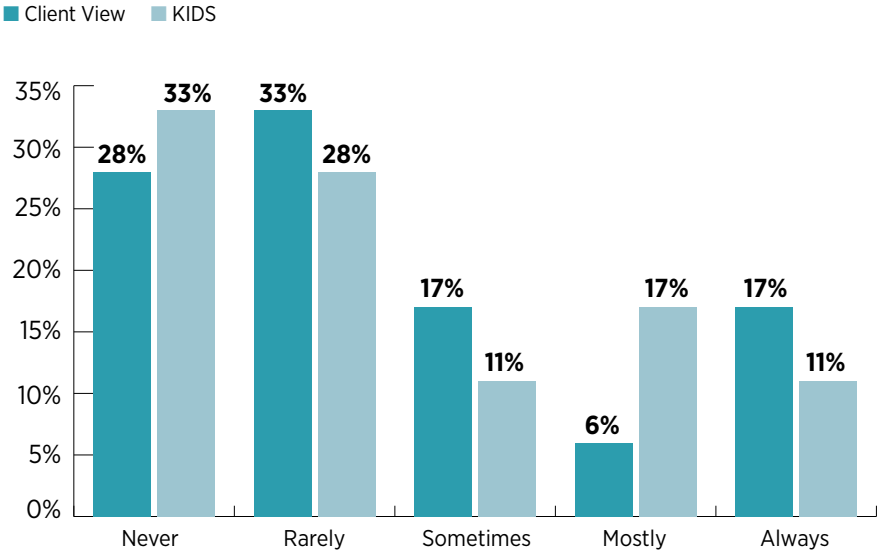
Post-Implementation: Reactions and Process Improvements

The overall goal of implementing predictive risk modeling in Allegheny County was to improve decision-making by making it more data driven and creating a uniform approach to making decisions, while increasing the number of those responsible to carry out that activity. Prior to implementing PRM, supervisors were responsible for making final call screening decisions. Since implementation, all of the call screeners as well as the supervisors are involved in making call screening decisions.

Use of Data

The follow-up survey conducted of call screeners also asked about the extent to which they conduct a more thorough search in the data warehouse, as well as KIDS, this time with a focus on reports in which the Family Screening Score was high. A little more than 60% of the call screeners who responded to the survey noted they "rarely" or "never" conduct an additional search in ClientView or KIDS. **Figure 3** illustrates the extent to which call screeners report conducting subsequent searches in ClientView when the resulting score is high. Overall, full-time screeners were more likely to conduct additional searches in ClientView than part-time call staff. When asked to explain why additional searches are not done, most call screeners said that they had done the searches in the data warehouse earlier in the process and one call screener noted that "the score stands for what is pulled" by the PRM tool.

FIGURE 3: **Additional Searching Conducted When Family Screening Score is High**



Call Screener Attitudes and Beliefs

When asked how confident call screeners are in the tool's ability to accurately assess the risk of placing a child into out-of-home care or incurring a repeat re-referral of maltreatment, half of the call screeners said they were confident. Full-time call screeners were slightly less likely to agree than the part-time call screeners. One screener explained that lack of confidence in the PRM tool stems from its inability to take families' expected improvement or individual circumstances into account, for instance, when families are receiving services that are improving their situation. When asked how confident they were in the research that went into developing the tool, 11 of the 18 call screeners (61%) reported they were confident in the research that went into developing the tool.

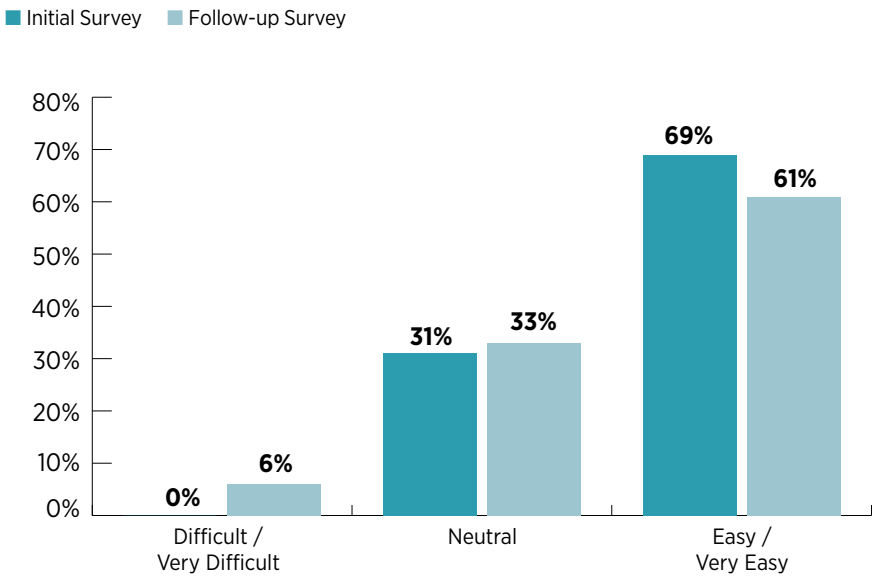
A series of statements were included in the follow up survey which were used to gauge the call screeners' understanding of the Family Screening Tool. As shown in **Table 3**, call screeners understand the intention of using the tool in making screening decisions.

TABLE 3: **Attitude of Call Screeners Toward the Allegheny Family Screening Tool**

PERCENTAGE OF CALL SCREENERS "AGREEING" OR "STRONGLY AGREEING" WITH STATEMENT	
I understand what the score is predicting.	94%
I understand how the score should relate to or inform screening decisions.	94%
I understand the content of the data sources helping to produce the score.	89%
I am adequately prepared to use the tool.	100%

Call screeners were also asked how easy or difficult it is to use the Family Screening Tool. Over 60% of the call screeners who responded to the surveys, regardless of when administered, find the Family Screening Tool "easy" or "very easy" to use, as displayed in **Figure 4**. Many of the respondents appreciate that the resulting score generated by the tool helps to validate their decision to screen in or out the General Protective Services referral for investigation.

FIGURE 4: **Ease of Navigation/Use of the Allegheny Family Screening Tool**



A graphical display, using a thermometer, is used to highlight the value of the score. Call screeners were asked in the first survey to indicate how useful the thermometer was to them. While many of the call screeners (38%) said they had "no opinion" about the graphic display, nearly half (44%) found the thermometer to be "somewhat helpful" or "helpful," explaining that it is straightforward and easy to read. An additional 19% said the graphic was "not helpful at all" or was only "limitedly helpful," noting the number could be larger and the color scheme could be improved. One call screener said that an actual number would suffice while another, who self-identified as a visual learner, liked everything about the graphic display.

Call screeners were also asked to express concerns they had about the screening tool. One of the call screeners, as noted earlier, stated that "the tool does not take the human element of judgment into account," while another stated that "the score frequently has nothing to do with what is actually going on with the situation at hand." Another said that call screeners are able to recognize information that needs to be updated, which the tool is unable to do and thus will generate a score that inaccurately portrays a family's circumstances.

Stakeholder Input

Following implementation of PRM in Allegheny County, stakeholders internal and external to DHS continue to have a positive reaction to Allegheny County's implementation of the Family Screening Tool. Their hope is that the tool will result in the increased safety of children and enable the County to be more proactive and less reactive in its case practice. An internal stakeholder noted that the tool should help with decision making, especially for borderline cases, such as when it is difficult to determine whether the case should be screened in for an investigation or screened out and possibly referred to community services. It should be noted that the goal of the PRM implementation is to use the tool to make decisions regarding every GPS referral.

The family court judge who participated in the post-implementation interviews stated that she supports use of the tool. While admittedly she does not have the knowledge needed to examine the algorithm used to generate the score, she has confidence in the people who presented the material at the community presentation she attended and in what the algorithm is meant to achieve. She stated it was clear that the County and its team of experts know what they are doing and that inviting stakeholders to the presentation demonstrates that DHS intends to be transparent in its implementation of the tool.

Another external stakeholder explained how her foundation was approached to fund the initiative, in part, based on a longstanding relationship with DHS. This stakeholder went on to say that DHS is going beyond what is required in terms of keeping the foundation informed and those at the foundation welcome the additional information and the County's transparency.

Both internal and external stakeholders noted during the interviews that there is a need to continue to keep community stakeholders informed. For example, one analyst internal to DHS thought that the community needs to remain involved and informed as the County moves forward with predictive analytics. An external stakeholder noted that the PRM community presentation has not evolved in the last two years; the same information is presented in the same manner at each meeting. One provider said it should be stressed that the score is just an additional piece of information to further assist with decision making—it is not the only factor considered. Another provider wanted to know what the "disaster plan" is for the Family Screening Tool, e.g., what safeguards are in place to ensure that transparency will continue in the future, regardless of who is overseeing the project.

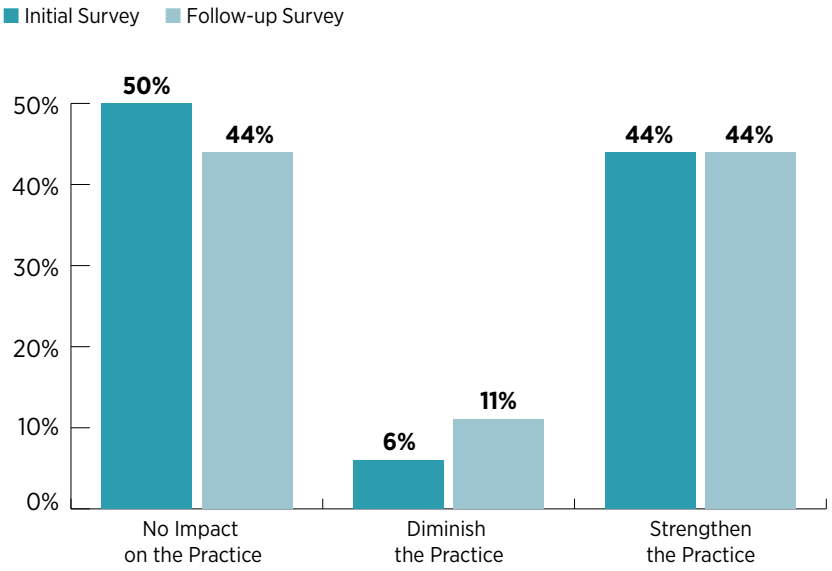
Process Improvements

Impact on Practice

The surveys administered to call screeners two months following implementation and then four to five months later gave them the opportunity to comment on whether they thought the tool would have an impact on the call screening practice. The follow-up survey also asked screeners to explain why or why not it has had an impact, with half (n=9) offering an explanation. While the initial survey found half (50%) of the call screeners did not anticipate PRM to have any impact, this changed slightly (44%) in the later survey. Some call screeners explained that mandatory

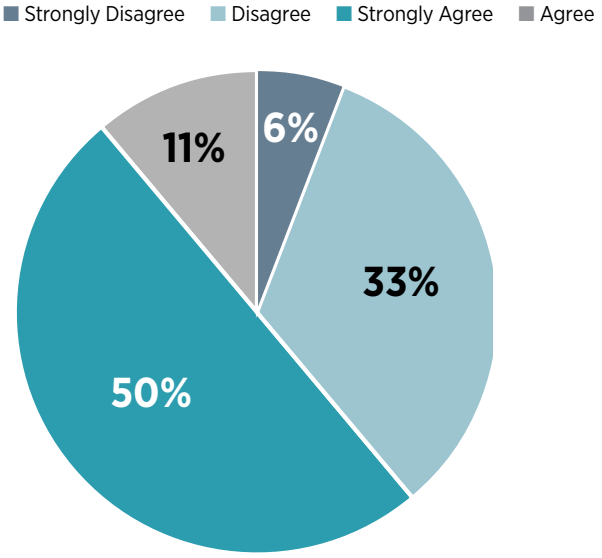
screen-ins based on a high score would impact practice. Others commented that scores should not be mandatory on active cases and decisions to assess a referral are still based on the allegations presented by a caller. As displayed in **Figure 5**, 44% of the call screeners overall, when results from both surveys are considered, thought using the tool would strengthen the call screen practice. Call screeners stated, "consistent decision making will be increased." One call screener from the initial survey responded PRM would diminish practice, while two responded as such to the follow-up survey. This may be due to the reported slowness of the system which may have become more of an irritant over time.

FIGURE 5: **Expected Impact on Call Screener Practice**



One of the stated goals of implementing PRM was to create a workplace culture that is more data-driven. Call screeners were asked their perspective on whether the use of the predictive risk modeling tool was shifting the workplace culture towards that goal. The follow-up survey administered to call screeners found that 61% of the screeners, as shown in **Figure 6**, either agreed or strongly agreed that the tool is creating a data-driven culture within the workplace. When this finding is considered along with the percentage of call screeners who said the tool would not impact call screener practice (44%), it is possible that call screeners already thought of the culture at Allegheny DHS as being data-driven. It appears the decisions being made by the screening unit prior to PRM implementation were based on good screening practices, with the tool now reinforcing those decisions through a systemized use of data.

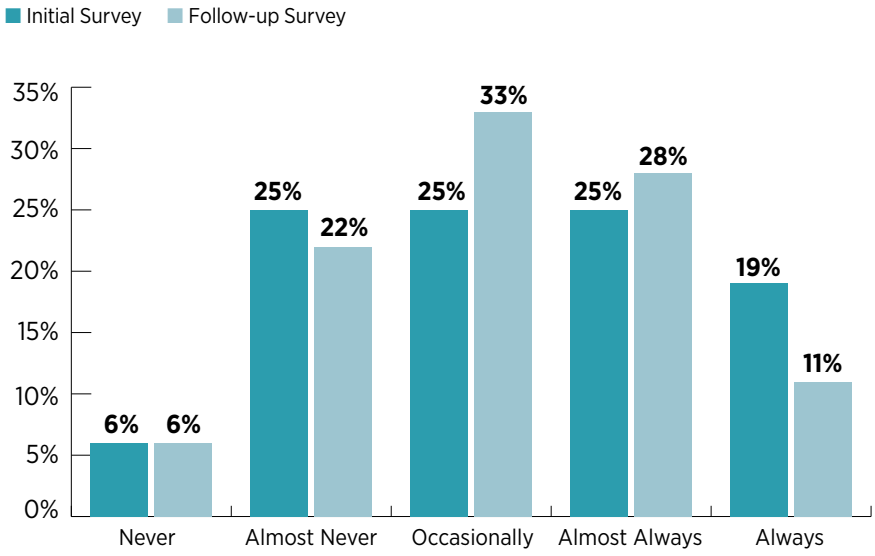
FIGURE 6: **Use of PRM Tool is Changing the Culture of Our Workplace to Be More Data-Driven**



Improved Screening Process

The survey administered to call screeners two months following implementation of the tool found that over 40% of the call screeners use the tool to inform their recommendation on a consistent basis, although a little less than a third (31%) reported they rarely use it, if at all. On the survey administered after enhancements were made to the tool, a smaller percentage of staff said they "always" used it, although a larger percentage "almost always" or "occasionally" use the tool to inform their recommendation. When asked, in the follow-up survey, to explain, three of the call screeners who did not use the tool to inform their recommendation said their recommendation was already formed prior to running the score. One call screener reported that he or she would discuss the score or concerns about the score with the supervisor while another said that he or she would provide the information to the supervisor to make the decision. One of the call screeners stated that "the score has nothing to do with assignment; active cases are always high or mandatory." **Figure 7** illustrates the shift in call screeners use of the tool to inform their recommendation.

FIGURE 7: **Use of Tool to Inform Recommendation**



Technical Improvements

Call screeners were also asked in the survey administered after the tool was enhanced how often a score was generated that seemed inaccurate given the information they had gathered about the family based on the data which are available and/or from what they had collected from the reporter. Nearly three-quarters (72%) of the call screeners noted they "occasionally" have encountered a score that seems to be inaccurate, while an additional 11% have frequently encountered an inaccurate score.

When asked to explain what they do when the score appears to be inaccurate, nearly half (n=9) said they notify a supervisor. Three staff, two full-time and one part-time call screener, reported they review KIDS and/or ClientView to inform their decision when they believe the score is not right. Another screener relied on the research he or she already completed, instead of relying on the tool to assist with that process. Yet another staff member reported contacting the technology staff regarding the inaccurate score.

Suggestions offered by call screeners during both surveys to improve the Allegheny Family Screening Tool were primarily related to technical issues. Several call screeners from the follow-up survey remarked the system is slow; staff noted it takes a long time for the score to generate, with the system sometimes "timing out." One staff member suggested the risk and safety boxes should not be locked after the score is viewed so that recommendations could be adjusted later in the process. Another screener suggested that the tool should also pull the family history with CYF into the narrative field which lists the factors used to produce the score, not just the program areas identified using ClientView.

While these suggestions may improve the functionality of the tool at some time in the future, there have been two technical issues reported that specifically affected the performance of the tool. A few call screeners noted concerns about the accuracy of the score when clients and Master Client Index (MCI) numbers[11] are duplicated within a referral. A similar issue was identified with the tool not being able to generate the score when MCI numbers were missing. These issues were corrected in November 2016; the follow-up survey identified that most screeners (83%) find the score is "mostly" or "always" clearly displayed.

11 Master Client Index or MCI numbers are used by DHS staff to identify individuals known to other agencies and providers across Allegheny County.

RECOMMENDATIONS

Major changes in policy and practice can be difficult to implement, particularly when the agency making the change is pioneering a new technological solution, such as predictive risk modeling. Allegheny County has chosen to implement a PRM tool to increase appropriate and consistent use of data to drive its decision-making. While the results of this process evaluation are encouraging, some recommendations are offered toward informing the implementation process moving forward.

1. *Maintain transparent communication with internal and external stakeholders.*
   Stakeholders overwhelmingly applauded the efforts that DHS has made to be transparent and to keep them informed throughout the implementation process. It will be important that this transparency continue.
2. *Increase user buy-in.*
   Less than half of the call screeners currently view predictive modeling as benefiting screening practice, though more than 60% agreed that the tool is creating a data-driven culture within the workplace. An opportunity exists to increase user buy-in.
3. *Continue to resolve technical issues as they arise, documenting solutions.*
   As changes and enhancements are made to the tool, they should be documented to inform further tool development, increasing the return on the technological investment.
4. *Develop benchmarks for implementing predictive risk modeling.*
   Benchmarks can be developed to foster buy-in and promote use of the tool for decision-making. Using the results of the process evaluation, Allegheny might consider developing benchmarks which would target an increase in the percentage of staff who use the tool on an ongoing basis. For example, one measure might challenge call screeners to consistently use the tool in their decision-making, e.g., by March 31, 2018 85% of all call screeners report using the tool always or almost always to inform their recommendation to screen in or out a GPS referral for assessment. Benchmarks to keep stakeholders informed might also be considered to ensure transparency does in fact occur, e.g., issue quarterly newsletters to external stakeholders to keep them informed on progress.

APPENDIX A: INITIAL SURVEY RESULTS

Demographics

	WORK STATUS		
CHARACTERISTICS OF CALL SCREENERS PARTICIPATING IN SURVEY	FULL TIME	PART TIME	UNKNOWN
Number of Call Screeners Surveyed by Work Status	12	4	0
Average Years Worked as a Call Screener by Work Status	8	5	0
Average Years Worked for Allegheny County by Work Status	14	7	0

Training

DID YOU RECEIVE TRAINING PRIOR TO USING THE ALLEGHENY FAMILY SCREENING TOOL?	#
Yes	14
No	0
No Answer	2
Total	16

HOW WELL DID THE TRAINING PREPARE YOU TO USE THE TOOL?	#
Not at all	1
Limitedly	1
No opinion	3
Somewhat	5
Very well	6
Total	16

HOW WELL DID THE TRAINING INCREASE YOUR UNDERSTANDING ABOUT HOW THE TOOL WORKS?	#
Not at all	1
Limitedly	1
No opinion	3
Somewhat	5
Very well	6
Total	16

Appendix A (continued)

WHICH PART OF THE TRAINING DID YOU FIND TO BE MOST HELPFUL?	#
No answer	2
Overview of Predictive Analytics/Predictive Risk Modelling	3
Application of Predictive Analytics in Allegheny County	5
Policy/Practice	1
Case scenarios	4
Process changes	0
KIDS Changes/Design	1
Total	16

Screening Tool Function and Visualization

HOW EASY/DIFFICULT IS IT FOR YOU TO NAVIGATE OR USE THE PRM TOOL?	#
Very difficult	0
Difficult	0
Neutral	5
Easy	4
Very easy	7
Total	16

HOW HELPFUL IS THE "THERMOMETER" VISUALIZATION?	#
Not helpful at all	2
Limitedly helpful	1
No opinion	6
Somewhat helpful	4
Very helpful	3
Total	16

Appendix A (continued)

Decision Making

HOW FREQUENTLY DO YOU USE THE PRM TOOL TO INFORM YOUR RECOMMENDATION (EXCLUDING MANDATORY REFERRALS)?	#
Never	1
Almost never	4
Occasionally/sometimes	4
Almost every time	4
Every time	3
Total	16

HOW HELPFUL IS THE PRM TOOL TO INFORM YOUR RECOMMENDATION?	
Not at all helpful	1
Limitedly helpful	2
Neutral/No opinion	6
Somewhat helpful	7
Very helpful	0
Total	16

HOW OFTEN DO YOU CONDUCT A SEARCH IN CLIENTVIEW AFTER VIEWING THE DYNAMIC TEXT ("MAD LIBS")?	#
Never	1
Almost never	3
Occasionally/sometimes	4
Almost every time	4
Every time	4
Total	16

WHAT IMPACT DO YOU THINK THE ALLEGHENY FAMILY SCREENING TOOL WILL HAVE ON THE CALL SCREEN PRACTICE?	
Strengthen the practice	7
Diminish the practice	1
No impact on the practice	8
Total	16

APPENDIX B: FOLLOW-UP SURVEY RESULTS

Demographics

	WORK STATUS		
CHARACTERISTICS OF CALL SCREENERS PARTICIPATING IN SURVEY	FULL TIME	PART TIME	UNKNOWN
Number of Call Screeners Surveyed by Work Status	11	6	1
Average Years Worked as a Call Screener by Work Status	9	6	1
Average Years Worked for Allegheny County by Work Status	15	9	3

Experience and Attitudes with Using the Allegheny Family Screening Tool

HOW EASY/DIFFICULT IS IT FOR YOU TO NAVIGATE OR USE THE ALLEGHENY FAMILY SCREENING TOOL?	#
Very Difficult	0
Difficult	1
Neutral	6
Easy	4
Very Easy	7
Total	18

HOW OFTEN HAVE YOU HAD A SCORE THAT SEEMS INACCURATE GIVEN THE FAMILY HISTORY YOU HAVE AVAILABLE OR COLLECTED DURING YOUR REVIEW OF THE CALL?	#
A great deal	0
A moderate amount	2
Occasionally	13
Rarely	3
Never	0
Total	18

Appendix B (continued)

AGREEMENT WITH STATEMENTS	ALWAYS	MOSTLY	SOMETIMES	RARELY	NEVER	TOTAL
The score is clearly displayed.	7	8	3	0	0	18
I go back and conduct a more thorough search in KIDS when the score is high.	2	3	2	5	6	18
I go back and conduct a more thorough search in ClientView when the score is high.	3	1	3	6	5	18
How frequently do you use the Tool to inform your recommendation (excluding mandatory referrals)?	2	5	6	4	1	18

AGREEMENT WITH STATEMENTS	NO ANSWER	STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE	TOTAL
I am confident of the Tool's ability to accurately assess risk of future placement or re-referral.	0	1	8	7	2	18
I understand what the score is predicting.	0	0	1	9	8	18
The use of this tool is changing the culture of our workplace to be more data-driven.	0	1	6	9	2	18
I understand how the score should relate to or inform screening decisions.	0	0	1	13	4	18
I understand the content of the data sources helping to produce the score.	0	0	2	10	6	18
I am adequately prepared to use the Tool.	1	0	0	11	6	18
I am confident in the research that went into the development of this tool.	0	1	6	9	2	18

WHAT IMPACT DO YOU THINK THE ALLEGHENY FAMILY SCREENING TOOL WILL HAVE ON THE CALL SCREEN PRACTICE?	#
Strengthen the practice	8
Diminish the practice	2
No impact on the practice	8
Total	18

SECTION 5

Impact Evaluation Summary of the Allegheny Family Screening Tool

by the Allegheny County Department of Human Services

SUMMARY

The Allegheny Family Screening Tool (AFST) is a predictive risk model built and trained using County administrative child protection and service records.[1] Allegheny County implemented the AFST in 2016 as a decision-support tool, with the goal of improving both the accuracy and consistency of decisions made about referrals to the child maltreatment hotline.

A request for proposals was issued in December 2015 and in early 2016, the Allegheny County Department of Human Services (DHS) issued a competitive contract to Stanford University (principal investigator: Goldhaber-Fiebert) to design and conduct an independent evaluation of the impact of the AFST (along with associated policy changes) on the County's child maltreatment screening decisions.

The evaluation looks at Version 1 of the AFST and consists primarily of outcome comparisons for two groups of children: (1) the approximately 31,000 children who were referred for alleged maltreatment during the 18-month period before the AFST was implemented (January 1, 2015 through July 31, 2016, called "Pre-AFST [late]" in this report) and (2) the approximately 34,000 children referred after the AFST was fully implemented ("Post-AFST": December 1, 2016 through May 31, 2018). This report provides a summary of the findings; to read the full technical report, please see: Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office. Two peer reviewers provided critical feedback on earlier drafts of the evaluation report.

Evaluation findings are detailed in the sections that follow, and emerge from a set of methodologically strong, quasi-experimental methods (i.e., interrupted time series analyses, generalized linear models). Quasi-experimental methods refer to a type of evaluation approach used when it is not possible or desirable to implement a randomized controlled trial (RCT).

1 Vaithianathan R, Putnam-Hornstein E, Jiang N, Nand P, & Maloney T. (2017). Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County methodology and Implementation. Centre for Social Data Analytics: https://www.alleghenycounty.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=6442457403

While less robust than a gold-standard RCT, carefully designed quasi-experimental methods are considered the next-best approach to testing program impact. The County decided not to pursue an RCT primarily for practical reasons.

Key findings of the impact evaluation include:

1. **Overall, the AFST did not lead to increases in the rate of referred children screened-in for investigation.** Use of the tool appears to have resulted in a different pool of children screened-in for investigation (including more children who were deemed in need of child welfare intervention or supports, see below). But from the perspective of investigative workload, there was not a substantial increase in the number or proportion of children investigated among all children referred for maltreatment.
2. **Implementation of the AFST increased the identification of children determined to be in need of further child welfare intervention.** Use of the tool led to an increase in the screening-in of children who were subsequently determined to need further intervention or supports. Specifically, there was a statistically significant increase in the proportion of children screened-in whose child welfare case was then opened or, if no case was opened, were re-referred within 60 days. (Please note that investigators and supervisors making these case opening decisions remained blind to the score so this result reflects real change in the case-mix of families screened-in.)
3. **Use of the AFST did not lead to decreases in re-referral rates for children screened-out without investigation.** Re-referral rates among children screened-out stayed the same for children overall, with the exception of children who were 4-6 years of age. This was the age group directly affected by County changes to mandatory field policy screening protocols by age. Unfortunately, for this age group there was a slight but statistically significant increase in the likelihood of the Post-AFST group of children being re-referred.
4. **The AFST led to reductions in disparities of case opening rates between black and white children.** Prior to the introduction of the AFST, case-opening rates for black children were higher than for white children. During the Post-AFST period, increases in the rate of white **children determined to be in need of further child welfare intervention**, coupled with slight declines in the rate at which black children were screened-in for investigation, led to reductions in racial disparities. Specifically, there was an increase in the number of white children who had cases opened for services, reducing case disparities between black and white children.
5. **There was no evidence that the AFST resulted in greater screening consistency within individual call-screeners.** Specifically, for the subgroup of 11 call screeners who handled a substantial volume of both Pre-AFST and Post-AFST referrals, attempts were made to assess whether the AFST led to more "between-screener" consistency. Likewise, changes

in screening consistency by referred child's age group and racial group were also assessed. No impact was detected, although it should be noted that there was likely insufficient power to identify anything other than very large shifts.

METHODOLOGY

Implementation of the AFST

With the implementation of the AFST, call screeners in Allegheny County are now presented with a single Family Screening Score. The score is a standardized summary of available data, providing additional information to aid the call screener (and their supervisor) to make decisions regarding further investigation. Screening recommendations are made on any call which is classified as "general protective service" (GPS)[2] for all individuals currently residing in the same household as the alleged victim (alleged child/victim, biological mother and father of alleged victim, the perpetrator, other related and unrelated children in the home, and other adults in the home). Investigation recommendations are made by the County hotline staff (screeners and supervisors) and follow one of three courses: 1) Screen-out of a referral without any further evaluation or assessment, 2) Field screen of the referral to assess whether an investigation is warranted, or 3) Screen-in of a referral, which is synonymous with conducting a formal investigation. When a field screen (a home visit to assess the safety of the child[ren] and determine whether a formal investigation is warranted) is conducted, it is always followed by a decision to either screen-out or screen-in the referral.

At the time of referral, a re-referral and placement risk score is calculated for each child associated with the referral. The AFST, which is the only score the screeners see, is based on the maximum score (either re-referral or placement) across all children associated with the referral at the time of the screening call. The score ranges from 1 to 20 (where 20 is the highest "risk" and 1 is the lowest), indicating the ventile into which the AFST falls. A recommendation for "auto screen-in" occurs when the AFST falls above 18 for the placement score.[3]

Accompanying Protocol Changes with AFST Implementation

Several other systematic changes to the maltreatment referral screening process accompanied the full implementation of the AFST.

- **Field Screening.** First, the County's mandatory field screen policy was updated. Previously, households with at least one child under the age of 7 were required to be field screened, regardless of the family's history. With the implementation of the AFST, the maximum age for a mandatory field screen decreased from 7 to 4 years of age. In addition, the new mandatory field screen policy added the following three conditions: (1) all children who attend homeschool/cyber school receive a mandatory field screen regardless of age; (2) any family that has had 4 or more referrals in 2 years without any of the referrals being formally investigated receive a mandatory field screen; and (3) any other referrals where more information is necessary to make a final decision receive a field screen.[4]

2 Allegations fall under the state of Pennsylvania's Child Protective Service (CPS) statutes (23 Pa.C.S. § 6303) or General Protective Service (GPS) statutes (23 Pa.C.S. § 6334). CPS referrals include those made for child abuse, including physical and sexual abuse. CPS referrals must be investigated and require more urgent response times, often overlap with law enforcement and medical investigations, and lead to a determination of whether abuse occurred (that may result in perpetrators being registered in the state's ChildLine registry). GPS referrals include referrals made when there is a risk of harm. For example, neglect, truancy and substance use by parents would all fall under GPS referrals. GPS referrals may be investigated or screened out without further assessment, at the discretion of call screening staff. GPS investigations assess for risk and safety to ensure well-being of children and provide families with any supports they may need. GPS investigations cannot result in registry with the state's ChildLine registry. Both CPS and GPS referrals can result in a family having a case opened at the end of an investigation for ongoing services and supports. In 2017, 21 percent of DHS referrals were CPS referrals and 79 percent were GPS referrals.

3 Of note, a referral flagged for auto screen-in does not obligate the call-screener and/or supervisor to screen-in, but rather implies that this would be the default action. The concurrence rate for referrals marked as auto screen-in for the dates December 1, 2016–November 29, 2018 was 61 percent.

- **Screener Supervision.** Second, call screeners now make a recommendation about the decision to screen-in/out to his/her supervisor who has the responsibility for the ultimate decision. Prior to this set of policy and practice changes, the primary role of the call screening staff was to gather information to inform supervisor decision-making. Call screeners collected data from several databases and resources, including internal DHS systems (e.g., KIDS, Client View), courts, public assistance and criminal justice. Call screeners also spoke with the individual making the report and other key contacts (e.g., schools, doctors). The information collected was given to supervisors for final decision-making. Although the process between screeners and supervisors was collaborative, following implementation of the AFST, call screeners took on a greater role in making recommendations for screening decisions.

Evaluation Outcomes

The evaluation team defined three main outcomes to measure underlying effects of the AFST implementation on the County's maltreatment referral screening decisions. The choice of outcomes reflected the emphasis the County placed on evidence of changes to screening accuracy, as well as potential impacts on screening decisions by race/ethnicity. The main outcomes investigated were:

- **Overall rates of children screened-in for investigation**
  This outcome is intended to measure how the implementation of a predictive risk model impacted the flow of children referred for alleged maltreatment into investigations, with potential implications for workload and the system overall.

  A child is considered screened-in for investigation if the referral (i.e., household) that includes the child is advanced by the hotline screener and their supervisor for further investigation. Therefore, the rate of "screened-in for investigation" was defined as equal to the total number of children in referrals assigned to further investigation (numerator) divided by the total number of children in referrals (denominator), computed for referrals falling in each calendar month and for children in different age and racial/ethnic groups.

- **Likelihood a screened-out child had no re-referrals within 2 months**
  This outcome is intended to measure how AFST implementation impacted one feature of accuracy: do children who are screened-out appear in subsequent referrals? The assumption is that the absence of a near-term follow-up referral indicates there were no safety and/or well-being issues. A re-referral is assumed to indicate that there was a missed opportunity on the part of the County to have intervened with services earlier.

  A child is considered screened-out if the maltreatment referral that includes the child is not advanced by the hotline screener and supervisor for an in-person investigation. If no additional referral is made within 2 months of the index referral event, then the child is considered to have been screened out without a re-referral. It should be noted that a child can have more than one screen-out and re-referral over time, but only subsequent referrals within 2 months of a specific "index event" were examined. The rate of screen-outs with no

4 An average of a 1000 GPS referrals per month have been processed since the AFST was implemented and data about reasons for field screening started being collected. During this time period, about 10% of referrals were assigned to the child welfare unit that handles field screening. Of these field screened referrals, 82% indicated a child under 4 on the referrals as the reason and 16% expressed the discretionary desire for more information. The other rules—regarding home schooled children or 4+ recent referrals screened out—were used very rarely.

re-referral within a 2-month time window was defined as the number of children in referrals that were not advanced for further investigation and were not re-referred within 2 months (numerator), divided by the total number of children in referrals that were screened-out without investigation (denominator), computed for referrals falling in each calendar month and for children in different age and racial/ethnic groups. These analyses were repeated using a 6-month re-referral window as a robustness check.

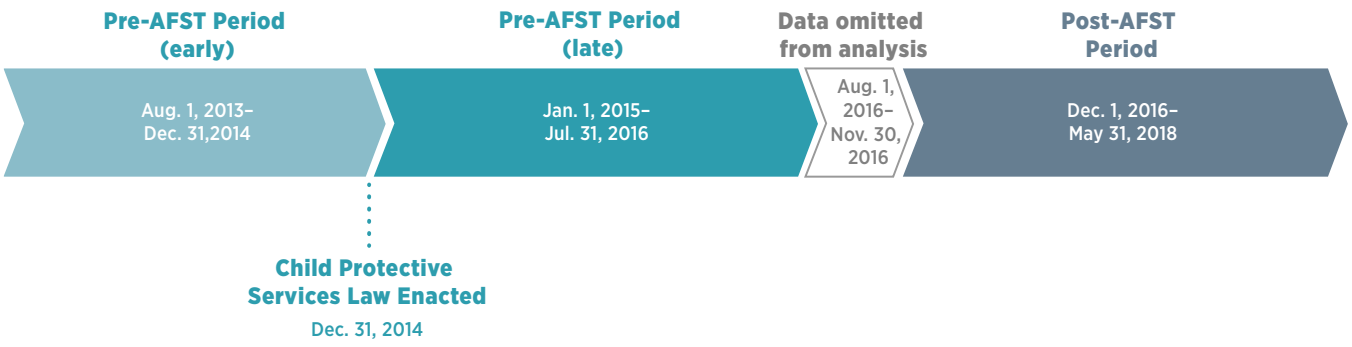
- **Likelihood a screened-in child had a case opened for services upon investigation, or had a re-referral within 2 months if no case was opened**
  This outcome is intended to measure how AFST implementation impacted one feature of accuracy: do children who are screened-in for investigation evidence safety and service needs requiring child protective services (i.e., case opened initially, or a re-referral if not)?

  A child is considered to have experienced this third outcome if a referral that includes the child is screened-in (i.e., advanced by the hotline screener and supervisor for investigation) and upon investigation, one of two things happens: (1) a child protective service case is opened by the investigating worker (indicating that safety concerns and service needs were identified); or (2) a child protective service case is not opened by the investigating worker and the child is re-referred within 2 months of the original referral (indicating safety concerns and service needs were identified by the hotline screener, but may have been improperly addressed by the investigating worker). The rate of screen-ins requiring services was defined as the total number of screened-in children meeting criteria 1 or 2 above (numerator), divided by the total number of children screened in for investigation (denominator), computed for referrals falling in each calendar month and for children in different age and racial/ethnic groups.

  It should be noted that case openings were deemed to provide a good measure for the purposes of evaluation because the investigator does not see the AFST score. As such, the investigating worker's decision is made independent of the score.

Evaluation Window

The entire evaluation window spans August 1, 2013 through May 31, 2018. For the purpose of the analyses, the data were divided into multiple periods:



- **Pre-AFST Period.** The "Pre-AFST Period" spans August 1, 2013 through July 31, 2016 and was divided in two parts:
  - The Pre-AFST Period (early) spans August 1, 2013 through December 31, 2014
  - The Pre-AFST Period (late) spans January 1, 2015 through July 31, 2016

  The evaluator's decision to divide the Pre-AFST Period is based on a set of amendments to the State of Pennsylvania's existing Child Protective Services Law, which became effective on December 31, 2014 and had the effect of altering a number of features of referrals to the call center.[5] The second (late) Pre-AFST Period served as the point of most comparison in the analysis.

- **Post-AFST Period.** The period after the full implementation period is termed the "Post-AFST Period" and spans December 1, 2016 through May 31, 2018. Outcomes during the Post-AFST Period are compared to outcomes in periods prior to this.

Notably, data for the period between August 1, 2016 and November 30, 2016 are omitted from all analyses. When the AFST was launched, an initial policy decision sought to restrict score generation to only individuals and families who could be substantively identified in prior county data (preventing scores from being displayed that were solely constructed from basic referral/geographic information when the family was otherwise unknown to DHS). This policy initially restricted scores to situations where a child on the call was positively identified with a prior county identifier (meaning that the child was previously known to DHS and had been assigned a unique identifying client number). Many children, most notably newborns and infants, who experience the highest rates of maltreatment and fatalities, often do not have system involvement, and therefore do not have their own county identifier, but their parents or caregivers may have significant current and prior system involvement. The initial design led to situations where known information about adults on the referral could not be used in generating a score if none of the children were recognized, and this was quickly deemed too restrictive. After November 30, 2016, scores could be generated for a referral if any individual named, child or adult, could be matched to a county identifier. Given the non-random nature of the children who did not receive AFST scores prior to December 1, 2016, analyses of this data cannot reliably attribute observed changes in outcomes pre- and post-implementation to the AFST score.

Data

All analyses use de-identified (anonymized) data relating to individuals named in maltreatment referrals made to Allegheny County's child protective services hotline. The data consist of information about individual household members including their race, legal sex and age. Additionally, the data identify the call screeners and supervisors associated with the referral and track previous referrals and investigations with child welfare and other child-serving systems from August 1, 2013 through May 31, 2018.

5 Child Protective Services Act, P.L.1240, No.206, 23 PA §§6301-6386 (2015).

The analytic dataset focuses on outcomes (described above) for children below 18 years at the time of referral.

The analytic dataset also contains several variables used in the analysis to control for child demographics (age, legal sex, race) and household characteristics (household counts and composition, socioeconomic status and maximum risk scores). For child's race, the evaluation used a categorical variable which included the category "Unable to Determine", when race was not coded as white, Black/African American, or other. Other control variables had complete data.

Analytic Approach

Three main types of analyses are reported for each of the three main outcomes. Comparisons of unadjusted population means and the Interrupted Time Series Analysis (ITSA) describe levels and changes in outcomes within the County's child maltreatment screening system overall, and for age and race/ethnic subgroups given existing trends in the Pre-AFST period (late) (January 1, 2015 through July 31, 2016) and the Post-AFST period (December 1, 2016 through May 31, 2018). Individual-level multivariate regression analyses focus on changes after adjustment for changes in referral case mix over time. The evaluation team also considered the AFST's effects on outcomes for subgroups of children defined in terms of their age group and racial/ethnic characterization. This enabled potential heterogeneity and disparities in the policy's effects across subgroups.

- **Unadjusted Population Means.** The simplest comparison performed was a comparison of unadjusted means for the Pre- and the Post-AFST periods, testing whether they are statistically different from one another using a two-sided t-test of equality of means.
- **Interrupted Time Series Analysis.** Changes in the level and trend of monthly rates of each outcome during the Pre- and Post- periods were assessed using an Interrupted Time Series Analysis. In this evaluation, the ITSA measures changes in both the level and slope of each outcome in the Post-AFST months in relation to the Pre-AFST months. The ITSA approach captures population-level changes in outcomes and trends after a policy change in comparison to the levels and trends prior to that change.
- **Child-Level Multivariate Regression Analysis.** Finally, the evaluators used multivariate individual-level regression analyses to assess the impact of the AFST on the predicted level of each outcome Pre- and Post-AFST, while adjusting for child and household characteristics. These analyses focus on estimates of the average effect of the AFST, adjusting for evolving case mix over time. The predictive margins presented in evaluation tables and figures can be interpreted as the average outcome if all children in the sample were in either the Pre-AFST or the Post-AFST time-frame, holding all other control variables constant.

RESULT HIGHLIGHTS

Overall rates of children screened-in for investigation

All children screened-in

- Prior to the implementation of the AFST, the number of maltreatment referrals had increased during the Pre-AFST period. This increase in total referrals corresponded with state law and policy changes that expanded mandatory reporting. With an increase in the number of referrals received, the fraction of all referrals screened-in for investigation began to decline. The AFST largely halted this decline in screened-in investigations for all groups. Even though the average screen-in level in the Pre-AFST period was higher than in the Post-AFST period, it is unknown whether screen-in rates would have continued to decline.

Children screened-in, by age group

- Prior to the implementation of the AFST, the fraction of referrals screened-in for investigation was declining for children in all subgroups except for ages 4 to 6, with larger declines observed in the oldest age group (13 to 17 years). The AFST largely halted these age-specific declines, most noticeably for children ages 7 years and older.

Children screened-in, by race

- Prior to the implementation of the AFST, the fraction of referrals screened-in for investigation were declining for children in all race groups, with larger declines observed for Black/African American children than for white children.
- The AFST largely halted all race-specific declines, most noticeably for Black/African American children.

Call screener consistency

- There was moderate consistency in the referral screen-in outcome across call screeners.
- Screen-in rates increased in the Post-AFST period for 7 of the 11 call screeners (4 of these were significant increases) and decreased for 4 of the 11 call screeners (none statistically significantly).
- The variance of call screener outcomes decreased for children in both Black/African American and white race groups, with a larger effect apparent in the Black/African American group (though not statistically significantly).

Likelihood a screened-out child had no re-referrals within 2 months

Re-referrals of all children

- In breaking down the AFST's effect on decisions to screen out children without investigation, there was a small increase in the overall rate at which screened-out children were re-referred.

Re-referrals by age group

- The increase in re-referral rate was concentrated among children in the 4-to-6-year-old age group. Among all other age groups, reductions in the likelihood of being re-referred after being screened out were non-significant.
- The observed increase in re-referral rates among 4-to-6-year-olds after the implementation of the AFST is likely due to corresponding changes in the County's policy regarding the maximum age for mandatory field screening. With implementation of the AFST, the County reduced the age for mandatory field screening from under 7 years of age to under 4 years of age. It may be that previous field screenings in this age group helped to identify more children for whom a screen-in for investigation was appropriate.

Re-referrals by race group

- Multiple analyses showed small increases in re-referral rates for both race subgroups, which were not significant for white children and only occasionally significant for Black/African American children.

Re-referrals by call screeners

- The absence of changes in re-referral rates for children screened-out was consistent across call-screeners, and variation between calls screeners in this outcome did not change significantly. The variance of call-screener-specific outcomes increased slightly (not statistically significantly) in the Post-AFST period compared to the Pre-AFST period.
- Results were similar when evaluators used a re-referral window of 6 months instead of two months.

Likelihood a screened-in child had a case opened for services upon investigation, or had a re-referral within 2 months if no case was opened

- The AFST increased the identification of higher-need children (measured as those children determined to be in need of further child welfare intervention, i.e., those who, after being screened-in, had cases opened for child protective services or, if no case was opened, had a re-referral within 2 months). It should again be noted that the investigating worker and supervisor, those making the decision to open a case, did not have access to the score.
- Increase in the identification of children determined to be in need of further child welfare intervention emerged across age and racial subgroups.
- While changes to screening-in higher-need children remained throughout the Post-AFST period, the initial improvement effect did attenuate somewhat over time.
- With a re-referral window of 6 months, the direction of the result was the same and there was somewhat less attenuation over time.
- The AFST had an immediate upward effect on the likelihood of screening-in children later determined to be in need of further child welfare intervention for both white and Black/African American children.

- For Black/African American children, the initial improvement effect of implementing the AFST attenuated over time, whereas for white children the effect was more persistent.
- Whereas prior to the AFST, Black/African American children had a higher rate of case openings after screen-in than white children, this disparity was reduced over time in the post-AFST period.
- The overall change in screen-ins for children determined to be in need of further child welfare intervention was consistent across call screeners, and variation between calls screeners in this outcome did not change significantly.

CONCLUSIONS

Overall, analyses documented that the AFST and associated policies:

- increased the accuracy of decisions about children screened-in for investigation and
- did not increase the number of children screened-in for investigation (as compared to the average during the pre-AFST period).

Among children screened-out without an investigation, there was a slight increase in the re-referral rates for children between 4 and 6 years of age, the group of children directly affected by corresponding policy changes to mandatory in-home assessments (i.e., field screening). The County will look at this finding carefully and will work with call screening staff to understand why some of the children in this age group who had a high AFST score, and who were later re-referred and screened-in, were screened-out at this initial referral.

One of the key topics addressed by the evaluation was the effect of the implementation of the AFST and surrounding policy changes on disparities in outcomes across race/ethnicity. Overall, changes in a type of "accuracy" measure (i.e., an increase in accuracy for children screened-in for investigation and a negligible or slight decrease in accuracy for children screened out) were consistently observed for both Black/African American and white children.

It should also be noted, however, that if community referrals are biased by race, then even appropriate screen-outs for Black/African American children might look "inappropriate" because they get re-referred. Therefore, we need to treat the racial differences in this outcome measure with caution. The fact that two-thirds of children who were defined as inaccurately screened-out in our analysis are Black/African American might be suggestive that broader (and more objective) definitions should be considered.

The AFST was hypothesized to result in greater screener consistency but the evaluation detected no such improvements. It's worth noting that the improvements would have to have been relatively large to be detected; however, there are also other possible explanations for the lack of improved consistency here. First, with the implementation of the AFST, the County enhanced the autonomy of the call screener who now makes a recommendation to their supervisor. Prior to the implementation of the AFST, decisions were collaborative in nature

but were ultimately made by the supervisor. This practice change enhanced the opportunity for variability in decision making, perhaps reducing any improvements that the AFST might have shown. Further, there is considerable lack of concurrence with the AFST by call screeners, limiting the ability for a tool like the AFST to have effect on consistency. For the period December 1, 2016–November 29, 2018, only 61 percent of the referrals that scored in the "mandatory" screen-in range were in fact screened in (Table 1). Therefore, the County will continue to try to work with call screeners to understand why they might be making these decisions. We hope to have more data and therefore more power to measure screener variation in the next stage of the evaluation.

TABLE 1: **Percent of children screened-in for investigation, by AFST risk level**

	PERCENT SCREENED-IN FOR INVESTIGATION
Mandatory	61%
High	47%
Medium	42%
Low	31%
No Score	23%
Total	41.4%

The evaluation team concluded that "the effects of implementing the AFST and surrounding policy changes show moderate improvements in accuracy of screen-ins with small decreases in the accuracy in screen-outs, a halt in the downward trend in pre-implementation screen-ins for investigation, no large or consistent differences across race/ethnic or age-specific subgroups in these outcomes, and no large or substantial differences in consistency across call-screeners."

The County is encouraged that the AFST has shown some effect on the accuracy of decision making and reductions in overall case opening disparities between black and white children, particularly in the face of implementation challenges (for a discussion of technical, practice, and policy challenges please see the FAQ). More importantly, there was no evidence of unintended adverse effects.

The evaluation aligns with what the leadership of the County have observed: that the tool has tremendous potential and that there are few, if any, unintended adverse effects given workers' willingness to use their own discretion in the screening decision. But implementation challenges were significant and persist; these must be overcome to maximize the impact of automated risk stratification tools.

We will continue to work to improve the model and its implementation. As of November 2018, the County released a model with significant enhancements (for more information on this, see Methodology, Version 2 and FAQ). We will also continue the evaluation and will ask Stanford University to consider streamlining their methods and examine:

- how previously-defined outcomes (defined in this evaluation summary) changed with the implementation of Version 2 of the tool, stratified by race/ethnicity, age, and AFST score.
- the impact of the AFST and associated policies on home removals.
- consistency of outcomes across supervisors (instead or in addition to examining call screeners, given decision making processes).
- the impact of the high- and low-risk protocols on decision making.

SECTION 6

Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office

Jeremy D. Goldhaber-Fiebert, PhD and Lea Prince, PhD, Stanford University
March 20, 2019

EXECUTIVE SUMMARY

Goal: The impact evaluation assesses how implementation of the screening score (Allegheny Family Screening Tool or AFST) within Allegheny County's child welfare office helped to:

- Improve accuracy of referrals by call screeners (increasing the fraction of children who screen-in with further action taken upon investigation and the fraction of children who screen-out with no re-referrals within 2 months)
- Maintain reasonable workload in terms of the rate of screen-ins (and subsequent investigations)
- Reduce disparities in terms of the above outcomes for similar children from age and race/ethnic subgroups
- Promote consistency in terms of the above outcomes across the call screeners

Approach: The evaluation uses a set of methodologically strong, quasi-experimental techniques (e.g., interrupted time series analyses, generalized linear models) to achieve the impact evaluation's goals. The study primarily compares outcomes for children involved in general protective service (GPS)[1] referrals in the 15 - 17 months after the full implementation of the AFST (December 1, 2016 through May 31, 2018) (~34,000 children) to outcomes for children involved in GPS referrals in the period before implementation, primarily January 1, 2015 through July 31, 2016 (~31,000 children). Further details of the approach appear in the **Methods** sections below.

Findings and Interpretation:

1. *Accuracy:* Implementation of the AFST and associated policies increased accuracy for children screened-in for investigation and may have slightly decreased accuracy for children screened-out.
    - Implementation of the AFST increased the proportion of children who screened-in for investigation and upon investigation either had further action taken or else were re-referred within 60 days. The larger initial effect appeared to partially attenuate over time.

1 Allegations fall under the state of Pennsylvania's Child Protective Service (CPS) statutes (23 Pa.C.S. § 6303) or General Protective Service (GPS) statutes (23 Pa.C.S. § 6334). CPS referrals include those made for child abuse, including physical and sexual abuse. CPS referrals must be investigated and require more urgent response times, often overlap with law enforcement and medical investigations, and lead to a determination of whether abuse occurred (that may result in perpetrators being registered in the state's ChildLine registry). GPS referrals include referrals made when there is a risk of harm. For example, neglect, truancy and substance use by parents would all fall under GPS referrals. GPS referrals may be investigated or screened out without further assessment, at the discretion of call screening staff. GPS investigations assess for risk and safety to ensure well-being of children and provide families with any supports they may need. GPS investigations cannot result in registry with the state's ChildLine registry. Both CPS and GPS referrals can result in a family having a case opened at the end of an investigation for ongoing services and supports. In 2017, 21 percent of DHS referrals were CPS referrals and 79 percent were GPS referrals.

Implementation of the AFST and associated policies may have slightly decreased accuracy in terms of the proportion of children screened-out who had no re-referrals within 60 days, with the majority of this small effect in children aged 4 to 6 years.

2. *Workload:* Implementation of the AFST and associated policies halted the downward trend in the rate of children screened-in for investigation.
3. *Disparities:*
    - *Accuracy:*
        - Further action taken or re-referral within 60 days after being screened-in for investigation: There were larger increases in accuracy of being screened-in and/or less attenuation of the effect over time for white children and children aged < 4 years. In contrast, the initial improvement in accuracy attenuated more rapidly for Black/African American children.
        - Re-referral within 60 days after being screened-out: There were greater losses of accuracy of being screened-out for children ages 4 to 6 years though the overall size of effect even in this age-group was relatively small. This may be due to concurrent changes in the mandatory in-home assessment (field screening) policy in terms of the maximum age being reduced from under 7 to under 4 years of age.
    - *Workload:* The effect of the AFST and associated policies of halting the downward trend in the rate of children screened-in for investigation was larger for older children (e.g., ages 13 to 17) and for Black/African American children. The effect was smallest for children ages 4 to 6 years where a screen-in for investigation may have replaced field screening which was no longer required for this age-group.
4. *Consistency:* For the subgroup of 11 call screeners handling a substantial volume of referrals in both the Pre-AFST and Post-AFST implementation periods, the AFST and associated policies did not significantly alter the consistency of outcomes relating to accuracy or workload across call screeners. Likewise, the AFST did not significantly alter age group-specific or race group-specific consistency for any of these outcomes. Of note, particularly for call screener consistency outcomes by age-group or race group, there was likely insufficient sample size (power) to detect changes.

CONTENTS

Methods *10*

- Overview *10*
- Implementation of the AFST *10*
- Other Changes Occurring with the Implementation of the AFST *10*
  - Outcomes *11*
    - Accuracy Outcomes *11*
    - Workload Outcomes *12*
    - Consistency Outcomes *13*
    - Disparities Outcomes *13*
- Hypothesized Effects of the Implementation of the AFST Score and Related Policies *14*

Data *15*

Analytic Approach *15*

- Overview *15*
- Description of Specific Methods *16*
- Covariates and Standard Errors *18*
- Description of Trends in the Study Population and Changes to Case Mix over Time *19*

Results *19*

- Accuracy Outcomes and Disparities in Accuracy Outcomes *19*
  - How did the AFST change accuracy for referrals screened-in? *19*
  - How did the AFST change accuracy for referrals screened-in related to children in different age-groups? *20*
  - How did the AFST change accuracy for referrals screened-in related to children in different race groups? *20*
  - How did the AFST change accuracy for referrals screened-out? *20*
  - How did the AFST change accuracy for referrals screened-out related to children in different age-groups? *21*
  - How did the AFST change accuracy for referrals screened-out related to children in different race groups? *21*

Workload Outcomes and Disparities in Workload Outcomes *21`*

How did the AFST change workload as measured by the fraction of referrals screened-in for further investigation? *21*

How did the AFST change workload related to children in different age-groups? *22*

How did the AFST change workload related to children in different race groups? *22*

Consistency Outcomes and Disparities in Consistency Outcomes *22*

How did the AFST change the consistency of accuracy for referrals screened-in by call screener? *22*

How did the AFST change the consistency of accuracy for referrals screened-in by call screener by age-groups? *23*

How did the AFST change the consistency of accuracy for referrals screened-in by call screener by race groups? *23*

How did the AFST change the consistency of accuracy for referrals screened-out by call screener? *23*

How did the AFST change the consistency of accuracy for referrals screened-out by call screener by age-groups? *23*

How did the AFST change the consistency of accuracy for referrals screened-out by call screener by race groups? *24*

How did the AFST change workload differentially by call screener? *24*

How did the AFST change workload differentially by call screener by age-groups? *24*

How did the AFST change workload differentially by call screener by race groups? *25*

Discussion, Conclusions, and Implications *25*

TABLES *28*

TABLE 1A: Summary Statistics, child characteristics *28*

TABLE 1B: Summary Statistics, household characteristics *29*

TABLE 2: Means of outcomes *30*

TABLE 3A: Accuracy of screen-in, ITSA analysis, all children *31*

TABLE 3B: Accuracy of screen-in, ITSA analysis, < 4 years old *31*

TABLE 3C: Accuracy of screen-in, ITSA analysis, 4 to 6 years old *32*

TABLE 3D: Accuracy of screen-in, ITSA analysis, 7 to 12 years old *32*

TABLE 3E: Accuracy of screen-in, ITSA analysis, 13 to 17 years old *33*

TABLE 3F: Accuracy of screen-in, ITSA analysis, White *33*

TABLE 3G: Accuracy of screen-in, ITSA analysis, Black/African American *34*

TABLE 4A: Accuracy of screen-in, adjusted analysis, all children *34*

TABLE 4B: Accuracy of screen-in, adjusted analysis, by age group *34*

TABLE 4C: Accuracy of screen-in, adjusted analysis, by race *35*

TABLE 5A: Accuracy of screen-out, ITSA analysis, all children *35*

TABLE 5B: Accuracy of screen-out, ITSA analysis, < 4 years old *36*

TABLE 5C: Accuracy of screen-out, ITSA analysis, 4 to 6 years old *36*

TABLE 5D: Accuracy of screen-out, ITSA analysis, 7 to 12 years old *37*

TABLE 5E: Accuracy of screen-out, ITSA analysis, 13 to 17 years old *37*

TABLE 5F: Accuracy of screen-out, ITSA analysis, White *38*

TABLE 5G: Accuracy of screen-out, ITSA analysis, Black/African American *38*

TABLE 6A: Accuracy of screen-out, adjusted analysis, all children *39*

TABLE 6B: Accuracy of screen-out, adjusted analysis, by age group *39*

TABLE 6C: Accuracy of screen-out, adjusted analysis, by race *40*

TABLE 7A: Workload, ITSA analysis, all children *40*

TABLE 7B: Workload, ITSA analysis, < 4 years old *41*

TABLE 7C: Workload, ITSA analysis, 4 to 6 years old *41*

TABLE 7D: Workload, ITSA analysis, 7 to 12 years old *42*

TABLE 7E: Workload, ITSA analysis, 13 to 17 years old *42*

TABLE 7F: Workload, ITSA analysis, White *43*

TABLE 7G: Workload, ITSA analysis, Black/African American *43*

TABLE 8A: Workload, adjusted analysis, all children *43*

TABLE 8B: Workload, adjusted analysis, by age-group *44*

TABLE 8C: Workload, adjusted analysis, by race *44*

TABLE 9: Means of Outcomes (1) - (3) for call screeners included and excluded from Outcome 4/Consistency analyses *45*

TABLE 10A: Consistency in accuracy of screen-in, adjusted analysis, for 11 included call screeners *45*

TABLE 10B: Consistency in accuracy of screen-in, adjusted analysis, for 11 included call screeners, by age-group *46*

TABLE 10C: Consistency in accuracy of screen-in, adjusted analysis, for 11 included call screeners, by race *47*

TABLE 11A: Means and variance of screener's predicted probability of accuracy of screen-in, adjusted analysis, for 11 included screeners *48*

TABLE 11B: Means and variance of screener's predicted probability of accuracy of screen-in, adjusted analysis, by age-group *48*

TABLE 11C: Means and variance of screener's predicted probability of accuracy of screen-in, adjusted analysis, by race *49*

TABLE 12A: Consistency in accuracy of screen-out, adjusted analysis, for 11 included call screener *49*

TABLE 12B: Consistency in accuracy of screen-out, adjusted analysis, for 11 included call screener, by age-group *50*

TABLE 12C: Consistency in accuracy of screen-out, adjusted analysis, for 11 included call screener, by race *51*

TABLE 13A: Means and variance of screener's predicted probability of accuracy of screen-out, adjusted analysis, for 11 included screener *52*

TABLE 13B: Means and variance of screener's predicted probability of accuracy of screen-out, adjusted analysis, by age-group *52*

TABLE 13C: Means and variance of screener's predicted probability of accuracy of screen-out, adjusted analysis, by race *53*

TABLE 14A: Consistency in workload, adjusted analysis, for 11 included call screener *53*

TABLE 14B: Consistency in workload, adjusted analysis, for 11 included call screeners, by age-group *54*

TABLE 14C: Consistency in workload, adjusted analysis, for 11 included call screener, by race *55*

TABLE 15A: Means and variance of screener's predicted workload, adjusted analysis, for 11 included screeners *56*

TABLE 15B: Means and variance of screener's predicted workload, adjusted analysis, by age-group *56*

TABLE 15C: Means and variance of screener's predicted workload, adjusted analysis, by race *57*

TABLE 16A: Estimated magnitude of monthly impact of AFST on accuracy of screen-in *57*

TABLE 16B: Estimated magnitude of impact of AFST on accuracy of screen-out *58*

TABLE 16C: Estimated magnitude of impact of AFST on workload *58*

FIGURES *59*

FIGURE 1. Example of the AFST Score *59*

FIGURE 2A: Accuracy of Screen-In, ITSA analysis *60*

FIGURE 2B: Accuracy of Screen-In, by age-group *60*

FIGURE 2C: Accuracy of Screen-In, by race *61*

FIGURE 3A: Accuracy of Screen-In, adjusted analysis *61*

FIGURE 3B: Accuracy of Screen-In, adjusted analysis, by age-groups *62*

FIGURE 3C: Accuracy of Screen-In, adjusted analysis, by race *62*

FIGURE 4A: Accuracy of Screen-Out, ITSA analysis *63*

FIGURE 4B: Accuracy of Screen-Out, ITSA analysis, by age-group *63*

FIGURE 4C: Accuracy of Screen-Out, ITSA analysis, by race *64*

FIGURE 5A: Accuracy of Screen-Out, adjusted analysis *64*

FIGURE 5B: Accuracy of Screen-Out, adjusted analysis, by age-group *65*

FIGURE 5C: Accuracy of Screen-Out, adjusted analysis, by race *65*

FIGURE 6A: Workload, ITSA Analysis *66*

FIGURE 6B: Workload, ITSA Analysis, by age-group *66*

FIGURE 6C: Workload, ITSA Analysis, by race *67*

FIGURE 7A: Workload, adjusted analysis *67*

FIGURE 7B: Workload, adjusted analysis, by age-group *68*

FIGURE 7C: Workload, adjusted analysis, by race *68*

FIGURE 8A: Consistency of accuracy of screen-in for 11 call screeners, adjusted analysis *69*

FIGURE 8B: Consistency of accuracy of screen-in for 11 call screeners, adjusted analysis, by age-group *69*

FIGURE 8C: Consistency of accuracy of screen-in for 11 call screeners, adjusted analysis, by race *70*

FIGURE 9A: Consistency of accuracy of screen-out for 11 call screeners, adjusted analysis *70*

FIGURE 9B: Consistency of accuracy of screen-out for 11 call screeners, adjusted analysis, by age-group *71*

FIGURE 9C: Consistency of accuracy of screen-out for 11 call screeners, adjusted analysis, by race *71*

FIGURE 10A: Consistency of workload for 11 call screeners, adjusted analysis *72*

FIGURE 10B: Consistency of workload for 11 call screeners, adjusted analysis, by age-group *72*

FIGURE 10C: Consistency of workload for 11 call screeners, adjusted analysis, by race *73*

APPENDIX *74*

APPENDIX A1: Analytic dataset and variable construction *74*

APPENDIX A2: Notes on interrupted time-series (ITSA) *76*

Appendix Tables *77*

TABLE A1A: Accuracy of Screen-in, ITSA analysis, all children, 6 month re-referral window *77*

TABLE A1B: Accuracy of Screen-in, ITSA analysis, < 4 years old, 6 month re-referral window *77*

TABLE A1C: Accuracy of Screen-in, ITSA analysis, 4 to 6 years old, 6 month re-referral window *78*

TABLE A1D: Accuracy of Screen-in, ITSA analysis, 7 to 12 years old, 6 month re-referral window *78*

TABLE A1E: Accuracy of Screen-in, ITSA analysis, 13 to 17 years old, 6 month re-referral window *79*

TABLE A1F: Accuracy of Screen-in, ITSA analysis, White, 6 month re-referral window *79*

TABLE A1G: Accuracy of Screen-in, ITSA analysis, Black/African American, 6 month re-referral window *80*

TABLE A2A : Accuracy of screen-in, adjusted analysis, all children, 6 month re-referral window *80*

TABLE A2B: Accuracy of screen-in, adjusted analysis, by age group, 6 month re-referral window *80*

TABLE A2C: Accuracy of screen-in, adjusted analysis, by race, 6 month re-referral window *81*

TABLE A3A: Accuracy of screen-out, ITSA analysis, all children, 6 month re-referral window *81*

TABLE A3B: Accuracy of screen-out, ITSA analysis, < 4 years old, 6 month re-referral window *82*

TABLE A3C: Accuracy of screen-out, ITSA analysis, 4 to 6 years old, 6 month re-referral window *82*

TABLE A3D: Accuracy of screen-out, , ITSA analysis, 7 to 12 years old, 6 month re-referral window *83*

TABLE A3E: Accuracy of screen-out, ITSA analysis, 13 to 17 years old, 6 month re-referral window *83*

TABLE A3F: Accuracy of screen-out, ITSA analysis, White, 6 month re-referral window *84*

TABLE A3G: Accuracy of screen-out, ITSA analysis, Black/African American, 6 month re-referral window *84*

TABLE A4A: Accuracy of screen-out, adjusted analysis, all children, 6 month re-referral window *84*

TABLE A4B: Accuracy of screen-out, adjusted analysis, by age group, 6 month re-referral window *85*

TABLE A4C: Accuracy of screen-out, adjusted analysis, by race, 6 month re-referral window *85*

TABLE APPENDIX A5: Regression results for further action or no further action and re-referral within 60 days, conditional on screen-in (Outcome 1: accuracy of screen-in) *86*

TABLE APPENDIX A6: Regression results for no re-referral within 60 days, conditional on screen-out (Outcome 2: accuracy of screen-out) *88*

TABLE APPENDIX A7: Regression results for screen-in (Outcome 3: workload) *90*

Appendix Figures *92*

APPENDIX FIGURE 1A: Total children in referral calls, by month *92*

APPENDIX FIGURE 1B: Total children in referral calls, by month and age-group *92*

APPENDIX FIGURE 1C: Total children in referral calls, by month and race *93*

APPENDIX FIGURE 2A: Predicted probability of accuracy of screen-in, consistency across 11 call screeners, adjusted analysis *93*

APPENDIX FIGURE 2B: Predicted probability of accuracy of screen-in, consistency across 11 call screeners, adjusted analysis, by age-group *94*

APPENDIX FIGURE 2C: Predicted probability of accuracy of screen-in, consistency across 11 call screeners, adjusted analysis, by race *94*

APPENDIX FIGURE 3A: Predicted probability of accuracy of screen-out, consistency across 11 call screeners, adjusted analysis *95*

APPENDIX FIGURE 3B: Predicted probability of accuracy of screen-out, consistency across 11 call screeners, adjusted analysis, by age-group *95*

APPENDIX FIGURE 3C: Predicted probability of accuracy of screen-out, consistency across 11 call screeners, adjusted analysis, by race *96*

APPENDIX FIGURE 4A: Predicted probability of workload, consistency across 11 call screeners, adjusted analysis *96*

APPENDIX FIGURE 4B: Predicted probability of workload, consistency across 11 call screeners, adjusted analysis, by age-group *97*

APPENDIX FIGURE 4C: Predicted probability of workload, consistency across 11 call screeners, adjusted analysis, by race *97*

METHODS

Overview

The impact evaluation analyzes how the introduction of a screening score (AFST) for use by Allegheny County Child Welfare Office's intake office as part of the decision-making process for children involved in GPS referrals combined with a set of policy and practice changes affected several important outcome measures relating to accuracy, workload, disparities in accuracy and workload, and consistency in these outcomes across call screeners. The sections below describe the AFST and its implementation, the outcome measures used, the policy changes the evaluation accounts for, the data used in the evaluation, and the analytic approaches chosen to perform the evaluation along with their rationale.

Implementation of the AFST

For each referral (involving one or more child in a household) after the implementation of the AFST, call screeners are presented with a visual which either indicates a mandatory screen-in or displays the AFST score (**Figure 1**). The latter is presented as a tool to aid the call screener in making recommendations about screening decisions regarding further investigation, along with the set of tools that the screeners used prior to implementation. Screening recommendations are made on any referral which is classified as GPS. Generation of the score is based on data related to the individual clients for each referral, which includes the victim child(ren), siblings, parents, legal guardians, perpetrators, and potentially unrelated children and adults in the home. Recommendations are made by the Allegheny County hotline staff (screeners and supervisors) and follow one of three courses: 1) Screen-out of a referral without any further evaluation or assessment, 2) Field screen of the referral to assess whether an investigation is warranted, or 3) Screen-in of a referral, which is synonymous with conducting a formal investigation. A field screen refers to an in-home assessment at the referral household.[2] A field screen is always followed by a decision to either screen-out or screen-in the referral.

At the time of the referral, each individual associated with the referral is assigned both a re-referral and a placement risk score. The AFST, which is the only score that the screener has access to, is based on the maximum score (either re-referral or placement) across all individuals associated with the referral at the time that the referral first occurs. The score has a range of 1 to 20 (where 20 is the highest "risk" and 1 is the lowest), indicating the quantile into which the AFST falls. An "auto screen-in" occurs when the AFST falls above 18 for the placement score.[3]

Other Changes Occurring with the Implementation of the AFST

Several other systematic changes to the call screening process accompanied the full implementation of the AFST. First, the mandatory field screen policy was updated, for households which are not already slated to screen-in and which meet a set of criteria. The maximum age for a mandatory field screen decreased from 7 (previously, households with at least one child age 0 through age 6 years required a field screen, regardless of history) to 4 years of age. In addition, the new

2 A field screen is mandatory for a set of conditions; see Other Changes Occurring with the Implementation of the AFST below for details.

3 Of note, an auto screen-in does not obligate the call screener and/or supervisor to screen-in, but rather implies that this would be the default action.

mandatory field screen policy adds the following three conditions: (1) all children who attend home school/cyber school receive a mandatory field screen regardless of age; (2) any family that has had 4 or more referrals in 2 years without any of the referrals being formally investigated; and (3) anything else where more information is necessary to make a final decision. Importantly, second, call screeners make a recommendation about the decision to screen-in/out to their supervisor who has the responsibility for the ultimate decision. Prior to this set of policy changes, the process of assessment and recommendation involving call screeners and supervisors differed from that described above. Prior to the implementation of the AFST, the primary role of the call screening staff was to gather information to inform supervisor decision-making. Call screeners collected data from several databases and resources, including internal DHS systems (KIDS, Client View), courts, public assistance and criminal justice. Call screeners also spoke with the individual making the report and other key contacts (e.g., schools, doctors). The information collected was given to supervisors for final decision-making. Although the process between screeners and supervisors was collaborative, following implementation of the AFST, call screeners took on a greater role in making recommendations for screening decisions and this process was incorporated into the KIDS system.

Outcomes

We selected outcomes to measure underlying effects of the AFST implementation in terms of accuracy of the call screening process for children involved in referrals, workload entering the system, and disparities across children's age and race/ethnicity in terms of accuracy and workload. We also examined the consistency of these outcomes across call screeners. Specifically, the analysis examines how the implementation of the AFST may have impacted these multiple outcomes, including:

ACCURACY OUTCOMES

Outcome 1 — Likelihood of a screen-in with action taken upon investigation or no further action taken and a re-referral within 2 months: A child is considered to be in this category if a referral that includes the child is screened-in and upon investigation the disposition is further action taken or there is no further action taken and a re-referral occurs within the 2-month time window starting from when the call was referred. "Further action" is defined by the referral service decision and occurs when a referral accepts for service or connects to either an open case or connects to a closed case and is re-opened for service. Because of the re-referral period, the last 2 months (April and May 2018) of calls in the data are not allowed as "index events" since there will not be complete follow-up in the data for re-referrals. These last 2 months of data are only used to determine whether re-referrals occurred or not for calls in the prior months. The rate of such screen-ins was defined as the number of children falling into this category divided by the total number of children screened-in for investigation, computed for referrals falling in each calendar month. We repeat the Outcome 1 analysis, using a 6-month re-referral window, as a robustness check and include the results in the Appendix. This outcome is computed for all children. *AFST-related changes in Outcome 1 are intended to measure how AFST implementation*

impacted one feature of accuracy: do those who are screened-in for investigation have further action taken or if not do re-referrals within several months indicate ongoing issues that the initial screen-in may have been sensitive to?

Outcome 2 — Likelihood of a screen-out with no re-referrals within a 2-month time window: A child is considered screened-out if the report that includes the child is not referred by the call screener for further investigation. If another referral occurs within the time window (i.e., within 2 months of the referral call), then the child is considered to have been screened-out with a re-referral. Because of the re-referral period, the last 2 months (April and May 2018) of calls in the data are not allowed as "index events" since there will not be complete follow-up in the data for re-referrals. These last 2 months of data are only used to determine whether re-referrals occurred or not for calls in the prior months. Of note, a child can have more than one screen-out and re-referral over time, but the "index event" of a call that can be considered a screen-out is only assessed for any subsequent calls outside of the re-referral time window of the previous "index event". We define the rate of screen-outs with no re-referrals within a 2-month time window as the number of children in reports who were not referred by the call screener for further investigation and were subsequently not re-referred within a given number of months divided by the total number of children in reports, computed for referrals falling in each calendar month. This outcome is computed for all children. We repeat the Outcome 2 analysis, using a 6-month re-referral window, as a robustness check and include the results in the appendix. AFST-related changes in Outcome 2 are intended to measure how AFST implementation impacted one feature of accuracy: do those who are screened-out remain unassociated with subsequent referrals?

An analogy can be made between these accuracy outcomes for both screen-ins and screen-outs and the more general concepts and language of screening test assessment. In general screening test terminology, test accuracy is measured based on sensitivity (i.e., true positive fractions) which is the percentage of those subjects with the underlying condition who test positive and specificity (i.e., the true negative fraction) which is the percentage of those subjects without the underlying condition who test negative. Ultimately, the ideal is to have a test with a high positive predictive value, the fraction of test positives that have the underlying condition and the fraction of test negatives who do not have the underlying condition. In our context, accuracy for screen-ins (i.e., test positives) is the fraction of children who screen-in that have further action taken upon investigation (an indicator of the underlying condition). Likewise, accuracy for screen-outs (i.e., test negatives) is the fraction of children who screen-out that have no re-referrals for 2 months (an indicator of the absence of the underlying condition). The hypothesized effect of the AFST score would be to increase both of these.

WORKLOAD OUTCOMES

Outcome 3 — Rates of calls screened-in for investigation: A child is considered screened-in for investigation if the referral (household) that includes the child is referred by the call screener for further investigation. Therefore, we define the rate of "screened-in for investigation" as equal to

the number of children in referrals assigned to further investigation divided by the total number of children in referrals, computed for calls falling in each time interval (i.e., calendar month). This outcome is computed for all children. AFST-related changes in Outcomes 3 are intended to measure how AFST implementation impacted the workload entering the investigative system.

CONSISTENCY OUTCOMES

Outcome 4 — Consistency in call screener actions, as related to Outcomes 1–3: We estimate Outcomes 1–3 by call screener. This outcome is computed for calls screeners associated with at least 350 referrals before and after the implementation of the AFST to enable reasonably stable call screener specific estimates. AFST-related changes in these call screener-specific outcomes are intended to measure how AFST implementation impacted consistency in accuracy and in workload (as measured by Outcomes 1–3).

DISPARITIES OUTCOMES

Across Outcomes 1–4: We estimate Outcomes 1-4 for age and race/ethnic subgroups to examine how AFST-related changes in them differed across these subgroups. Such AFST-related changes are intended to measure how AFST implementation impacted disparities.

Stratification of outcomes by age-group included four age-groups: <4 years, 4 to 6 years, 7 to 12 years, 13 to 17 years.[4] While other/undetermined race categories are controlled for in the main analysis for Outcomes 1–3, the disparities analysis is limited to stratification by white and Black/African only, which include over 90% of children included in referrals.[5]

The entire period of analysis spans August 1, 2013 through May 31, 2018 and focuses on how outcome measures changed as a result of the full implementation of the AFST (December 1, 2016). Hence, for the analyses we divide the data into multiple periods. The period after the full implementation period is termed the "Post-AFST Period" and its outcomes are compared to outcomes in periods prior to this. The time prior to the implementation of the AFST spans August 1, 2013 through July 31, 2016 and is divided in two parts. It is divided into the period August 1, 2013 through December 31, 2014 and the period January 1, 2015 through July 31, 2016. The decision to divide the of pre-AFST period is based on a set of amendments to the State of Pennsylvania's existing Child Protective Services Law, which became effective on December 31, 2014 and had the effect of altering a number of features of referrals to the call center.[6] We focus on this second, later pre-AFST period as the point of comparison in our analysis, and refer to this as the "Pre-AFST" from here. Notably, data for the period between August 1, 2016 and November 30, 2016 are omitted from all analyses. When the AFST was launched, an initial policy decision sought to restrict score generation to only individuals and families who could be substantively identified in prior county data (preventing scores from being displayed that were solely constructed from basic referral/geographic information when the family is otherwise unknown to DHS). The first iteration of this policy initially restricted scores to only situations where a scored child on the call needed to be positively identified with a prior county identifier. Many children, most notably newborns and young babies who are most at risk, often do not have

4 In the individual-level analyses described further below, the analytic models adjust for household composition, including the number of children within a household who are under 1 years of age to control for unobserved effects of very young children on Outcomes 1–3. Due to strong interest in outcomes related to the youngest children, future planned analyses will disaggregate the current age-specific subgroups such that children <1 years and children 1–4 years old will be included in two separate categories when sufficient subgroup-specific sample size has accrued.

5 A child was coded as "Black/African American" if his/her race was Black, African/American or mixed Black or African American, at the time of the referral. For outcomes which incorporate re-referrals, race was coded based on the race recorded in the index referral.

6 Child Protective Services Act, P.L.1240, No.206, 23 PA §§6301-6386 (2015)

system involvement, but their parents or caregivers may have significant current and prior system involvement. The initial design led to situations where known information about adults on the referral could not be used in generating a score if none of the children were recognized, and this was quickly deemed too restrictive . After November 30, 2016, scores could be generated on a call if any individual on the call had a past county identifier. Given the reason for the differential missing-ness of the AFST score, analyses of these data cannot rely on missing at random assumptions, as selection mechanisms could operate to make outcomes higher or lower than both the pre-implementation and post-implementation levels that would not be genuinely attributable to the AFST score implementation per se.

Hypothesized Effects of the Implementation of the AFST Score and Related Policies

Accuracy Outcomes (Outcomes 1 and 2): We hypothesized that if implementation performed as expected, both accuracy outcomes would increase (i.e., a higher proportion of screen-ins with have further action taken upon investigation and a higher proportion of screen-outs would have no re-referrals in the subsequent 2 months). The rationale for this hypothesis is that the score is intended to quantitatively integrate a great deal of available information that is predictive of placement and re-referral probabilities, with such information presumed relevant for screen-in/out decisions.

Workload Outcome (Outcome 3): We hypothesized that there would be no substantial change in workload if implementation performed as expected, though we acknowledged that this hypothesis was weaker than for some of our other outcomes. We believed that the mix of which referrals were screened-in and which were screened-out might change with the hypothesized improvements in accuracy but had no reason to believe that there would be a higher (lower) proportion of total calls screening-in due to the implementation.

Consistency Outcome (Outcome 4 [Outcomes 1–3 by Call Screener]): We hypothesized that if implementation performed as expected, consistency in both accuracy and workload outcomes across call screeners would increase (i.e., variation in outcomes across call screeners would decrease). The rationale for this hypothesis is that the AFST score will be the same for a referral regardless of to which call screener it is displayed, and hence, it will provide a regularized/standardized input/measure to help to inform the screen-in/out decisions.

Disparity Outcomes (Across Outcomes 1–4): We hypothesized that if implementation performed as expected, disparities in outcomes across age and race groups would diminish, though we acknowledged that this hypothesis than for some of our other outcomes. For example, if in the pre-implementation period children in two age groups had similar rates of screen-ins that upon investigation led to further action, but the AFST was better able to predict for one age group than another, then in the post-implementation one age group's accuracy would increase differentially from the other's and the disparity between groups could widen.

Data

All analyses use de-identified data relating to those involved in referrals to Allegheny County's call center. The data consist of information about individual household members including race, legal sex, and age. Additionally, the data enumerate the call screeners and supervisors associated with the referral and track previous referrals and investigations with child welfare and other child-serving systems from August 1, 2013 through May 31, 2018.

The analytic dataset focuses on outcomes (described above) for children below age 18 years at the time of the referral. The analytic dataset also retains data, specifically relating to re-referrals, for children who turned 18 between the time of the initial referral call and the end of a 2-month window. The analytic dataset focuses on children with a referral type of GPS and excludes children with other referral types since GPS referrals allow discretion regarding the call screen decision relative to CPS-type referrals, which are mandated as screen-in.

The analytic dataset also contains several variables used in the analysis to control for child (age, legal sex, race) and household (household counts and composition, socioeconomic status and maximum risk scores) characteristics. For child race, we used a categorical variable which included the category "Unable to Determine", when race was not coded as white, Black/African American or other. Other control variables had complete data. See Appendix A1 for detail on the construction of variables.

Analytic Approach

Overview

For Outcomes 1 to 3, we report three main types of analyses, described in detail below. The first is a comparison of unadjusted population means for the Pre-AFST Period (January 1, 2015 through July 31, 2016) and the Post-AFST Period (December 1, 2016 through May 31, 2018). The second is an analysis of changes in the level and trend of monthly rates of the outcomes in these periods using an Interrupted Time Series Analysis (ITSA). The third is an individual-level multivariate regression analyses to estimate the impact of the AFST on the predicted level of each outcome both Pre-and Post-AFST. For each of these analyses, we consider the effect of the policy for the overall population of children with GPS referrals during the analytic period. For Outcome 4, we focus on the individual-level multivariate regression analyses and how they differed across call screeners and how these differences across call screeners changed in the Post-AFST period.[7] To determine how outcomes for individual call screeners changed after the AFST, we examine the Pre-Post-AFST difference in the predicted outcomes. To determine how the AFST may have regularized and made outcomes more consistent for similar children across call screeners as a group, we compare the Pre- versus Post-AFST variance of predicted outcomes. Of note, the evaluation has the least power to detect differences in call screener consistency. Finally, we consider the effects of the policy implementation on Outcomes 1–4 on subgroups of children defined by their age-group and race/ethnicity characterization. This enables us to consider potential heterogeneity and disparities in the policy's effects across these subgroups. We perform all analyses using Stata (v14) software.

7 To determine consistency between and across call screeners, it is necessary to adjust for differences in case mixes, and therefore, the individual-level multivariate analysis is the appropriate one to focus on. Furthermore, the frequency of referrals per month, by screener, is not consistently high enough to provide statistically meaningful interpretation of call screener-specific ITSA analysis results.

The rationale for performing multiple analyses for Outcomes 1 to 3 is that findings from them are complementary and help to highlight various features of the AFST's effects. Comparisons of unadjusted means and the ITSA analyses describe levels and changes in outcomes within the system overall and for age and race/ethnic subgroups given existing trends in the Pre-AFST period. Individual-level regression analyses focus on changes in levels after adjustment for changes in referral case-mix over time. See "Description of Trends in the Study Population and Changes to Case Mix over Time" below for further details.

Description of Specific Methods

First, the simplest comparison we perform is the comparison of unadjusted means in the Pre-AFST and Post-AFST period, testing whether they are statistically different from one another using a two-sided t-test of equality of means.

Second, Interrupted Time Series Analysis is a proven quasi-experimental research design which is particularly useful in the evaluation of a program change when a randomized trial is infeasible and/or unethical. In the evaluation of the AFST, we perform ITSA on a series of monthly rates of each outcome divided into policy periods as described above. The ITSA measures changes in both the level and slope of each outcome in the Post-AFST months in relation to the Pre-AFST months. The ITSA approach captures population-level changes in outcomes and trends after a policy change in comparison to the levels and trends prior to that change. For our application, the biggest strength of the ITSA approach is the ability to test changes in trends because we observe clear time trends in outcomes in the Pre-AFST data. Making causal inference with ITSA relies on several assumptions with the most important being that the rates of change in outcomes from other causes (secular trends) are much slower than changes due to the abrupt implementation of the policy of interest (see **Appendix A2**).[8]

For our outcomes, we made ITSA model estimates of the form:

$$Outcome_t = \beta_0 + \beta_1 time_t + \beta_2 policy_{1t} + \beta_3 policy_{1t}\ x\ time_{1t} + \beta_4 policy_{2t} + \beta_5 policy_{2t}\ x\ time_{2t} + e_t$$

where $Outcome_t$ is the outcome variable (monthly mean), $time_t$ is the time since the start of the data series, $policy_{1t}$ and $policy_{2t}$ are binary indicators (0 prior to policy implementation, otherwise 1) for: 1) December 31, 2014 amendments to the State of Pennsylvania's existing Child Protective Services Law; 2) December 1, 2016 full implementation of the AFST. The coefficient β_0 represents the intercept, or starting level of the outcome; β_1 represents the slope (or trend) in the outcome prior to any of the policies considered, β_2 and β_4 represent the change in intercept (the immediate impact on levels caused by each policy) and β_3 and β_5 capture the difference in the trends after each policy, respectively; and e_t represents stochastic error. In the ITSA analysis, outcomes are modeled as rates calculated by month (t). The multi-period/multi-policy cumulative changes to level/intercept and trend can be estimated as an extension to this ITSA model and tested for statistical significance. We use the *ITSA* command in Stata.[9]

8 Within the context of ITSA, it is possible to use either an Ordinary Least Squares (OLS) or autoregressive integrated moving-average model (ARIMA) approach. The default itsa command in Stata uses an OLS regression model instead of an ARIMA model because OLS tends to be more flexible and interpretable in an interrupted time-series setting than an ARIMA (Box and Jenkins 1976; Velicer and Harrop 1983) (see Also **Appendix A2**).

9 The itsa command allows the user to specify the number of lags to control for autocorrelation. We test for the correctness of our specification using actest which performs the Cumby-Huizinga general specification test of serial correlation.

We estimate similar *ITSA* models for age-specific subgroups of children and for race/ethnic subgroups.

Third, we used multivariate individual-level regression analyses to assess outcomes while adjusting for child and household characteristics. Specifically, we estimate Generalized Linear Models (glm in Stata) with a logit link enabling greater flexibility in the distribution of the error term than a standard logit model. We run our analysis at the level of any child involved in any call (i.e., not only the child for which the call was made):

$$Outcome_{it} = \beta_0 + \beta_1 X_i + \beta_2 policy_t + e_{it}$$

Where $Outcome_{it}$ is a binary indicator equaling 1 if the child is either screened-in with further action taken upon investigation or re-referral within 2 months (Outcome 1), screened out with no re-referral within 2 months (Outcome 2), or screened in (Outcome 3). X_i is a vector of child and household characteristics and $policy_t$ is a binary indicator for calls after the AFST was implemented and e_{it} is a stochastic error term. The multivariate analysis is at the child level (i) with time (t) represented as continuous days across the Pre- and Post- AFST periods. The policy period is coded as 0 if the referral took place between January 1, 2015 and July 31, 2016 and coded as 1 if the referral call took place after December 1, 2016.

We also examine how race and age-groups are differentially impacted by the AFST, with the following model:

$$Outcome_{it} = \beta_0 + \beta_1 X_i + \beta_2 policy_t + \beta_3 policy_t x Var_i + e_{it}$$

Where Var_i is either race- or age-specific subgroup. Note, race- or age-specific subgroups are also included in the full vector of child/household characteristics, X_i.[10]

We use the margins command in Stata to compute the predicted level of each outcome, both Pre-and Post-AFST implementation. These analyses do not evaluate Pre-AFST or Post-AFST time trends as in the ITSA analyses but rather, they focus on estimates of the average effect of the AFST adjusting for evolving case mix over time. The predictive margins presented in tables and figures of the results can be interpreted as the average outcome if all children in the sample were in either the Pre-AFST or the Post-AFST period, holding all other control variables as they happen to be.

10 Because there is specific interest in outcomes related to black/African American children versus white children, the model testing differential impacts by race excludes children coded as "other" or "undetermined" (~10%). In all other models, all children were included in the analytic sample.

To examine consistency of call screener actions, related to Outcomes 1–3, we limit our sample to those call screeners with ≥ 350 calls in each of the Pre- and Post-AFST and perform a similar set of multivariate analyses using the model:

$$Outcome_{ijt} = \beta_0 + \beta_1 X_i + \beta_2 policy_t + \beta_3 Screener_j + \beta_4 policy_t x\, Screener_j + e_{ijt}$$

Where $Screener_j$ is a screener fixed-effects variable, which defines outcomes based for each individual call screener (j).

For the consistency outcomes, we are interested in how much the screener effects varied in the Pre-AFST period (i.e., variance of the predicted means/margins) and how much they varied in the Post-AFST period and testing for whether changes in variance were statistically significant.

Again, we examine how race- and age-groups are differentially impacted by the AFST by call screener with the following model:

$$Outcome_{ijt} = \beta_0 + \beta_1 X_i + \beta_2 policy_t + \beta_3 Screener_j + \beta_4 policy_t x\, Screener_j + \beta_5 policy_t x\, Var_i + \beta_6 Screener_j x\, Var_i + \beta_7 policy_t x\, Screener_j x Var_i + e_{ijt}$$

As above, we estimate call screener variance and test for changes in call screener variance in performance on the outcomes within each group.

Covariates and Standard Errors

For the models involving multivariate adjustments, definitions of the included covariates are as follows. Child characteristics at the time of the referral include the child's age at referral, race, and legal sex. Household characteristics include a risk score category (low, medium, high, mandatory) based on the household maximum of either the computer generated referral or the placement score, after cutoffs are applied (using the Pre-AFST algorithm to create risk score for Post-AFST children), the household composition, including total persons in the household at the time of the call in these categories: less than 1 year, 1 to 5 years, 6 to 12 years, 13 to 17 years, parents of victim, other adults; a binary variable indicating if the mean age of adults in the household is: 18 to 29 years, 30 to 49 years, 50 to 65 years, or 66+ years, or "no adult age is listed"; and a set of binary indicators for which of five poverty categories the household's zip code belongs to or if the household does not have data for zip code. When covariate information is not complete, variables indicating unknown or unable to determine are used as described above and in the Appendices. Because the outcome is the same for all children in a referral (household), and because we expect some correlation in outcomes among individual call screeners, we cluster the standard errors at the level of the call screener identification number for Outcomes 1–3. Because the consistencies outcome and related disparity outcomes are at the are at the call screener level, we cluster standard errors at the referral level for these analyses.

Description of Trends in the Study Population and Changes to Case Mix over Time

The importance of evaluating changes in trends in outcomes is highlighted by trends in the study population over the periods prior to and after the implementation of the AFST (January 1, 2015 through May 31, 2016 [the Pre-AFST Period] and December 1, 2016 through May 31, 2018 [the Post-AFST Period]). In principle, outcomes like investigational workload or call screener accuracy could be related to overall volume of referrals. If there is a limit on the number of investigations that can be conducted in a month then for days/months with high levels of referrals this limit may force a lower screen-in rate. If there is a limit on the amount of time that call screeners have, then for days/months with high levels of referrals, this will tend to reduce the time the screeners can spend on each call which could decrease accuracy of triaging calls. In fact, the monthly volume of children involved in GPS calls increased between the Pre-AFST period (~1,600 children per month) to (~1,900 children per month) with bigger increases in call volumes involving older children than younger children and with bigger increases for Black/African American children than white children. Hence, trends and changes in trends are important to consider (**Appendix Figures 1a–1c**).

The importance of evaluating outcomes adjusting for case mix is highlighted by changes in the case mix of the study population over time. In principle, outcomes like accuracy could be related to case mix. For example, if the system is more accurate for some subgroups of children then increases in the prevalence of that group could lead to an estimated increase in the unadjusted outcome which might otherwise be attributed to AFST. We examine individual child characteristics in the Pre-AFST and Post-AFST periods (**Table 1a**) as well as household characteristics in these periods (**Table 1b**) for changes in case mix. In general, most individual child and household measures stay similar though there is an increase in the prevalence of Black/African American children in referrals.

RESULTS

Accuracy Outcomes and Disparities in Accuracy Outcomes

In this section, we describe how the implementation of the AFST and related policies changed accuracy in two ways. First, we focus on accuracy for children screening in as measured by the proportion of children with referrals which screened-in for investigation that had further action taken, or if not, had a re-referral within 2 months. Second, we focus on accuracy for children screened out as measured by the proportion of children with referrals which screened-out who had no re-referral calls within 2 months. We also examine differential impacts of the AFST on accuracy by age-specific subgroups and race-specific subgroups to assess disparities in accuracy outcomes.

How did the AFST change accuracy for referrals screened-in?

The AFST increased the accuracy for referrals that screened-in as measured by an increase in the percentage of children screened-in for investigation who had further action taken or, if not had a

re-referral within 2 months. While this improvement in accuracy of being screened-in remained higher throughout the Post-AFST period, the initial improvement effect of implementing the AFST did attenuate somewhat over time. With a re-referral window of 6 months, the direction of the result was the same and there was somewhat less attenuation over time. See **Tables 2**, **3a**, **4a**, **Figures 2a**, **3a**, **Appendix Table A1a**, **Appendix Table A2a** and **Appendix Table A5** for numerical details.

How did the AFST change accuracy for referrals screened-in related to children in different age-groups?

The AFST increased the accuracy for referrals that screened-in for children over age 4, as measured by an increase in the percentage of children in each age-group screened-in for investigation who had further action taken or, if not, had a re-referral within 2 months. While this age subgroup-specific improvement in accuracy of being screened-in remained higher throughout the Post-AFST period, the initial improvement effect of implementing the AFST in each subgroup did attenuate somewhat over time. With a re-referral window of 6 months, the direction of the result was the same and there was somewhat less attenuation over time. See **Tables 2**, **3b-e**, **4b**, **Figures 2b**, **3b**, **Appendix Tables A1b–A1e**, **Appendix Table A2b** and **Appendix Table A5** for numerical details.

How did the AFST change accuracy for referrals screened-in related to children in different race groups?

The AFST had an immediate upward effect on the accuracy for referrals that screened-in for both white and Black/African American children, as measured by an increase in the percentage of children in each race subgroup screened-in for investigation who had further action taken or, if not, had a re-referral within 2 months. For Black/African American children, the initial improvement effect of implementing the AFST attenuated over time, such that there was no significant overall increase in accuracy for Black/African American children when compared to White children. With a re-referral window of 6 months, there was a slightly larger increase in the accuracy of screening-in for Black/African American children, although the increase was not significantly different than zero. See **Tables 2**, **3f–g**, **4c**, **Figures 2c**, **3c**, **Appendix Tables A1f–A1g**, **Appendix Table A2c** and **Appendix Table A5** for numerical details.

How did the AFST change accuracy for referrals screened-out?

The AFST had little effect on accuracy for referrals that screened-out as measured by a decrease in the percentage of children screened-out who had a re-referral within 2 months. The multiple analyses showed small decreases in the accuracy of screening out but only sometimes found this decrease to be statistically significant. Prior to the implementation of the AFST, the accuracy for referrals screened-out was increasing slightly, a trend that the AFST largely halted. The results were similar when we used a re-referral window of 6 months. See T**ables 2, 5a**, **6a**, **Figures 4a**, **5a**, **Appendix Table A3a**, **Appendix Table A4a** and **Appendix Table A6** for numerical details.

How did the AFST change accuracy for referrals screened-out related to children in different age-groups?

In breaking down the AFST's effect on accuracy of being screened-out by age-group, the largest decrease in accuracy occurred in children aged 4 to 6 years. While there were small, non-significant reductions in the accuracy of being screened-out for all age-groups, the larger effect in 4 to 6 year-olds may be due to changes in the policy regarding the maximum age for mandatory field screening which was reduced from under 7 years if age to under 4 years of age in the Post-AFST period, where previously the field screening in this age-group helped to identify more children in this age-group for whom being screened-in for investigation was appropriate. The results were similar when we used a re-referral window of 6 months. See **Tables 2**, **5b–e**, **6b**, **Figures 4b**, **5b**, **Appendix Table A3b–A3e**, **Appendix Table A4b** and **Appendix Table A6** for numerical details.

How did the AFST change accuracy for referrals screened-out related to children in different race groups?

The AFST had little effect on accuracy for referrals that screened-out for both white and Black/African American children, as measured by a decrease in the percentage of children screened-out who had a re-referral within 2 months. The multiple analyses showed small decreases for both race subgroups, which were not significant for white children and only occasionally significant for Black/African American children. The results were similar when we used a re-referral window of 6 months. See **Tables 2**, **5f-g**, **6c**, **Figures 4c**, **5c**, **Appendix Table A3f–A3g**, **Appendix Table A4c** and **Appendix Table A6** for numerical details.

Workload Outcomes and Disparities in Workload Outcomes

In this section, we describe how the implementation of the AFST and related policies changed workload in terms of the fraction of children who screened in and hence had investigations conducted. We also examine differential impacts of the AFST on workload by age-specific subgroups and race-specific subgroups to assess disparities in workload outcomes.

How did the AFST change workload as measured by the fraction of referrals screened-in for further investigation?

Prior to the implementation of the AFST, the fraction of referrals screened-in for investigation were declining. The AFST largely halted this decline. Hence, even though the average level in the Pre-AFST period was higher than in the Post-AFST period, it may well be the case that had the AFST not been implemented, screen-in rates could have continued to decline and been lower than were observed in the Post-AFST period. See **Tables 2**, **7a**, **8a**, **Figures 6a**, **7a** and **Appendix Table A7** for numerical details.

How did the AFST change workload related to children in different age-groups?
Prior to the implementation of the AFST, the fraction of referrals screened-in for investigation were declining for children in all subgroups except for age 4 to 6, with larger declines observed in the oldest age-group (13 to 17). The AFST largely halted these age-specific declines, most noticeably for children aged 7 years and older. Therefore, despite age subgroup-specific declines in levels from the Pre-AFST and Post-AFST periods, it may well have been the case that the subgroup-specific levels in the Post-AFST period could have been even lower without the implementation of the AFST. See **Tables 2**, **7b–e**, **8b**, **Figures 6b**, **7b** and **Appendix Table A7** for numerical details.

How did the AFST change workload related to children in different race groups?
Prior to the implementation of the AFST, the fraction of referrals screened-in for investigation were declining for children in all race groups with larger declines observed in Black / African American children compared to those in white children. The AFST largely halted these race-specific declines, most noticeably for Black / African American children. Therefore, despite race subgroup-specific declines in levels from the Pre-AFST and Post-AFST periods, it may well have been the case that the subgroup-specific levels in the Post-AFST period could have been even lower without the implementation of the AFST. See **Tables 2, 7f–g**, **8c**, **Figures 6c**, **7c** and **Appendix Table A7** for numerical details.

Consistency Outcomes and Disparities in Consistency Outcomes

In this section, we examine the consistency of the AFST's effects on accuracy and workload across call screeners. We examine whether the magnitudes of AFST's effects differed across call screeners and specifically whether AFST decreased their variation in outcomes in the Post-AFST period relative to variation the Pre-AFST period. Finally, we examine whether the AFST's variation in outcomes and change in variation in outcomes differed for referrals involving children of age- or race-specific subgroups. Of note, this analysis is restricted to the 11 call screeners with at least 350 referrals in both the Pre-AFST and Post-AFST periods.[11] Hence, outcomes are not directly comparable to the outcomes for all call screeners reported to this point. See **Table 9**, which shows comparison of the unadjusted means for call screeners included and those excluded from this analysis. While most outcomes are quite similar between the two groups of call screeners, the Post-AFST value for Outcome 3 (fraction of referrals screened-in) are substantially higher for the group included in this analysis than they are for call screeners not included in this analysis, highlighting that this analysis is relevant only for examination of the consistency of Outcomes 1–3 (and not the outcomes themselves), and across only the part of the call screener workforce that has been stable over time.

11 Further, all analyses exclude the post-AFST months on April and May 2018 (referrals made in these months were excluded for Outcomes 1 and 2, so we exclude them for Outcome 1 to maintain consistency among included call screeners).

How did the AFST change the consistency of accuracy for referrals screened-in by call screener?
The overall increase in accuracy of a screen-in was consistent across call screeners, and variation between calls screeners in this outcome did not change significantly. Accuracy of being screened-in increased in the Post-AFST period compared to the Pre-AFST period for 10 of the 11 call

screeners (statistically significantly so in 1 call screener) and decreased in 1 call screener (not statistically significantly). The variance of call screener-specific outcomes decreased (not statistically significantly) in the Post-AFST period compared to the Pre-AFST period. See **Tables 10a**, **11a**, **Figures 8a**, and **Appendix Figure 2a** for numerical details.

How did the AFST change the consistency of accuracy for referrals screened-in by call screener by age-groups?

The overall increase in accuracy of a screen-in for children of different age-groups was consistent across call screeners and was generally larger for children in older age-groups, and variation between call screeners in this outcome did not change significantly for children of any age-group. Most call screeners increased in accuracy for screen-ins for children in each age-group. Very few increases were statistically significant at the call screener/age-group level given the small sample sizes. The variance of call screener specific outcomes decreased (not statistically significantly) in the Post-AFST period compared to the Pre-AFST period for children in all age-groups. See **Tables 10b**, **11b**, **Figures 8b**, and **Appendix Figure 2b** for numerical details.

How did the AFST change the consistency of accuracy for referrals screened-in by call screener by race groups?

The overall increase in accuracy of a screen-in for both white and Black/African American children was reasonably consistent across call screeners, though average increases were smaller for Black/African American children across call screeners. Variation between calls screeners in this outcome did not change significantly for referrals in either race subgroup. Accuracy increased for 10 of the 11 call screeners for white children (statistically significantly so in 2 call screeners) and decreased for 1 call screener (not statistically significantly). Accuracy increased for 7 of the 11 call screeners for Black/African American children (statistically significantly so in 1 call screener) and decreased in 4 of the 11 call screeners (not statistically significantly). The variance of call screener outcomes increased for both race groups (not statistically significantly). See **Tables 10c**, **11c**, **Figure 8c**, and **Appendix Figure 2c** for numerical details.

How did the AFST change the consistency of accuracy for referrals screened-out by call screener?

The overall decrease in accuracy of screen-outs was consistent across call screeners, and variation between calls screeners in this outcome did not change significantly. Accuracy of being screened-out decreased in the Post-AFST period compared to the Pre-AFST period for 11 of the 11 call screeners (statistically significantly so in 1 call screener). The variance of call screener-specific outcomes increased (not statistically significantly) in the Post-AFST period compared to the Pre-AFST period. See **Tables 12a**, **13a**, **Figure 9a**, and **Appendix Figure 3a** for numerical details.

How did the AFST change the consistency of accuracy for referrals screened-out by call screener by age-groups?

The decrease in accuracy of screen-outs for children of different age-groups was generally more concentrated for children in younger age-groups, particularly for those age 4-6 years,

was consistent across call screeners, and variation between call screeners in this outcome did not change significantly for children of any age-group. While most call screeners had decreases in accuracy of screen-outs for younger children especially, very few of these decreases were statistically significant (and none of the increases were). Sample size and frequency of outcome at the call screener/age-group level meant that the analyses of this outcome were underpowered. The variance of call screener specific outcomes increased (statistically significant in 7 to 12 year olds) in the Post-AFST period compared to the Pre-AFST period for children in all age-groups less than 13 years and decreased minimally for the age-group 13 to 17 years (not statistically significantly). See **Tables 12b**, **13b**, **Figure 9b**, and **Appendix Figure 3b** for numerical details.

How did the AFST change the consistency of accuracy for referrals screened-out by call screener by race groups?

The overall decrease in accuracy of screen-outs for both white and Black/African American children was reasonably consistent across call screeners, though average decreases were somewhat larger for Black/African American children across call screeners. Variation between call screeners in this outcome did not change significantly for referrals in either race subgroup. Accuracy decreased for 8 of the 11 call screeners for white children (statistically significantly so in 1 call screener) and increased for 3 call screeners (not statistically significantly). Accuracy decreased for 8 of the 11 call screeners for Black/African American children (statistically significantly so in 1 call screener) and increased in 3 of the 11 call screeners (not statistically significantly). The variance of call screener outcomes increased for both race groups (not statistically significantly). See **Tables 12c**, **13c**, **Figure 9c**, and **Appendix Figure 3c** for numerical details.

How did the AFST change workload differentially by call screener?

The overall finding described above on workload (fraction of referrals screening in for investigation) was that the average level decreased from the Pre-AFST period to the Post-AFST period, though because of strong declining trends in workload in the Pre-AFST period, the Post-AFST levels may have been higher than what they would have been without the implementation of the AFST. There was moderate consistency in the workload outcome across call screeners. Workload increased in the Post-AFST period for 7 of the 11 call screeners (4 of these were significant increases) and decreased for 4 of the 11 call screeners (none statistically significantly). The variance of call screener-specific outcomes decreased (not statistically significantly) in the Post-AFST period compared to the Pre-AFST period. See **Tables 14a**, **15a**, **Figure 10a**, and **Appendix Figure 4a** for numerical details.

How did the AFST change workload differentially by call screener by age-groups?

The by-screener increase in workload for children of different age-groups was generally more concentrated in the middle age-groups, was consistent across call screeners, and variation between call screeners in this outcome did not change significantly for children of any age-group. Very few of the increases in workload were statistically significant, and none of the decreases in workload were statistically significant. Sample size at the call screener/age-group

level meant that the analyses of this outcome were underpowered. The variance of call screener-specific outcomes decreased for the younger age-groups (0 to 6 years) and increased of the older age-groups (not statistically significantly) between the Pre-AFST and the Post-AFST period. See **Tables 14b**, **15b**, **Figure 10b**, and **Appendix Figure 4b** for numerical details.

How did the AFST change workload differentially by call screener by race groups?
The overall change in workload for calls involving white and those involving Black/African American children was reasonably consistent across call screeners. Variation between calls screeners in this outcome did not change significantly for referrals in either race subgroup. Workload increased for 7 of the 11 call screeners for white children (statistically significantly so in 3 call screeners) and decreased for 3 call screeners (statistically significantly so in 1 call screener). Workload increased for 7 of the 11 call screeners for Black/African American children (statistically significantly so in 1 call screener) and decreased for 4 call screeners (not statistically significantly). The variance of call screener outcomes decreased for both race groups, with a larger effect apparent in the Black/African American group (not statistically significantly). See **Tables 14c**, **15c**, **Figure 10c**, and **Appendix Figure 4c** for numerical details.

DISCUSSION, CONCLUSIONS, AND IMPLICATIONS

We evaluated the impact of the AFST screening score implementation within Allegheny County's child welfare office in terms of its effect on accuracy, workload, disparity and consistency outcomes for children involved in GPS referrals. Overall, our analyses showed that the AFST and associated policies increased accuracy for children screened-in for investigation and may have slightly decreased accuracy for children screened-out. Improvements in accuracy attenuated somewhat over time post-implementation. The AFST and associated policies also stopped the downward trend in the rate of children screened-in for investigation. Age- and race-specific subgroup analyses showed that screen-in accuracy improvements were largest and/or had less attenuation over time for white children and children aged < 4 years. Loss of accuracy in screening-out was concentrated most in children ages 4 to 6 years though the overall size of effect even in this age-group was relatively small. This may be due to concurrent changes in the mandatory in-home assessment (field screening) policy in terms of the maximum age being reduced from under 7 to under 4 years of age. Effects were generally consistent across call screeners.

As with all such evaluations, methodological choices and assumptions were required. Below we discuss a number of these and the considerations that justified them as well as their potential limitations.

Accuracy-related outcomes for both screen-ins and screen-outs are defined partly based on whether subsequent referral calls are made within a given time window. In principle, additional referral calls could come shortly after the index call regarding the same incident which would potentially influence the accuracy measure. In our analyses this did not turn out to be the case. The number of index calls for which additional calls occurred within 1 day is less than 1%, within

2 days is less than 2% and within 1 week less than 3%. When we made a robustness check to exclude these calls, our main results did not change in any substantial way.

Analyses examining consistency across call screeners used a cut-off of having at least 350 calls in the pre- and post-AFST periods. While this ensures that estimates of outcomes and effects made at the call screener level tend to be more stable, the cut-off may appear arbitrary and certainly excludes call screeners with fewer calls. In fact, the distribution of calls taken by call screeners is bi-modal with a group of call screeners taking well below 200 calls in total and another taking well over 500. Hence a range of cut-off values would yield the same set of call screeners for analyses.

We analyzed outcomes stratified into age-specific groups, including a young age group of children aged 4 years and under. However, even within this age group, there could be additional heterogeneity, especially at the younger end since infants below the age of 1 year are of concern to agencies given their inherent vulnerability. While we include household composition as control variables in our multivariate analyses – specifically the count of children in the household age 1 year or below – we did not stratify our outcomes by this finer age category as the number of children in this finer age category is insufficient to provide precise effect estimates. Future analyses with more months of follow-up are planned to examine outcomes in finer age groups.

One of the outcomes we examined was the effect of the implementation of the AFST and surrounding policy changes on disparities in outcomes across race/ethnic and age-specific subgroups. It is important to note that true underlying rates of neglect and maltreatment for each of these subgroups is unknown and hence increases/decreases in a given measured system outcome (e.g., screen-ins) of one subgroup relative to another in principle could represent either a widening or a narrowing of a disparity (e.g., in terms of children experiencing actual neglect or maltreatment having the referral investigated). Given that the key assumption of the analysis is that changes in underlying conditions like rates of neglect and maltreatment are substantially slower than the change of implementing the AFST and surrounding policies, examining how outcomes changed from the pre- to the post-implementation period within groups are illustrative for exploring whether the use of the tool within the system led to bigger changes for some groups relative to others. Proper interpretation of such results critically depends on the ability of AFST to detect actual neglect or maltreatment in each group and for workers to act accordingly.

The goal of the evaluation of the effects of the AFST and surrounding policies was to provide a set of measures that are meaningful and important. However, the evaluation makes no claim or judgement about the relative importance of one outcome related to another. Specifically, we analyzed multiple outcomes including accuracy measures for screening in and screening out, workload, consistency across call screeners, and differences in these outcomes by age and race/ethnic subgroups. The stated goal of the AFST implementation is primarily that of improving accuracy. If achieving increased accuracy also involved increases in the number of calls screening-in for investigation, this would not necessarily imply that the AFST implementation was unsuccessful.

Rather, it might imply that additional allocation of investigative resources is required to sustain improvements due to the AFST—a finding that is relevant for other systems considering implementing similar tools. We encourage interpretation of findings across outcomes in a holistic way and with reference to the stated goals and constraints of child-serving systems.

The assessment of effects of the implementation in the evaluation relies on quasi-experimental methods as direct randomization was not feasible. One strength of the evaluation is that it uses multiple quasi-experimental methods – interrupted time series analysis (ITSA) as well as multivariate regression analyses indicators for the timing of the implementation—with findings quite consistent across these methods. As noted in the methods, a key assumption is that the estimated effect (changes pre to post-implementation) can be attributed primarily to the implementation of AFST because other changes (e.g., changes in case mix over time) are much slower and less abrupt than the implementation itself. The multivariate regression adjusts for many features of case mix explicitly. Yet, for both methods, if unmeasured features change relatively abruptly, it is in principle possible that the estimated effect is not attributable entirely to the AFST implementation.

To provide some context in terms of how many children may be affected by the AFST, estimates of child-counts for Outcomes 1–3 are presented in **Tables 16a–16c**. These estimates are based on the predicted probabilities of a given outcome estimated in the adjusted analyses, the related confidence intervals and the mean monthly total counts of children in referrals, children who screen-in and children who screen-out (over both the Pre- and Post- AFST). Roughly 24 more children each month screen-in accurately after the AFST, with over half of these children in the 7 to 12 year old age range and almost all of these children in the white race group. Roughly 11 more children who screen-out are done so inaccurately each month (though this result is not statistically significant) with ~2/3 of these children falling into the Black/African American race group (although the results are not statistically significant for any breakdown of age or race). Roughly 53 fewer children included in referrals screen-in each month (not significant) with over half of these falling in the 13 to 17-year age range and ~2/3 of these children in the black race group.

In conclusion, our evaluation of the effects of implementing the AFST and surrounding policy changes shows moderate improvements in accuracy of screen-ins with small decreases in the accuracy in screen-outs, a halt in the downward trend in pre-implementation screen-ins for investigation, no large or consistent differences across race/ethnic or age-specific subgroups in these outcomes, and no large or substantial differences in consistency across call screeners. As with the initial phases of most large-scale real-world system changes, implementation challenges arose, and one can speculate as to whether the achievable effects without such challenges could have been larger. In sum, the AFST appears to have had a modest positive effect on some screening outcomes that can be determined via process measures. Ultimately, Allegheny County and other systems considering the use tools like the AFST will need to consider how such metrics relate to their core goals (e.g., safety) and how achieving these effects relate to their costs and resource constraints both in terms of implementing the tool and the downstream impacts that such a tool can have.

TABLES

TABLE 1A: **Summary Statistics, child characteristics**

	PRE-AFST (JANUARY 1, 2015 - JULY 31, 2016)			POST-AFST (DECEMBER 1, 20165 - MAY 31, 2018)			P-VALUE*
	MEAN	95% CI		MEAN	95% CI		
Legal sex							
Male	50.72%	50.17%	51.28%	50.36%	49.82%	50.89%	0.352
Female	48.98%	48.43%	49.54%	48.76%	48.23%	49.30%	0.574
Other	0.29%	0.23%	0.36%	0.88%	0.78%	0.98%	0.000
Race							
Black/African American	46.70%	46.15%	47.26%	50.99%	50.45%	51.52%	0.000
White	41.03%	40.48%	41.58%	41.97%	41.44%	42.49%	0.015
Other	12.27%	11.90%	12.63%	7.05%	6.78%	7.32%	0.000
Age-group							
< 4 years	22.88%	22.42%	23.35%	21.95%	21.51%	22.39%	0.004
4–6 years	17.83%	17.40%	18.25%	16.64%	16.24%	17.04%	0.000
7–12 years	34.60%	34.07%	35.13%	36.07%	35.56%	36.58%	0.000
13–17 years	24.69%	24.21%	25.17%	25.34%	24.88%	25.80%	0.055

*Sample sizes are 31,190 (Pre-AFST) and 33,966 (Post-AFST). The child is considered "Black or African American" if their race is coded as "Black or African American" or "Black or African American" mixed with another race. *P-value is the two-sided p-value based on a two-sample t-test of the equality of means.*

TABLE 1B: Summary Statistics, household characteristics

	PRE-AFST (JANUARY 1, 2015–JULY 31, 2016)				POST-AFST (DECEMBER 1, 2016–MAY 31, 2018)				P-VALUE*
	MEAN	STANDARD ERROR	95% CI		MEAN	STANDARD ERROR	95% CI		
Risk score category									
Mandatory	22.60%		22.14%	23.07%	24.84%		24.38%	25.30%	0.000
High	34.06%		33.53%	34.58%	35.55%		35.04%	36.06%	0.000
Medium	24.14%		23.67%	24.62%	22.45%		22.01%	22.90%	0.000
Low	18.41%		17.98%	18.84%	16.92%		16.52%	17.32%	0.000
No score	0.79%		0.69%	0.88%	0.24%		0.18%	0.29%	0.000
Household poverty category (zip code)									
Wealthiest	24.89%		24.41%	25.37%	25.52%		25.06%	25.99%	0.063
Wealthier	20.90%		20.45%	21.36%	19.73%		19.30%	20.15%	0.000
Middle	10.31%		9.97%	10.65%	9.89%		9.57%	10.21%	0.074
Poor	25.60%		25.11%	26.08%	24.30%		23.84%	24.75%	0.000
Poorest	15.33%		14.93%	15.73%	15.44%		15.05%	15.82%	0.705
No zip code information	2.97%		2.78%	3.15%	5.13%		4.89%	5.36%	0.000
Mean age of household adults									
18 - 29 years	21.69%		21.23%	22.15%	20.09%		19.66%	20.52%	0.000
30 - 49 years	69.40%		68.89%	69.91%	71.14%		70.65%	71.62%	0.000
50 - 65 years	4.98%		4.74%	5.22%	5.47%		5.23%	5.72%	0.005
66 years–max	0.38%		0.31%	0.45%	0.28%		0.23%	0.34%	0.028
No adult age information	3.55%		3.35%	3.76%	3.02%		2.84%	3.20%	0.000
Household composition (counts)									
# parents	1.327	0.006			1.282	0.006			0.000
# other adults	1.488	0.005			1.589	0.005			0.000
# age 13–17	0.697	0.005			0.726	0.005			0.000
# age 6–12	1.278	0.006			1.301	0.006			0.007
# age–5	0.805	0.005			0.780	0.005			0.001
# age < 1	0.176	0.002			0.177	0.002			0.849

*All means are for entire sample of all referred children. Sample sizes are 31,190 (Pre-AFST) and 33,966 (Post-AFST). Risk scores categories are based on the maximum risk score within a given referral (household) of either the referral or the placement risk score. Risk bins were calculated using raw risk scores, and bin cutoffs were provided by Allegheny. Individual households have their zip codes categorized into poverty categories based on the American Community Survey (2008–2012) and its determination of the percentage of all households living below the poverty line, as follows: Poorest (>= 25%); Poor (20% to <25%); Mid (15% to <20%); Wealthier (10% to <15%); and Wealthiest (0% to <10%). *P-value is the two-sided p-value based on a two-sample t-test of the equality of means.*

TABLE 2: **Means of outcomes**

	PRE-AFST (JANUARY 1, 2015–JULY 31, 2016)				POST-AFST (DECEMBER 1, 2016–MAY 31, 2018)				P-VALUE*
	MEAN	N	95% CI		MEAN	N	95% CI		
Outcome (1) Accuracy of screen-in: Screen-in with further action taken or re-referral within 60 days									
All children	42.85%	15,016	42.06%	43.64%	46.61%	14,599	45.80%	47.42%	0.000
< 4 years	44.01%	3,947	42.46%	45.56%	45.18%	3,805	43.60%	46.76%	0.301
4 to 6 years	42.68%	2,570	40.77%	44.60%	45.97%	2,482	44.01%	47.93%	0.019
7 to 12 years	40.96%	4,997	39.60%	42.33%	45.93%	5,034	44.55%	47.30%	0.000
13 to 17years	44.37%	3,502	42.73%	46.02%	49.82%	3,278	48.10%	51.53%	0.000
White	39.26%	5,589	37.98%	40.54%	46.35%	5,685	45.05%	47.65%	0.000
Black/ African American	47.28%	7,715	46.17%	48.40%	47.47%	8,091	46.38%	48.56%	0.813
Outcome (2) Accuracy of screen-out: Screen-out with no re-referral within 60 days									
All children	85.02%	14,676	84.45%	85.60%	84.25%	16,433	83.69%	84.80%	0.094
< 4 years	85.49%	2,861	84.20%	86.79%	84.89%	2,979	83.61%	86.18%	0.519
4 to 6 years	85.13%	2,696	83.78%	86.47%	83.71%	2,683	82.31%	85.11%	0.153
7 to 12 years	84.78%	5,269	83.81%	85.75%	84.11%	6,143	83.20%	85.03%	0.327
13 to 17years	84.73%	3,850	83.59%	85.86%	84.45%	4,629	83.40%	85.49%	0.721
White	84.05%	6,488	83.16%	84.94%	83.79%	7,247	82.94%	84.64%	0.678
Black/ African American	84.42%	6,221	83.52%	85.33%	82.99%	7,726	82.15%	83.83%	0.023
Outcome (3) Workload: Screen-in									
All children	48.23%	31,176	47.67%	48.78%	46.19%	33,524	45.65%	46.72%	0.000
< 4 years	55.39%	7,133	54.24%	56.54%	55.22%	7,296	54.08%	56.36%	0.839
4 to 6 years	46.32%	5,559	45.01%	47.63%	47.07%	5,573	45.76%	48.38%	0.431
7 to 12 years	46.38%	10,789	45.44%	47.32%	44.23%	12,119	43.34%	45.11%	0.001
13 to 17years	45.55%	7,695	44.44%	46.66%	40.67%	8,536	39.63%	41.72%	0.000
White	43.69%	12,794	42.83%	44.55%	42.92%	14,067	42.10%	43.73%	0.200
Black/ African American	53.11%	14,559	52.30%	53.93%	50.28%	17,082	49.53%	51.03%	0.000

*Because outcomes are not often finalized on the referral date, we censor the Post-AFST period call-outcome variable at May 31, 2018. To allow complete follow-up for the second and third outcomes (re-referral within 60 days), we only included referrals through March 31, 2018 for the Post-AFST so that April and May data could be used to verify that re-referrals had or had not occurred. Screen-ins (the first outcome) include all children (< 18 years) in all GPS referrals. For the second outcome (screen-out: no re-referrals), any referral call within the 60-day window after the index referral was considered to determine whether a re-referral had occurred. Subsequent referrals outside the window were considered new "index events" for this analysis. The third outcome (screen-in: further action) includes all children who were screened-in at index referral and had a processed "service decision". *P-value is the two-sided p-value based on a two-sample t-test of the equality of means.*

TABLE 3A: Accuracy of screen-in, ITSA analysis, all children

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	54.05%	0.000	51.14%	56.96%
Pre-2014 Policy	Trend	-0.66	0.000	-1.01	-0.32
2014 Policy	Change in level	2.42	0.413	-3.48	8.33
Post 2014 policy, Pre-AFST	Change in trend	0.47	0.025	0.06	0.88
AFST implementation	Change in level	10.19	0.000	7.19	13.19
Post-AFST	Change in trend	-0.28	0.149	-0.66	0.10
Total trend in screen-in/further action rates Pre-AFST		-0.19	0.161	-0.47	0.08
Total trend in screen-in/further action rates Post-AFST		-0.47	0.001	-0.74	-0.21

Note: change in trend is expressed in percentage points/month.

TABLE 3B: Accuracy of screen-in, ITSA analysis, < 4 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	55.88%	0.000	50.85%	60.91%
Pre-2014 Policy	Trend	-0.54	0.015	-0.97	-0.11
2014 Policy	Change in level	-2.01	0.504	-8.00	3.99
Post 2014 policy, Pre-AFST	Change in trend	0.52	0.076	-0.06	1.10
AFST implementation	Change in level	4.85	0.130	-1.48	11.19
Post-AFST	Change in trend	-0.35	0.311	-1.05	0.34
Total trend in screen-in/further action rates Pre-AFST		-0.02	0.917	-0.41	0.37
Total trend in screen-in/further action rates Post-AFST		-0.37	0.198	-0.95	0.20

Note: change in trend is expressed in percentage points/month.

TABLE 3C: Accuracy of screen-in, ITSA analysis, 4 to 6 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	53.37%	0.000	49.37%	57.36%
Pre-2014 Policy	Trend	-0.80	0.000	-1.20	-0.39
2014 Policy	Change in level	6.02	0.068	-0.47	12.50
Post 2014 policy, Pre-AFST	Change in trend	0.55	0.043	0.02	1.07
AFST implementation	Change in level	9.16	0.000	4.83	13.48
Post-AFST	Change in trend	-0.08	0.739	-0.53	0.38
Total trend in screen-in/further action rates Pre-AFST		-0.25	0.143	-0.59	0.09
Total trend in screen-in/further action rates Post-AFST		-0.33	0.034	-0.63	-0.03

Note: change in trend is expressed in percentage points/month.

TABLE 3D: Accuracy of screen-in, ITSA analysis, 7 to 12 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	50.69%	0.000	45.91%	55.47%
Pre-2014 Policy	Trend	-0.60	0.054	-1.21	0.01
2014 Policy	Change in level	3.74	0.359	-4.39	11.87
Post 2014 policy, Pre-AFST	Change in trend	0.30	0.362	-0.36	0.96
AFST implementation	Change in level	13.43	0.000	9.46	17.41
Post-AFST	Change in trend	-0.28	0.302	-0.83	0.26
Total trend in screen-in/further action rates Pre-AFST		-0.30	0.021	-0.55	-0.05
Total trend in screen-in/further action rates Post-AFST		-0.58	0.021	-1.07	-0.09

Note: change in trend is expressed in percentage points/month.

TABLE 3E: Accuracy of screen-in, ITSA analysis, 13 to 17 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	56.94%	0.000	51.95%	61.94%
Pre-2014 Policy	Trend	-0.75	0.000	-1.12	-0.37
2014 Policy	Change in level	2.02	0.445	-3.25	7.28
Post 2014 policy, Pre-AFST	Change in trend	0.62	0.032	0.05	1.18
AFST implementation	Change in level	11.81	0.000	5.97	17.66
Post-AFST	Change in trend	-0.42	0.128	-0.96	0.12
Total trend in screen-in/further action rates Pre-AFST		-0.13	0.530	-0.55	0.29
Total trend in screen-in/further action rates Post-AFST		-0.55	0.003	-0.90	-0.21

Note: change in trend is expressed in percentage points/month.

TABLE 3F: Accuracy of screen-in, ITSA analysis, White

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	47.22%	0.000	40.98%	53.46%
Pre-2014 Policy	Trend	-0.57	0.063	-1.17	0.03
2014 Policy	Change in level	2.18	0.556	-5.22	9.58
Post 2014 policy, Pre-AFST	Change in trend	0.57	0.123	-0.16	1.31
AFST implementation	Change in level	10.02	0.005	3.13	16.91
Post-AFST	Change in trend	-0.26	0.419	-0.90	0.38
Total trend in screen-in/further action rates Pre-AFST		0.01	0.972	-0.42	0.44
Total trend in screen-in/further action rates Post-AFST		-0.25	0.292	-0.73	0.22

Note: change in trend is expressed in percentage points/month.

TABLE 3G: Accuracy of screen-in, ITSA analysis, Black/African American

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	58.87%	0.000	52.83%	64.91%
Pre-2014 Policy	Trend	-0.67	0.026	-1.26	-0.08
2014 Policy	Change in level	2.82	0.483	-5.20	10.83
Post 2014 policy, Pre-AFST	Change in trend	0.41	0.260	-0.31	1.13
AFST implementation	Change in level	8.05	0.001	3.28	12.83
Post-AFST	Change in trend	-0.32	0.314	-0.94	0.31
Total trend in screen-in/further action rates Pre-AFST		-0.26	0.207	-0.68	0.15
Total trend in screen-in/further action rates Post-AFST		-0.58	0.016	-1.05	-0.11

Note: change in trend is expressed in percentage points/month.

TABLE 4A: Accuracy of screen-in, adjusted analysis, all children

	PREDICTED PROBABILITY OF A SCREEN-IN WITH FURTHER ACTION	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST	43.68%	0.000	42.20%	45.15%
Post-AFST	46.56%	0.000	45.04%	48.08%
DIFF (Post - Pre)	2.88%	0.003	0.95%	4.81%

TABLE 4B: Accuracy of screen-in, adjusted analysis, by age group

	PREDICTED PROBABILITY OF A SCREEN-IN WITH FURTHER ACTION	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST				
< 4 years	43.43%	0.000	41.42%	45.44%
4–6 years	43.80%	0.000	41.51%	46.10%
7–12 years	43.82%	0.000	42.00%	45.64%
13–17 years	43.66%	0.000	41.76%	45.57%
Post-AFST				
< 4 years	43.31%	0.000	41.39%	45.22%
4–6 years	46.80%	0.000	44.46%	49.13%
7–12 years	48.46%	0.000	46.69%	50.23%
13–17 years	47.34%	0.000	45.39%	49.29%
Difference Post–Pre				
< 4 years	-0.12%	0.906	-2.19%	1.94%
4–6 years	2.99%	0.088	-0.44%	6.43%
7–12 years	4.64%	0.000	2.12%	7.16%
13–17 years	3.67%	0.019	0.60%	6.75%

TABLE 4C: **Accuracy of screen-in, adjusted analysis, by race**

	PREDICTED PROBABILITY OF A SCREEN-IN WITH FURTHER ACTION*	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
Pre-AFST				
White	43.01%	0.000	40.90%	45.13%
Black/African American	45.51%	0.000	43.62%	47.39%
Post-AFST				
White	49.49%	0.000	47.32%	51.65%
Black/African American	45.44%	0.000	43.55%	47.33%
Difference Post–Pre				
White	6.47%	0.000	3.62%	9.32%
Black/African American	-0.07%	0.961	-2.73%	2.60%
Difference Black–White				
Pre-AFST	2.49%	0.088	-0.37%	5.36%
Post-AFST	-4.05%	0.006	-6.95%	-1.14%

TABLE 5A: **Accuracy of screen-out, ITSA analysis, all children**

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	84.85%	0.000	82.19%	87.51%
Pre-2014 Policy	Trend	-0.07	0.596	-0.32	0.18
2014 Policy	Change in level	0.20	0.876	-2.34	2.73
Post 2014 policy, Pre-AFST	Change in trend	0.20	0.155	-0.08	0.47
AFST implementation	Change in level	-2.45	0.065	-5.06	0.16
Post-AFST	Change in trend	-0.18	0.170	-0.45	0.08
Total trend in screen-out/no re-referral rates Pre-AFST		0.13	0.016	0.02	0.23
Total trend in screen-out/no re-referral rates Post-AFST		-0.05	0.658	-0.29	0.19

Note: change in trend is expressed in percentage points/month.

TABLE 5B: Accuracy of screen-out, ITSA analysis, < 4 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	86.29%	0.000	83.87%	88.71%
Pre-2014 Policy	Trend	0.00	0.981	-0.23	0.23
2014 Policy	Change in level	-1.87	0.263	-5.18	1.45
Post 2014 policy, Pre-AFST	Change in trend	0.12	0.511	-0.25	0.50
AFST implementation	Change in level	-3.85	0.143	-9.06	1.35
Post-AFST	Change in trend	0.01	0.983	-0.52	0.54
Total trend in screen-out/no re-referral rates Pre-AFST		0.12	0.417	-0.18	0.42
Total trend in screen-out/no re-referral rates Post-AFST		0.13	0.562	-0.31	0.57

Note: change in trend is expressed in percentage points/month.

TABLE 5C: Accuracy of screen-out, ITSA analysis, 4 to 6 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	83.97%	0.000	78.79%	89.15%
Pre-2014 Policy	Trend	0.07	0.763	-0.39	0.53
2014 Policy	Change in level	-1.95	0.398	-6.53	2.64
Post 2014 policy, Pre-AFST	Change in trend	0.16	0.509	-0.32	0.64
AFST implementation	Change in level	-4.74	0.007	-8.11	-1.37
Post-AFST	Change in trend	-0.22	0.148	-0.52	0.08
Total trend in screen-out/no re-referral rates Pre-AFST		0.23	0.005	0.07	0.39
Total trend in screen-out/no re-referral rates Post-AFST		0.01	0.952	-0.25	0.27

Note: change in trend is expressed in percentage points/month.

TABLE 5D: Accuracy of screen-out, ITSA analysis, 7 to 12 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	82.47%	0.000	78.42%	86.52%
Pre-2014 Policy	Trend	0.02	0.922	-0.38	0.42
2014 Policy	Change in level	0.70	0.730	-3.33	4.72
Post 2014 policy, Pre-AFST	Change in trend	0.13	0.518	-0.27	0.54
AFST implementation	Change in level	-1.91	0.154	-4.56	0.74
Post-AFST	Change in trend	-0.32	0.076	-0.67	0.03
Total trend in screen-out/no re-referral rates Pre-AFST		0.15	0.001	0.07	0.23
Total trend in screen-out/no re-referral rates Post-AFST		-0.16	0.336	-0.50	0.18

Note: change in trend is expressed in percentage points/month.

TABLE 5E: Accuracy of screen-out, ITSA analysis, 13 to 17 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	87.39%	0.000	83.98%	90.80%
Pre-2014 Policy	Trend	-0.35	0.123	-0.79	0.10
2014 Policy	Change in level	3.26	0.283	-2.78	9.31
Post 2014 policy, Pre-AFST	Change in trend	0.38	0.151	-0.14	0.90
AFST implementation	Change in level	-0.38	0.855	-4.54	3.79
Post-AFST	Change in trend	-0.14	0.507	-0.55	0.28
Total trend in screen-out/no re-referral rates Pre-AFST		0.03	0.814	-0.24	0.30
Total trend in screen-out/no re-referral rates Post-AFST		-0.11	0.499	-0.42	0.21

Note: change in trend is expressed in percentage points/month.

TABLE 5F: Accuracy of screen-out, ITSA analysis, White

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	84.11%	0.000	80.88%	87.33%
Pre-2014 Policy	Trend	0.02	0.909	-0.37	0.42
2014 Policy	Change in level	-3.04	0.261	-8.41	2.34
Post 2014 policy, Pre-AFST	Change in trend	0.28	0.231	-0.18	0.73
AFST implementation	Change in level	-1.85	0.238	-4.97	1.27
Post-AFST	Change in trend	-0.62	0.004	-1.04	-0.21
Total trend in screen-out/no re-referral rates Pre-AFST		0.30	0.012	0.07	0.53
Total trend in screen-out/no re-referral rates Post-AFST		-0.33	0.061	-0.67	0.02

Note: change in trend is expressed in percentage points/month.

TABLE 5G: Accuracy of screen-out, ITSA analysis, Black/African American

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	84.69%	0.000	80.70%	88.67%
Pre-2014 Policy	Trend	-0.21	0.242	-0.57	0.15
2014 Policy	Change in level	3.70	0.121	-1.02	8.41
Post 2014 policy, Pre-AFST	Change in trend	0.20	0.384	-0.25	0.65
AFST implementation	Change in level	-3.20	0.136	-7.46	1.05
Post-AFST	Change in trend	0.18	0.380	-0.23	0.60
Total trend in screen-out/no re-referral rates Pre-AFST		-0.02	0.909	-0.29	0.26
Total trend in screen-out/no re-referral rates Post-AFST		0.17	0.293	-0.15	0.49

Note: change in trend is expressed in percentage points/month.

TABLE 6A: **Accuracy of screen-out, adjusted analysis, all children**

	PREDICTED PROBABILITY OF A SCREEN-OUT WITH NO REREFERRAL	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
Pre-AFST	84.69%	0.000	84.01%	85.36%
Post-AFST	83.51%	0.000	82.46%	84.57%
DIFF (Post–Pre)	-1.17%	0.073	-2.46%	0.11%

TABLE 6B: **Accuracy of screen-out, adjusted analysis, by age group**

	PREDICTED PROBABILITY OF A SCREEN-OUT WITH NO REREFERRAL	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
Pre-AFST				
< 4 years	85.44%	0.000	83.82%	87.06%
4–6 years	84.97%	0.000	83.63%	86.31%
7–12 years	84.80%	0.000	84.02%	85.57%
13–17 years	83.77%	0.000	82.63%	84.92%
Post-AFST				
< 4 years	84.14%	0.000	82.17%	86.11%
4–6 years	82.66%	0.000	80.63%	84.69%
7–12 years	83.52%	0.000	82.03%	85.01%
13–17 years	83.56%	0.000	82.19%	84.93%
Difference Post–Pre				
< 4 years	-1.30%	0.344	-4.00%	1.40%
4–6 years	-2.31%	0.075	-4.86%	0.24%
7–12 years	-1.28%	0.192	-3.21%	0.64%
13–17 years	-0.21%	0.827	-2.11%	1.68%

TABLE 6C: **Accuracy of screen-out, adjusted analysis, by race**

	PREDICTED PROBABILITY OF A SCREEN-OUT WITH NO REREFERRAL	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
Pre-AFST				
White	82.65%	0.000	81.50%	83.79%
Black/African American	85.30%	0.000	84.13%	86.47%
Post-AFST				
White	81.63%	0.000	79.84%	83.42%
Black/African American	83.43%	0.000	81.83%	85.03%
Difference Post–Pre				
White	-1.01%	0.322	-3.02%	0.99%
Black/African American	-1.87%	0.067	-3.87%	0.13%
Difference Black–White				
Pre-AFST	2.65%	0.004	0.86%	4.44%
Post-AFST	1.79%	0.173	-0.78%	4.37%

TABLE 7A: **Workload, ITSA analysis, all children**

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	48.93%	0.000	45.57%	52.28%
Pre-2014 Policy	Trend	0.13	0.415	-0.19	0.46
2014 Policy	Change in level	0.80	0.720	-3.68	5.28
Post 2014 policy, Pre-AFST	Change in trend	-0.58	0.006	-0.99	-0.18
AFST implementation	Change in level	2.16	0.188	-1.09	5.40
Post-AFST	Change in trend	0.44	0.004	0.15	0.73
Total trend in screen-in rates Pre-AFST		-0.45	0.000	-0.69	-0.21
Total trend in screen-in rates Post-AFST		-0.01	0.914	-0.17	0.16

Note: change in trend is expressed in percentage points/month.

TABLE 7B: Workload, ITSA analysis, < 4 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	54.25%	0.000	51.35%	57.16%
Pre-2014 Policy	Trend	-0.03	0.852	-0.34	0.28
2014 Policy	Change in level	4.42	0.053	-0.06	8.91
Post 2014 policy, Pre-AFST	Change in trend	-0.29	0.162	-0.70	0.12
AFST implementation	Change in level	4.07	0.097	-0.76	8.90
Post-AFST	Change in trend	0.08	0.675	-0.30	0.46
Total trend in screen-in rates Pre-AFST		-0.32	0.021	-0.59	-0.05
Total trend in screen-in rates Post-AFST		-0.24	0.087	-0.51	0.04

Note: change in trend is expressed in percentage points/month.

TABLE 7C: Workload, ITSA analysis, 4 to 6 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	46.02%	0.000	41.73%	50.32%
Pre-2014 Policy	Trend	0.16	0.442	-0.26	0.59
2014 Policy	Change in level	-0.59	0.868	-7.71	6.53
Post 2014 policy, Pre-AFST	Change in trend	-0.42	0.182	-1.05	0.21
AFST implementation	Change in level	3.03	0.225	-1.92	7.98
Post-AFST	Change in trend	0.26	0.330	-0.27	0.79
Total trend in screen-in rates Pre-AFST		-0.26	0.266	-0.72	0.20
Total trend in screen-in rates Post-AFST		0.00	0.995	-0.26	0.26

Note: change in trend is expressed in percentage points/month.

TABLE 7D: Workload, ITSA analysis, 7 to 12 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	47.30%	0.000	43.35%	51.25%
Pre-2014 Policy	Trend	0.09	0.676	-0.32	0.50
2014 Policy	Change in level	0.25	0.922	-4.82	5.32
Post 2014 policy, Pre-AFST	Change in trend	-0.39	0.109	-0.87	0.09
AFST implementation	Change in level	-0.38	0.855	-4.57	3.80
Post-AFST	Change in trend	0.42	0.022	0.06	0.78
Total trend in screen-in rates Pre-AFST		-0.31	0.019	-0.56	-0.05
Total trend in screen-in rates Post-AFST		0.12	0.364	-0.14	0.37

Note: change in trend is expressed in percentage points/month.

TABLE 7E: Workload, ITSA analysis, 13 to 17 years old

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	48.74%	0.000	42.85%	54.64%
Pre-2014 Policy	Trend	0.33	0.189	-0.17	0.84
2014 Policy	Change in level	-1.20	0.689	-7.19	4.79
Post 2014 policy, Pre-AFST	Change in trend	-1.24	0.000	-1.81	-0.67
AFST implementation	Change in level	3.60	0.089	-0.57	7.78
Post-AFST	Change in trend	0.97	0.000	0.60	1.35
Total trend in screen-in rates Pre-AFST		-0.91	0.000	-1.17	-0.64
Total trend in screen-in rates Post-AFST		0.06	0.643	-0.21	0.33

Note: change in trend is expressed in percentage points/month.

TABLE 7F: Workload, ITSA analysis, White

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	44.22%	0.000	40.66%	47.77%
Pre-2014 Policy	Trend	0.04	0.834	-0.33	0.41
2014 Policy	Change in level	0.70	0.794	-4.63	6.03
Post 2014 policy, Pre-AFST	Change in trend	-0.26	0.262	-0.73	0.20
AFST implementation	Change in level	1.24	0.559	-2.99	5.47
Post-AFST	Change in trend	0.20	0.309	-0.19	0.58
Total trend in screen-in rates Pre-AFST		-0.22	0.112	-0.50	0.05
Total trend in screen-in rates Post-AFST		-0.03	0.844	-0.29	0.24

Note: change in trend is expressed in percentage points/month.

TABLE 7G: Workload, ITSA analysis, Black/African American

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	54.83%	0.000	51.47%	58.18%
Pre-2014 Policy	Trend	0.20	0.169	-0.09	0.49
2014 Policy	Change in level	0.50	0.818	-3.88	4.89
Post 2014 policy, Pre-AFST	Change in trend	-0.87	0.000	-1.27	-0.48
AFST implementation	Change in level	3.40	0.078	-0.40	7.20
Post-AFST	Change in trend	0.68	0.000	0.34	1.02
Total trend in screen-in rates Pre-AFST		-0.67	0.000	-0.94	-0.40
Total trend in screen-in rates Post-AFST		0.01	0.940	-0.20	0.21

Note: change in trend is expressed in percentage points/month.

TABLE 8A: Workload, adjusted analysis, all children

	PREDICTED PROBABILITY OF A SCREEN-IN	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST	48.75%	0.000	46.84%	50.66%
Post-AFST	45.70%	0.000	42.67%	48.73%
DIFF (Post–Pre)	-3.05%	0.017	-6.47%	0.36%

TABLE 8B: **Workload, adjusted analysis, by age-group**

	PREDICTED PROBABILITY OF A SCREEN-IN	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
Pre-AFST				
< 4 years	50.16%	0.000	47.75%	52.58%
4–6 years	47.84%	0.000	45.62%	50.05%
7–12 years	49.44%	0.000	47.47%	51.40%
13–17 years	47.25%	0.000	45.12%	49.38%
Post-AFST				
< 4 years	48.46%	0.000	46.09%	50.83%
4–6 years	47.65%	0.000	44.35%	50.94%
7–12 years	46.39%	0.000	42.92%	49.86%
13–17 years	40.99%	0.000	37.68%	44.30%
Difference Post–Pre				
< 4 years	-1.71%	0.233	-4.51%	1.09%
4–6 years	-0.19%	0.913	-3.62%	3.24%
7–12 years	-3.05%	0.139	-7.09%	0.99%
13–17 years	-6.26%	0.003	-10.44%	-2.08%

TABLE 8C: **Workload, adjusted analysis, by race**

	PREDICTED PROBABILITY OF A SCREEN-IN	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
Pre-AFST				
White	48.64%	0.000	46.31%	50.97%
Black/African American	49.98%	0.000	47.59%	52.36%
Post-AFST				
White	46.82%	0.000	43.84%	49.80%
Black/African American	46.03%	0.000	42.84%	49.22%
Difference Post–Pre				
White	-1.82%	0.349	-5.63%	1.99%
Black/African American	-3.95%	0.040	-7.72%	-0.18%
Difference Black–White				
Pre-AFST	1.34%	0.322	-1.31%	3.98%
Post-AFST	-0.79%	0.226	-2.08%	0.49%

TABLE 9: **Means of Outcomes (1)–(3) for call screeners included and excluded from Outcome 4/Consistency analyses**

	PRE-AFST (JANUARY 1, 2015–JULY 31, 2016)				**POST-AFST** (DECEMBER 1, 2016–MAY 31, 2018)				
	MEAN	N	95% CI		MEAN	N	95% CI		P-VALUE*
Outcome (1) Accuracy of screen-in (Screen-in with further action taken or re-referral within 60 days)									
Screeners excluded	45.12%	4,978	43.74%	46.50%	45.77%	4,352	44.29%	47.25%	0.527
Screeners included	41.73%	10,038	40.77%	42.70%	46.97%	10,247	46.00%	47.94%	0.000
Outcome (2) Accuracy of screen-out (Screen-out with no re-referral within 60 days)									
Screeners excluded	84.43%	4,862	83.41%	85.45%	86.36%	7,641	85.59%	87.13%	0.003
Screeners included	85.23%	9,808	84.52%	85.93%	82.47%	8,787	81.68%	83.27%	0.000
Outcome (3) Workload (Screen-in)									
Screeners excluded	48.25%	10,328	47.28%	49.21%	36.05%	13,117	35.23%	36.87%	0.000
Screeners included	48.22%	20,848	47.54%	48.89%	52.70%	20,407	52.02%	53.39%	0.000

*Screeners excluded are those with less than 350 referral calls in either the pre-AFST or the post-AFST. *P-value is the two-sided p-value based on a two-sample t-test of the equality of means.*

TABLE 10A: **Consistency in accuracy of screen-in, adjusted analysis, for 11 included call screeners**

SCREENER	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
1	7.58	0.038	0.43	14.74
2	4.94	0.145	-1.71	11.59
3	8.66	0.005	2.65	14.66
4	6.04	0.136	-1.91	14.00
5	5.34	0.149	-1.92	12.60
6	4.18	0.229	-2.62	10.98
7	5.56	0.172	-2.41	13.54
8	0.46	0.944	-12.38	13.29
9	3.28	0.459	-5.41	11.98
10	2.99	0.582	-7.66	13.64
11	-0.55	0.914	-10.55	9.45

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on all children in the sample for screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 10B: **Consistency in accuracy of screen-in, adjusted analysis, for 11 included call screeners, by age-group**

SCREENER	AGE-GROUP	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
1	< 4 years	-0.37	0.940	-9.98	9.24
1	4 to 6 years	11.91	0.050	0.01	23.80
1	7 to 12 years	11.21	0.020	1.79	20.63
1	13 to 17 years	8.17	0.128	-2.35	18.69
2	< 4 years	0.59	0.899	-8.61	9.80
2	4 to 6 years	7.44	0.147	-2.62	17.50
2	7 to 12 years	6.64	0.156	-2.54	15.81
2	13 to 17 years	5.40	0.297	-4.76	15.57
3	< 4 years	3.83	0.401	-5.11	12.77
3	4 to 6 years	6.92	0.168	-2.91	16.76
3	7 to 12 years	9.61	0.015	1.88	17.34
3	13 to 17 years	13.25	0.004	4.29	22.22
4	< 4 years	-1.80	0.747	-12.75	9.14
4	4 to 6 years	2.93	0.663	-10.23	16.09
4	7 to 12 years	12.16	0.025	1.56	22.77
4	13 to 17 years	7.80	0.199	-4.11	19.72
5	< 4 years	0.35	0.946	-9.85	10.56
5	4 to 6 years	14.55	0.013	3.05	26.04
5	7 to 12 years	9.05	0.069	-0.70	18.81
5	13 to 17 years	-0.71	0.900	-11.73	10.32
6	< 4 years	-0.96	0.839	-10.25	8.33
6	4 to 6 years	-2.83	0.592	-13.18	7.52
6	7 to 12 years	6.08	0.184	-2.89	15.05
6	13 to 17 years	12.33	0.020	1.95	22.71
7	< 4 years	1.36	0.820	-10.40	13.12
7	4 to 6 years	12.77	0.053	-0.19	25.73
7	7 to 12 years	4.51	0.400	-5.99	15.02
7	13 to 17 years	5.83	0.296	-5.11	16.78
8	< 4 years	3.85	0.630	-11.85	19.55
8	4 to 6 years	3.44	0.758	-18.42	25.29
8	7 to 12 years	-1.11	0.898	-18.14	15.92
8	13 to 17 years	-3.35	0.749	-23.90	17.19
9	< 4 years	5.05	0.361	-5.79	15.90
9	4 to 6 years	6.37	0.323	-6.26	19.01
9	7 to 12 years	1.69	0.768	-9.59	12.97
9	13 to 17 years	0.50	0.945	-13.74	14.74

SCREENER	AGE-GROUP	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.] LOWER	UPPER
10	< 4 years	-2.23	0.745	-15.67	11.21
10	4 to 6 years	10.04	0.206	-5.52	25.61
10	7 to 12 years	3.66	0.659	-12.60	19.91
10	13 to 17 years	4.73	0.613	-13.60	23.05
11	< 4 years	-16.45	0.032	-31.49	-1.41
11	4 to 6 years	-0.24	0.975	-15.06	14.58
11	7 to 12 years	7.72	0.234	-4.98	20.41
11	13 to 17 years	5.34	0.509	-10.52	21.21

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on the entire sample for age-group, screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 10C: Consistency in accuracy of screen-in, adjusted analysis, for 11 included call screeners, by race

SCREENER	RACE	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.] LOWER	UPPER
1	white	12.68	0.013	2.63	22.72
1	Black/African American	5.50	0.274	-4.36	15.37
2	white	6.33	0.213	-3.63	16.29
2	Black/African American	3.30	0.479	-5.83	12.43
3	white	3.87	0.401	-5.17	12.91
3	Black/African American	11.88	0.004	3.84	19.92
4	white	-0.79	0.894	-12.40	10.83
4	Black/African American	10.38	0.071	-0.88	21.64
5	white	11.68	0.029	1.17	22.19
5	Black/African American	1.14	0.828	-9.15	11.42
6	white	9.24	0.08	-1.11	19.59
6	Black/African American	1.89	0.697	-7.62	11.40
7	white	10.64	0.073	-1.01	22.29
7	Black/African American	0.23	0.969	-11.11	11.56
8	white	8.08	0.436	-12.25	28.42
8	Black/African American	-3.32	0.71	-20.82	14.19
9	white	10.34	0.117	-2.57	23.25
9	Black/African American	-3.22	0.594	-15.05	8.61
10	white	8.74	0.203	-4.71	22.19
10	Black/African American	-2.03	0.817	-19.28	15.21
11	white	5.71	0.456	-9.30	20.73
11	Black/African American	-3.02	0.661	-16.49	10.46

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on the entire sample for race, screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 11A: **Means and variance of screener's predicted probability of accuracy of screen-in, adjusted analysis, for 11 included screeners**

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST	11	42.48	0.67	2.22	40.99	43.97
Post-AFST	11	46.89	0.92	3.06	44.83	48.95
Difference Post–Pre	11	-4.41	0.84	2.77	-6.27	-2.55
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST) = Mean (post-AFST)					p value=0.000	
Levene's test of equality of variance between groups—testing variance of predicted margins (i.e. variance in level of outcome)						
Ho: Var (pre-AFST) = Var (post-AFST)					p value=0.375	

TABLE 11B: **Means and variance of screener's predicted probability of accuracy of screen-in, adjusted analysis, by age-group**

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST (< 4 years)	11	43.73	1.47	4.89	40.45	47.01
Post-AFST (< 4 years)	11	43.11	1.09	3.61	40.69	45.54
Pre-AFST (4 to 6 years)	11	40.07	1.52	5.03	36.69	43.45
Post-AFST (4 to 6 years)	11	46.74	1.43	4.75	43.54	49.93
Pre-AFST (7 to 12 years)	11	42.41	1.07	3.55	40.03	44.80
Post-AFST (7 to 12 years)	11	48.89	0.92	3.06	46.83	50.95
Pre-AFST (13 to 17 years)	11	42.98	1.27	4.23	40.14	45.82
Post-AFST (13 to 17 years)	11	48.37	0.82	2.73	46.54	50.2019
Difference Pre–Post						
< 4 years	11	0.62	1.74	5.77	-3.26	4.49
4 to 6 years	11	-6.66	1.65	5.49	-10.35	-2.98
7 to 12 years	11	-6.47	1.23	4.06	-9.20	-3.74
13 to 17 years	11	-5.39	1.54	5.12	-8.83	-1.95
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST/< 4 years) = Mean (post-AFST/< 4 years)					p-value=	0.731
Ho: Mean (pre-AFST/4 to 6 years) = Mean (post-AFST/4 to 6 years)					p-value=	0.002
Ho: Mean (pre-AFST/7 to 12 years) = Mean (post-AFST/7 to 12years)					p-value=	0.000
Ho: Mean (pre-AFST/13 to 17 years) = Mean (post-AFST/13 to 17 years)					p-value=	0.006
Levene's test of equality of variance between groups						
Ho: Variance (pre-AFST/< 4 years) = Variance (post-AFST/< 4 years)					p-value =	0.375
Ho: Variance (pre-AFST/4 to 6 years) = Variance (post-AFST/4 to 6 years)					p-value =	0.642
Ho: Variance (pre-AFST/7 to 12 years) =Variance (post-AFST/7 to 12years)					p-value =	0.492
Ho: Variance (pre-AFST/13 to 17 years) = Variance (post-AFST/13 to 17 years)					p-value =	0.067

TABLE 11C: Means and variance of screener's predicted probability of accuracy of screen-in, adjusted analysis, by race

	N	MEAN	STD. ERR.	STD. DEV.		[95% CI]
Pre-AFST (white)	11	41.05	1.21	4.03	38.35	43.76
Post-AFST (white)	11	48.92	1.60	5.29	45.36	52.47
Pre-AFST(Black)	11	44.59	1.08	3.60	42.17	47.00
Post-AFST(Black)	11	46.65	1.12	3.73	44.15	49.16
Difference Pre–Post						
White	11	-7.87	1.18	3.90	-10.49	-5.25
Black/African American	11	-2.07	1.60	5.32	-5.64	1.50
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST/white) = Mean (post-AFST/white)					p-value=	0.000
Ho: Mean (pre-AFST/Black) = Mean (post-AFST/Black)					p-value=	0.226
Levene's test of equality of variance between groups						
Ho: Variance(pre-AFST/white) = Variance(post-AFST/white)					p-value=	0.309
Ho: Variance(pre-AFST/Black) = Variance(post-AFST/Black)					p-value=	0.862

TABLE 12A: Consistency in accuracy of screen-out, adjusted analysis, for 11 included call screeners

SCREENER	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
1	-2.08	0.488	-7.96	3.80
2	-0.49	0.855	-5.72	4.75
3	-1.41	0.545	-5.98	3.16
4	-3.69	0.286	-10.46	3.09
5	-5.04	0.085	-10.78	0.70
6	-2.15	0.399	-7.13	2.84
7	-0.85	0.798	-7.33	5.64
8	-0.89	0.824	-8.71	6.94
9	-1.76	0.585	-8.07	4.55
10	-9.57	0.013	-17.11	-2.02
11	-1.03	0.771	-7.97	5.91

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on all children in the sample for screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 12B: **Consistency in accuracy of screen-out, adjusted analysis, for 11 included call screeners, by age-group**

SCREENER	AGE-GROUP	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
1	< 4 years	-0.19	0.969	-9.64	9.26
1	4 to 6 years	-10.65	0.051	-21.34	0.04
1	7 to 12 years	3.46	0.317	-3.31	10.22
1	13 to 17 years	-4.83	0.309	-14.13	4.47
2	< 4 years	-1.97	0.687	-11.58	7.63
2	4 to 6 years	-5.65	0.255	-15.37	4.07
2	7 to 12 years	2.41	0.483	-4.32	9.14
2	13 to 17 years	-0.02	0.994	-7.09	7.04
3	< 4 years	1.55	0.632	-4.79	7.89
3	4 to 6 years	-2.66	0.487	-10.17	4.84
3	7 to 12 years	-1.79	0.582	-8.16	4.58
3	13 to 17 years	-2.36	0.511	-9.40	4.68
4	< 4 years	-4.77	0.350	-14.79	5.24
4	4 to 6 years	-8.90	0.151	-21.05	3.26
4	7 to 12 years	-5.10	0.264	-14.06	3.85
4	13 to 17 years	1.38	0.798	-9.18	11.94
5	< 4 years	-7.87	0.118	-17.75	2.00
5	4 to 6 years	-7.80	0.125	-17.77	2.17
5	7 to 12 years	-0.67	0.853	-7.77	6.42
5	13 to 17 years	-6.81	0.100	-14.92	1.30
6	< 4 years	-7.00	0.124	-15.91	1.91
6	4 to 6 years	-8.65	0.076	-18.22	0.92
6	7 to 12 years	1.61	0.594	-4.31	7.54
6	13 to 17 years	-0.57	0.874	-7.64	6.50
7	< 4 years	8.36	0.123	-2.26	18.98
7	4 to 6 years	3.19	0.577	-8.03	14.41
7	7 to 12 years	-9.08	0.045	-17.95	-0.21
7	13 to 17 years	-0.23	0.965	-10.49	10.03
8	< 4 years	-0.60	0.921	-12.43	11.23
8	4 to 6 years	0.83	0.893	-11.24	12.91
8	7 to 12 years	-7.30	0.213	-18.78	4.18
8	13 to 17 years	6.43	0.427	-9.43	22.29
9	< 4 years	-17.85	0.005	-30.19	-5.52
9	4 to 6 years	-3.27	0.510	-12.99	6.46
9	7 to 12 years	2.53	0.560	-5.96	11.02
9	13 to 17 years	2.89	0.460	-4.77	10.55

SCREENER	AGE-GROUP	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
10	< 4 years	-1.95	0.793	-16.56	12.66
10	4 to 6 years	-18.71	0.001	-29.95	-7.46
10	7 to 12 years	-10.93	0.036	-21.12	-0.73
10	13 to 17 years	-8.51	0.088	-18.30	1.28
11	< 4 years	1.50	0.799	-10.06	13.06
11	4 to 6 years	0.92	0.856	-9.03	10.88
11	7 to 12 years	-5.82	0.181	-14.33	2.70
11	13 to 17 years	2.61	0.670	-9.41	14.63

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on the entire sample for age-group, screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 12C: **Consistency in accuracy of screen-out, adjusted analysis, for 11 included call screeners, by race**

SCREENER	RACE	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
1	white	-0.38	0.932	-9.22	8.45
1	Black/African American	-4.36	0.321	-12.97	4.25
2	white	-3.10	0.479	-11.70	5.49
2	Black/African American	0.86	0.805	-5.95	7.67
3	white	-4.47	0.198	-11.27	2.34
3	Black/African American	1.16	0.730	-5.41	7.73
4	white	-4.70	0.411	-15.91	6.51
4	Black/African American	-0.86	0.848	-9.66	7.93
5	white	-9.22	0.037	-17.88	-0.56
5	Black/African American	-2.54	0.517	-10.24	5.15
6	white	-2.04	0.589	-9.43	5.36
6	Black/African American	-3.89	0.320	-11.55	3.78
7	white	-4.96	0.311	-14.58	4.65
7	Black/African American	4.24	0.411	-5.88	14.35
8	white	7.11	0.278	-5.74	19.97
8	Black/African American	-7.18	0.169	-17.41	3.06
9	white	0.46	0.919	-8.39	9.32
9	Black/African American	-4.06	0.399	-13.48	5.37
10	white	-1.81	0.783	-14.71	11.09
10	Black/African American	-16.43	0.001	-26.45	-6.40
11	white	4.28	0.402	-5.72	14.27
11	Black/African American	-7.03	0.190	-17.56	3.49

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on the entire sample for race, screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 13A: **Means and variance of screener's predicted probability of accuracy of screen-out, adjusted analysis, for 11 included screeners**

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST	11	85.02	0.43	1.43	84.07	85.98
Post-AFST	11	82.39	0.63	2.09	80.99	83.80
Difference Pre–Post	11	2.63	0.80	2.67	0.84	4.42
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST) = Mean (post-AFST)					pvalue=	0.008
Levene's test of equality of variance between groups - testing variance of predicted margins (i.e. variance in level of outcome)						
Ho: Var (pre-AFST) = Var (post-AFST)					pvalue=	0.296

TABLE 13B: **Means and variance of screener's predicted probability of accuracy of screen-out, adjusted analysis, by age-group**

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST (< 4 years)	11	85.25	1.24	4.10	82.49	88.00
Post-AFST (< 4 years)	11	82.45	1.53	5.07	79.04	85.85
Pre-AFST (4 to 6 years)	11	86.17	0.85	2.80	84.28	88.05
Post-AFST (4 to 6 years)	11	80.59	1.42	4.69	77.44	83.74
Pre-AFST (7 to 12 years)	11	85.54	0.62	2.04	84.16	86.91
Post-AFST (7 to 12 years)	11	82.75	1.22	4.06	80.02	85.48
Pre-AFST (13 to 17 years)	11	83.54	1.03	3.43	81.23	85.84
Post-AFST (13 to 17 years)	11	82.63	0.95	3.15	80.51	84.74
Difference Pre - Post						
< 4 years	11	2.80	2.02	6.70	-1.70	7.30
4 to 6 years	11	5.58	1.90	6.30	1.34	9.81
7 to 12 years	11	2.79	1.54	5.10	-0.64	6.21
13 to 17 years	11	0.91	1.34	4.44	-2.07	3.90
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST/< 4 years) = Mean (post-AFST/< 4 years)					p-value=	0.196
Ho: Mean (pre-AFST/4 to 6 years) = Mean (post-AFST/4 to 6 years)					p-value=	0.015
Ho: Mean (pre-AFST/7 to 12 years) = Mean (post-AFST/7 to 12years)					p-value=	0.100
Ho: Mean (pre-AFST/13 to 17 years) = Mean (post-AFST/13 to 17 years)					p-value=	0.512
Levene's test of equality of variance between groups						
Ho: Variance (pre-AFST/< 4 years) = Variance (post-AFST/< 4 years)					p-value =	0.346
Ho: Variance (pre-AFST/4 to 6 years) = Variance (post-AFST/4 to 6 years)					p-value =	0.062
Ho: Variance (pre-AFST/7 to 12 years) =Variance (post-AFST/7 to 12years)					p-value =	0.036
Ho: Variance (pre-AFST/13 to 17 years) = Variance (post-AFST/13 to 17 years)					p-value =	0.779

TABLE 13C: **Means and variance of screener's predicted probability of accuracy of screen-out, adjusted analysis, by race**

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST (white)	11	82.73	0.52	1.73	81.56	83.89
Post-AFST (white)	11	81.02	1.05	3.49	78.67	83.36
Pre-AFST(Black)	11	85.82	0.96	3.19	83.68	87.97
Post-AFST(Black)	11	82.18	1.10	3.65	79.72	84.63
Difference Pre–Post						
White	11	1.71	1.37	4.54	-1.33	4.76
Black/African American	11	3.65	1.66	5.50	-0.05	7.34
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST/white) = Mean (post-AFST/white)					p-value=	0.239
Ho: Mean (pre-AFST/Black) = Mean (post-AFST/Black)					p-value=	0.053
Levene's test of equality of variance between groups						
Ho: Variance(pre-AFST/white) = Variance(post-AFST/white)					p-value=	0.051
Ho: Variance(pre-AFST/Black) = Variance(post-AFST/Black)					p-value=	0.693

TABLE 14A: **Consistency in workload, adjusted analysis, for 11 included call screeners**

SCREENER	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.] LOWER	UPPER
1	6.75	0.008	1.74	11.76
2	1.27	0.613	-3.66	6.21
3	2.82	0.193	-1.43	7.08
4	-3.39	0.269	-9.40	2.62
5	4.75	0.066	-0.31	9.82
6	-0.78	0.742	-5.44	3.88
7	5.28	0.054	-0.10	10.67
8	-5.97	0.184	-14.76	2.83
9	7.47	0.01	1.80	13.14
10	-5.85	0.117	-13.17	1.47
11	7.42	0.048	0.06	14.78

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on all children in the sample for screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 14B: **Consistency in workload, adjusted analysis, for 11 included call screeners, by age-group**

SCREENER	AGE-GROUP	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
1	< 4 years	1.60	0.662	-5.59	8.80
1	4 to 6 years	0.05	0.990	-8.21	8.32
1	7 to 12 years	12.08	0.000	5.67	18.49
1	13 to 17 years	9.16	0.021	1.37	16.94
2	< 4 years	-0.15	0.968	-7.63	7.33
2	4 to 6 years	4.15	0.283	-3.42	11.72
2	7 to 12 years	2.55	0.440	-3.92	9.01
2	13 to 17 years	-1.52	0.677	-8.67	5.63
3	< 4 years	-0.36	0.909	-6.62	5.89
3	4 to 6 years	8.97	0.009	2.23	15.71
3	7 to 12 years	4.96	0.080	-0.59	10.51
3	13 to 17 years	-2.05	0.534	-8.53	4.42
4	< 4 years	-5.81	0.212	-14.94	3.32
4	4 to 6 years	2.78	0.581	-7.08	12.64
4	7 to 12 years	-0.52	0.894	-8.13	7.10
4	13 to 17 years	-9.72	0.034	-18.70	-0.74
5	< 4 years	-0.18	0.962	-7.69	7.32
5	4 to 6 years	10.59	0.013	2.27	18.90
5	7 to 12 years	5.55	0.099	-1.05	12.15
5	13 to 17 years	3.81	0.307	-3.50	11.11
6	< 4 years	0.43	0.905	-6.66	7.53
6	4 to 6 years	-1.43	0.705	-8.83	5.97
6	7 to 12 years	1.26	0.680	-4.74	7.26
6	13 to 17 years	-4.55	0.194	-11.42	2.32
7	< 4 years	4.07	0.375	-4.92	13.06
7	4 to 6 years	4.33	0.331	-4.40	13.07
7	7 to 12 years	7.29	0.040	0.35	14.23
7	13 to 17 years	3.78	0.339	-3.96	11.52
8	< 4 years	-3.24	0.598	-15.28	8.80
8	4 to 6 years	10.14	0.163	-4.11	24.40
8	7 to 12 years	-8.91	0.118	-20.08	2.26
8	13 to 17 years	-17.23	0.020	-31.74	-2.71
9	< 4 years	12.24	0.004	4.02	20.45
9	4 to 6 years	7.03	0.103	-1.41	15.48
9	7 to 12 years	9.31	0.013	1.97	16.65
9	13 to 17 years	0.10	0.981	-8.46	8.67

SCREENER	AGE-GROUP	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
10	< 4 years	-3.29	0.541	-13.84	7.26
10	4 to 6 years	-1.06	0.859	-12.78	10.67
10	7 to 12 years	-5.22	0.290	-14.91	4.46
10	13 to 17 years	-10.80	0.074	-22.66	1.07
11	< 4 years	9.06	0.078	-1.00	19.12
11	4 to 6 years	8.88	0.089	-1.37	19.13
11	7 to 12 years	2.18	0.658	-7.49	11.85
11	13 to 17 years	12.99	0.021	1.93	24.06

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on the entire sample for age-group, screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 14C: Consistency in workload, adjusted analysis, for 11 included call screeners, by race

SCREENER	RACE	DIFFERENCE IN PROBABILITY (POST-PRE-AFST)	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
1	white	11.69	0.001	4.51	18.87
1	Black/African American	3.45	0.345	-3.72	10.62
2	white	2.81	0.431	-4.18	9.81
2	Black/African American	1.31	0.713	-5.65	8.27
3	white	2.13	0.483	-3.83	8.09
3	Black/African American	3.22	0.305	-2.93	9.38
4	white	-3.47	0.458	-12.62	5.68
4	Black/African American	-4.45	0.292	-12.73	3.83
5	white	10.89	0.002	3.90	17.89
5	Black/African American	0.10	0.979	-7.31	7.51
6	white	-0.25	0.943	-7.08	6.59
6	Black/African American	-2.82	0.408	-9.49	3.86
7	white	12.78	0.001	5.29	20.26
7	Black/African American	-2.79	0.489	-10.69	5.11
8	white	0.29	0.971	-15.02	15.60
8	Black/African American	-7.24	0.223	-18.86	4.39
9	white	7.03	0.078	-0.79	14.84
9	Black/African American	9.62	0.022	1.41	17.82
10	white	-16.18	0.002	-26.35	-6.00
10	Black/African American	4.62	0.387	-5.84	15.08
11	white	6.16	0.236	-4.02	16.34
11	Black/African American	9.68	0.082	-1.21	20.57

Predicted probabilities are calculated using the coefficient estimated in multivariate regression analysis and predicting the outcome on the entire sample for race, screener and pre- or post-AFST, holding all else constant. The difference in probability is expressed in percentage points. Standard errors were clustered at the call-referral level.

TABLE 15A: **Means and variance of screener's predicted workload, adjusted analysis, for 11 included screeners**

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST	11	49.35	1.20	3.97	46.68	52.02
Post-AFST	11	51.15	1.06	3.52	48.79	53.51
Difference Pre - Post	11	-1.80	1.55	5.13	-5.24	1.65
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST) = Mean (post-AFST)					pvalue=	0.272
Levene's test of equality of variance between groups—testing variance of predicted margins (i.e. variance in level of outcome)						
Ho: Var (pre-AFST) = Var (post-AFST)					pvalue=	0.673

TABLE 15B: **Means and variance of screener's predicted workload, adjusted analysis, by age-group**

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST (< 4 years)	11	51.29	1.34	4.43	48.31	54.27
Post-AFST (< 4 years)	11	52.59	0.87	2.88	50.66	54.53
Pre-AFST (4 to 6 years)	11	48.24	1.61	5.34	44.65	51.82
Post-AFST (4 to 6 years)	11	53.19	1.09	3.63	50.75	55.62
Pre-AFST (7 to 12 years)	11	49.95	1.17	3.88	47.34	52.56
Post-AFST (7 to 12 years)	11	52.73	1.45	4.82	49.49	55.97
Pre-AFST (13 to 17 years)	11	47.41	2.07	6.87	42.79	52.03
Post-AFST (13 to 17 years)	11	45.95	1.80	5.96	41.95	49.96
Difference Pre–Post						
< 4 years	11	-1.31	1.62	5.36	-4.91	2.30
4 to 6 years	11	-4.95	1.35	4.49	-7.96	-1.94
7 to 12 years	11	-2.78	1.85	6.13	-6.90	1.34
13 to 17 years	11	1.46	2.69	8.91	-4.53	7.44
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST/< 4 years) = Mean (post-AFST/< 4 years)					p-value=	0.438
Ho: Mean (pre-AFST/4 to 6 years) = Mean (post-AFST/4 to 6 years)					p-value=	0.004
Ho: Mean (pre-AFST/7 to 12 years) = Mean (post-AFST/7 to 12years)					p-value=	0.164
Ho: Mean (pre-AFST/13 to 17 years) = Mean (post-AFST/13 to 17 years)					p-value=	0.600
Levene's test of equality of variance between groups						
Ho: Variance (pre-AFST/< 4 years) = Variance (post-AFST/< 4 years)					p-value =	0.226
Ho: Variance (pre-AFST/4 to 6 years) = Variance (post-AFST/4 to 6 years)					p-value =	0.277
Ho: Variance (pre-AFST/7 to 12 years) =Variance (post-AFST/7 to 12years)					p-value =	0.578
Ho: Variance (pre-AFST/13 to 17 years) = Variance (post-AFST/13 to 17 years)					p-value =	0.616

TABLE 15C: **Means and variance of screener's predicted workload, adjusted analysis, by race**

	N	MEAN	STD. ERR.	STD. DEV.	[95% CI]	
Pre-AFST (white)	11	49.36	1.77	5.87	45.41	53.30
Post-AFST (white)	11	52.44	1.40	4.65	49.32	55.56
Pre-AFST(Black)	11	49.53	1.76	5.85	45.60	53.45
Post-AFST(Black)	11	50.86	0.99	3.27	48.67	53.06
Difference Pre–Post						
White	11	-3.08	2.50	8.29	-8.65	2.49
Black/African American	11	-1.34	1.65	5.47	-5.01	2.34
T-tests of means of predicted outcomes for screeners						
Ho: Mean (pre-AFST/white) = Mean (post-AFST/white)					p-value=	0.246
Ho: Mean (pre-AFST/Black) = Mean (post-AFST/Black)					p-value=	0.437
Levene's test of equality of variance between groups						
Ho: Variance(pre-AFST/white) = Variance(post-AFST/white)					p-value=	0.445
Ho: Variance(pre-AFST/Black) = Variance(post-AFST/Black)					p-value=	0.127

TABLE 16A: **Estimated magnitude of monthly impact of AFST on accuracy of screen-in**

	ESTIMATED TOTAL # OF CHILDREN WITH ACCURATE SCREEN-IN PER MONTH						ESTIMATED # OF CHILDREN IMPACTED BY AFST PER MONTH		
	PRE-AFST			POST-AFST			(POST-AFS–PRE-AFST)		
	N	LOWER BOUND	UPPER BOUND	N	LOWER BOUND	UPPER BOUND	N	LOWER BOUND	UPPER BOUND
all	**358**	**346**	**370**	**381**	**369**	**394**	**24**	**8**	**39**
< 4 years	93	89	98	93	89	97	0	-5	4
4–6 years	61	58	65	66	62	69	4	-1	9
7–12 years	123	118	128	136	131	141	13	6	20
13–17 years	83	79	86	89	86	93	7	1	13
white	135	129	142	156	148	164	20	13	27
black	201	191	212	201	193	209	0	-14	14

Estimates are based on predicted probabilities of accuracy of screen-in (Tables 4) and mean number of children screened-In per month over entire analysis period. The total average number of children screened-in per month over the entire analysis period is 819 with 26%, 17%, 34% & 23% for age groups < 4, 4-6, 7 - 12, and 13 - 17 years respectively and 38%, 53% for white and Black/African American, respectively.

TABLE 16B: Estimated magnitude of impact of AFST on accuracy of screen-out

	ESTIMATED TOTAL # CHILDREN WITH ACCURATE SCREEN-OUT PER MONTH						ESTIMATED #OF CHILDREN IMPACTED BY AFST PER MONTH		
	PRE-AFST			POST-AFST			(POST-AFST–PRE-AFST)		
	N	LOWER BOUND	UPPER BOUND	N	LOWER BOUND	UPPER BOUND	N	LOWER BOUND	UPPER BOUND
all	774	768	780	763	754	773	-11	-22	1
< 4 years	149	146	151	146	143	150	-2	-7	2
4–6 years	136	134	138	132	129	135	-4	-8	0
7–12 years	287	285	290	283	278	288	-4	-11	2
13–17 years	209	207	212	209	205	212	-1	-5	4
white	340	335	344	336	328	343	-4	-12	4
black	353	348	358	345	339	352	-8	-16	1

Estimates are based on predicted probabilities of accuracy of screen-out (Tables 6) and mean number of children screened-out per month over entire analysis period. The total average number of children screened-out per month over the entire analysis period is 914 with 19%, 17%, 37% & 27% for age groups < 4, 4-6, 7 - 12, and 13 - 17 years respectively and 45%, 46% for white and Black/African American, respectively.

TABLE 16C: Estimated magnitude of impact of AFST on workload

	ESTIMATED TOTAL # OF CHILDREN SCREENED-IN PER MONTH						ESTIMATED # OF CHILDREN IMPACTED BY AFST PER MONTH		
	PRE-AFST			POST-AFST			(POST-AFST–PRE-AFST)		
	N	LOWER BOUND	UPPER BOUND	N	LOWER BOUND	UPPER BOUND	N	LOWER BOUND	UPPER BOUND
all	**851**	**817**	**884**	**797**	**745**	**850**	**-53**	**-113**	**6**
< 4 years	198	188	207	191	182	200	-7	-18	4
4–6 years	145	138	152	144	134	154	-1	-11	10
7–12 years	307	295	320	289	267	310	-19	-44	6
13–17 years	208	199	218	181	166	195	-28	-46	-9
white	356	339	373	342	320	364	-13	-41	15
black	430	410	451	396	369	424	-34	-66	-2

Estimates are based on predicted probabilities of workload (Tables 8) and mean number of children in referrals per month over entire analysis period. The total average number of children in referrals per month over the entire analysis period is 1745 with 22%, 17%, 36% & 25% for age groups < 4, 4-6, 7 - 12, and 13 - 17 years respectively and 42%, 49% for white and Black/African American, respectively.

FIGURES

FIGURE 1: **Example of the AFST Score**

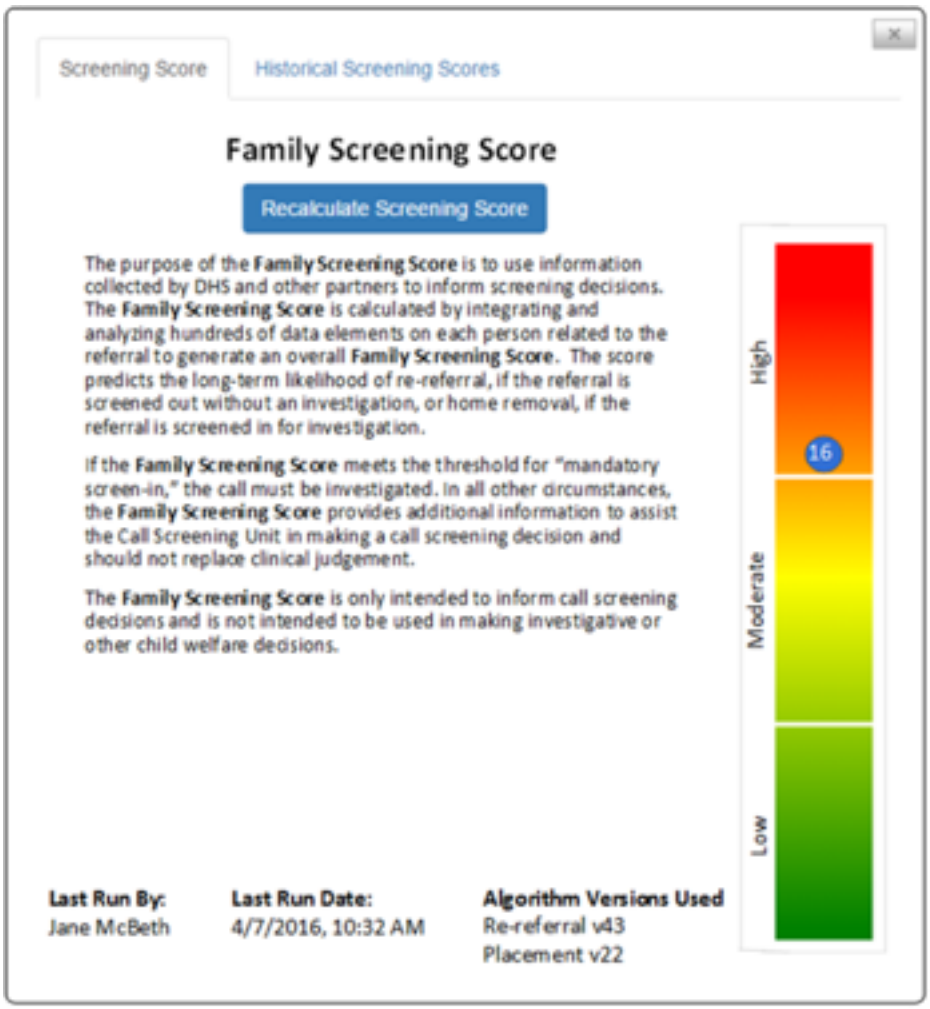
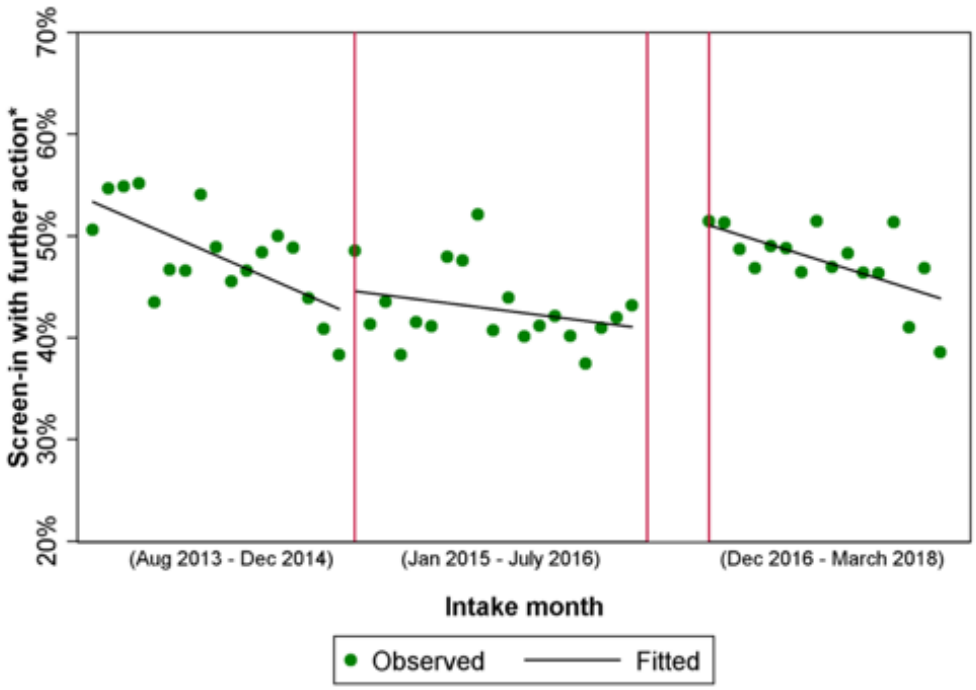
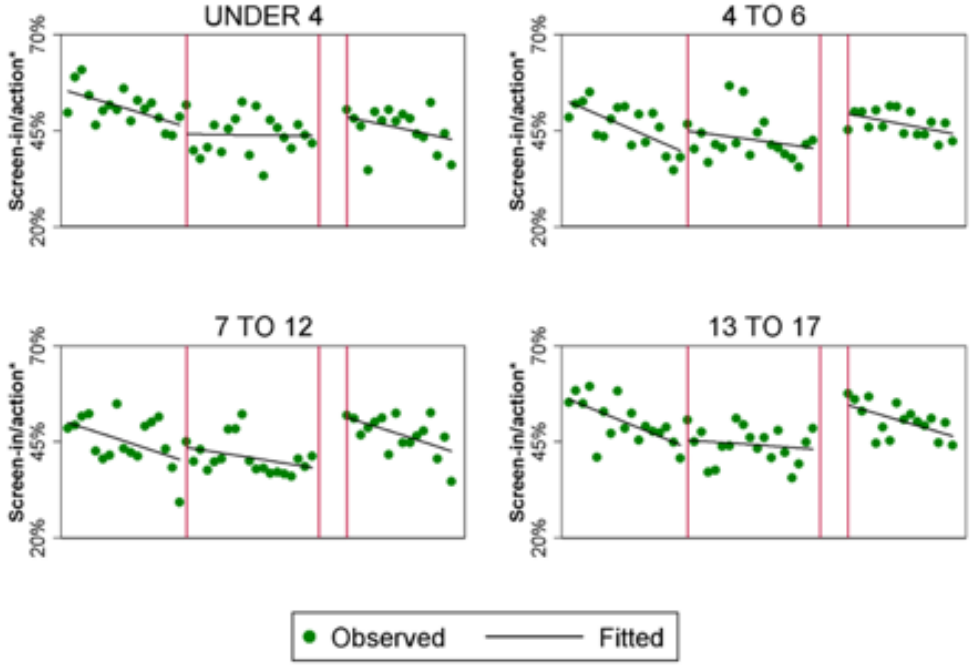


FIGURE 2A: **Accuracy of Screen-In, ITSA analysis**



*or Screen-in w/out further action & re-referral within 60 days

FIGURE 2B: **Accuracy of Screen-In, by age-group**



*or Screen-in w/out further action & re-referral within 60 days

FIGURE 2C: **Accuracy of Screen-In, by race**

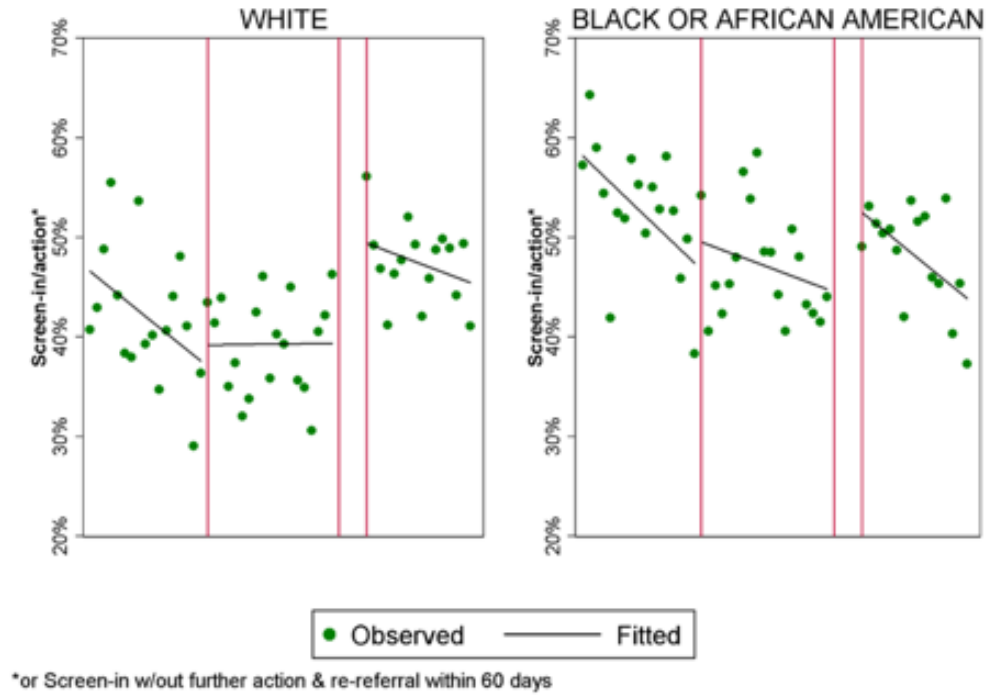


FIGURE 3A: **Accuracy of Screen-In, adjusted analysis**

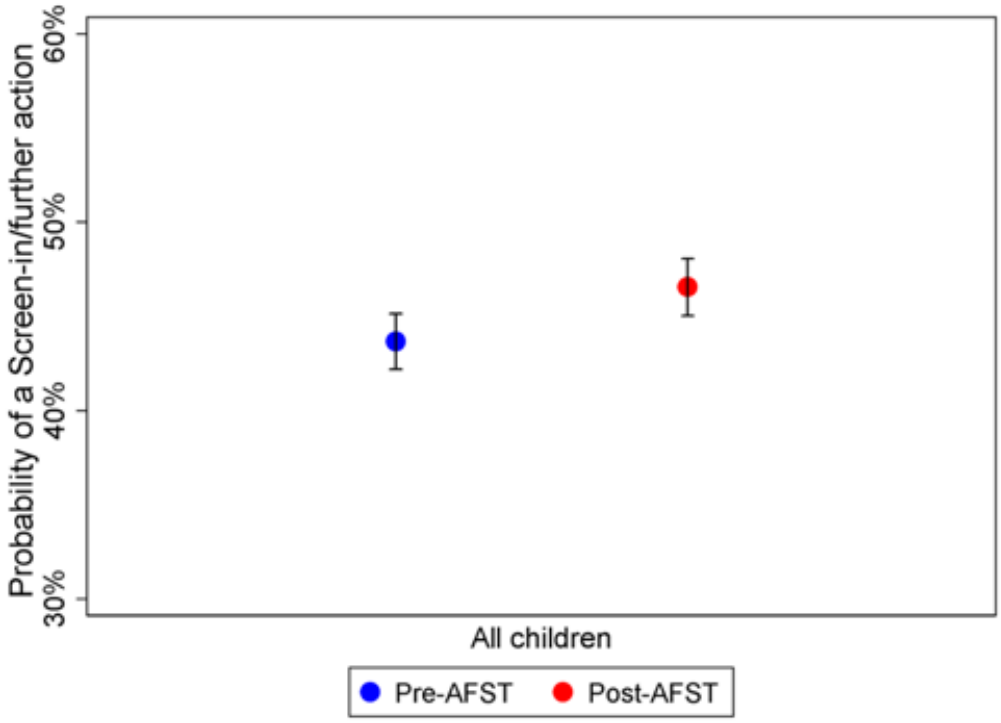


FIGURE 3B: **Accuracy of Screen-In, adjusted analysis, by age-groups**

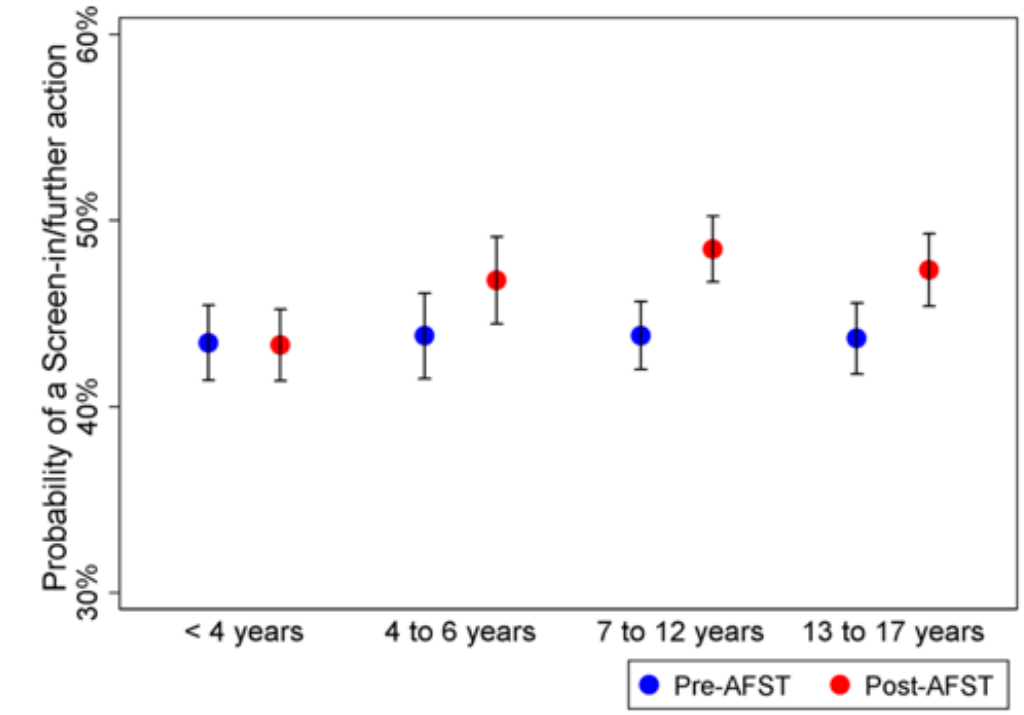


FIGURE 3C: **Accuracy of Screen-In, adjusted analysis, by race**

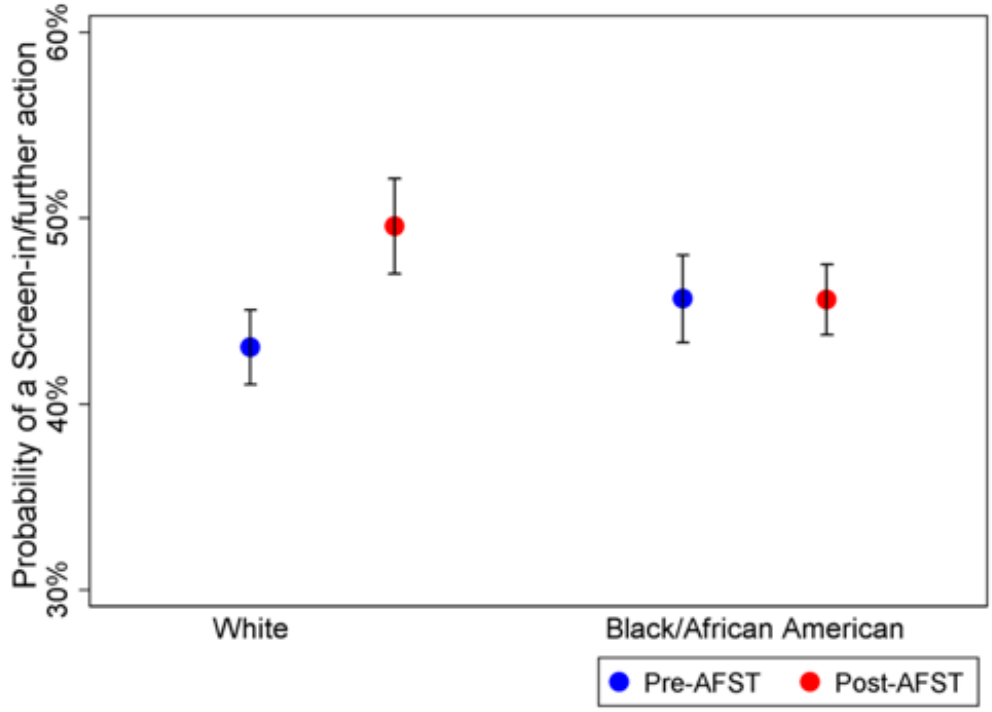


FIGURE 4A: **Accuracy of Screen-Out, ITSA analysis**

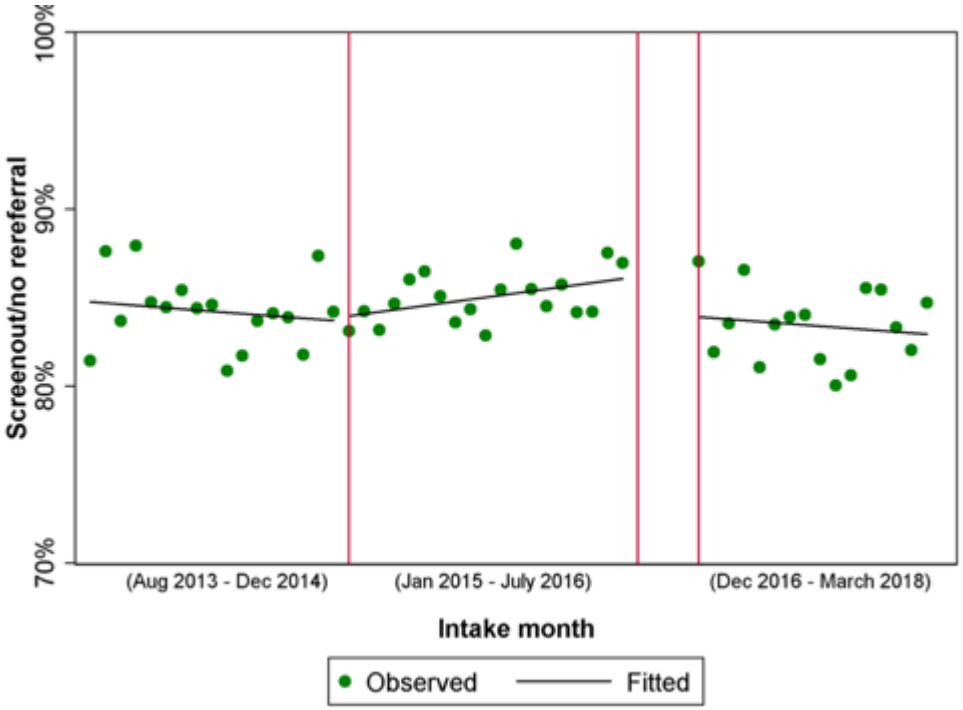


FIGURE 4B: **Accuracy of Screen-Out, ITSA analysis, by age-group**

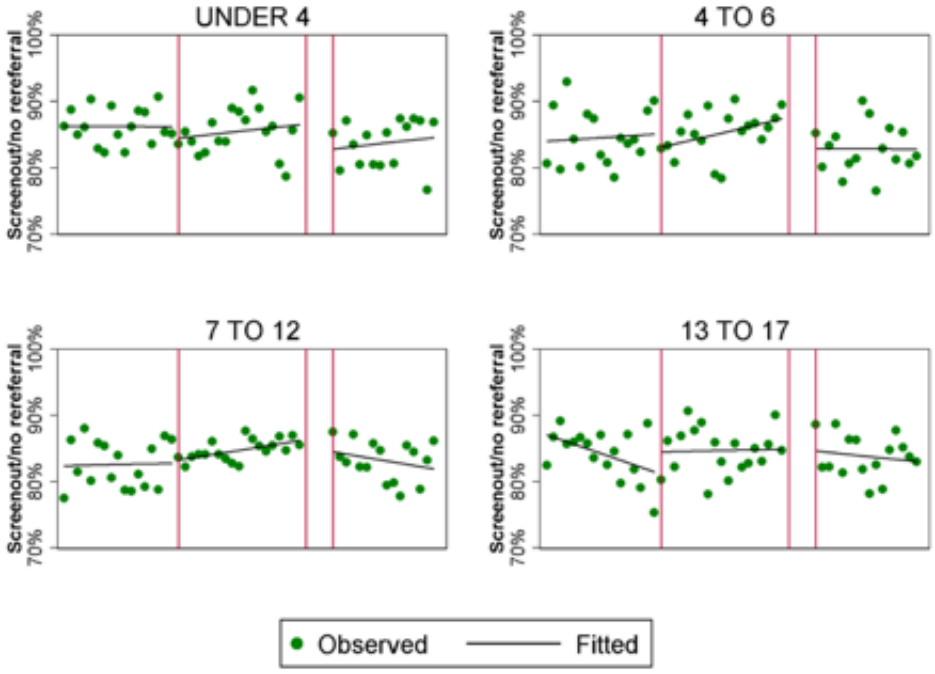


FIGURE 4C: **Accuracy of Screen-Out, ITSA analysis, by race**

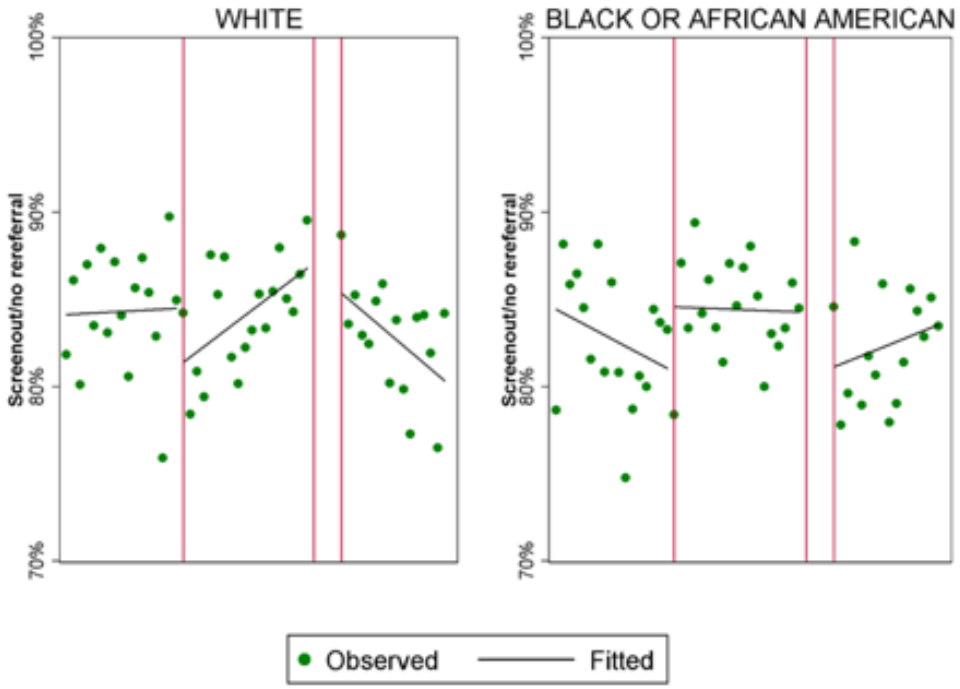


FIGURE 5A: **Accuracy of Screen-Out, adjusted analysis**

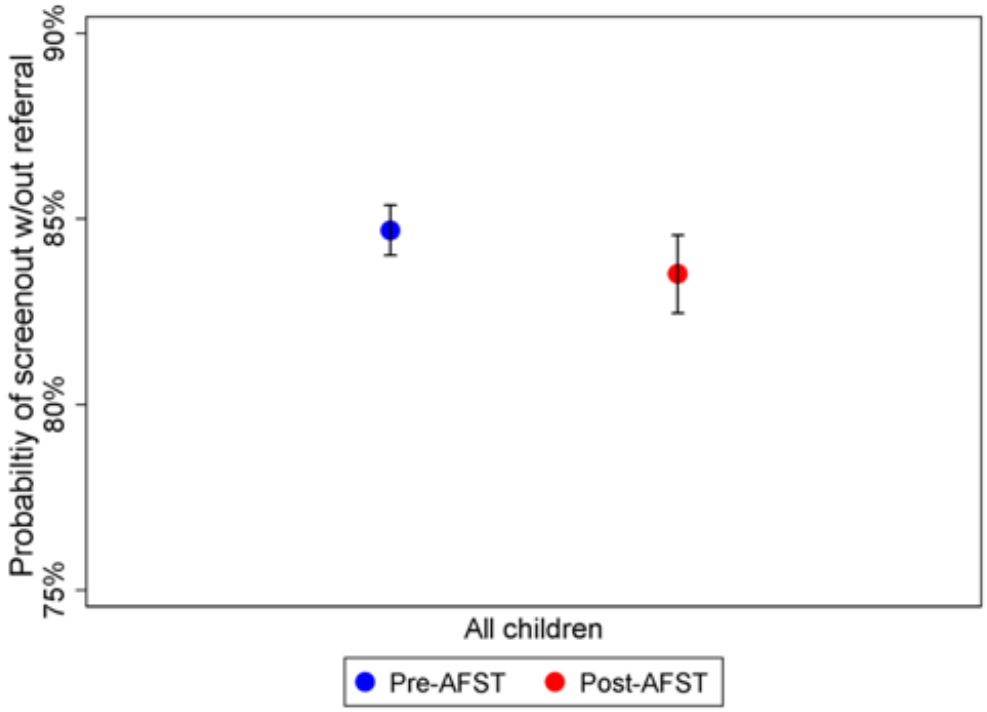


FIGURE 5B: **Accuracy of Screen-Out, adjusted analysis, by age-group**

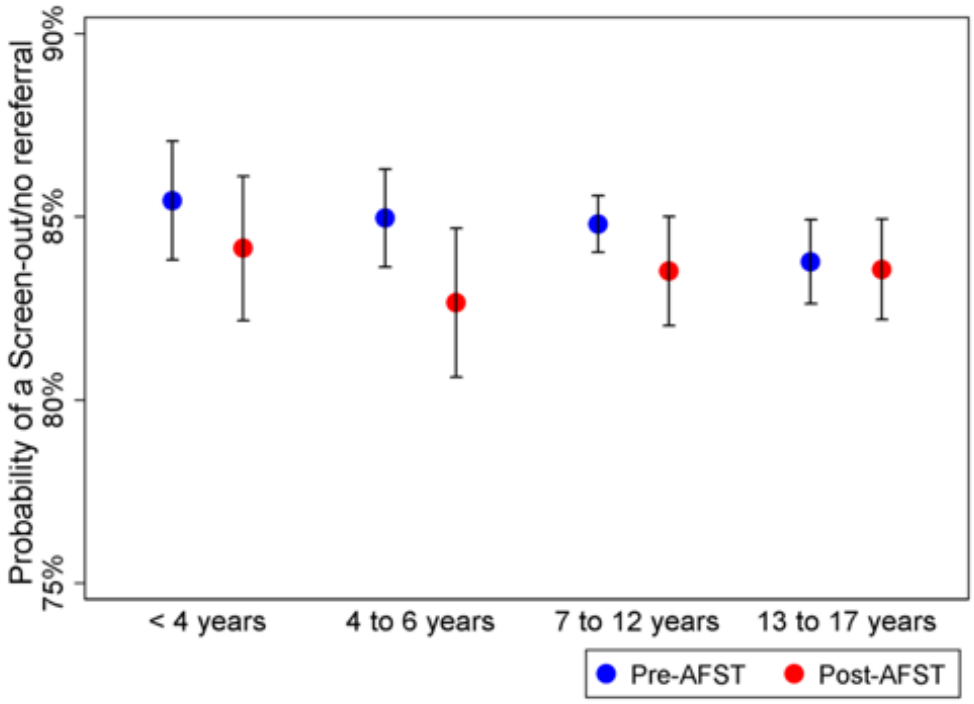


FIGURE 5C: **Accuracy of Screen-Out, adjusted analysis, by race**

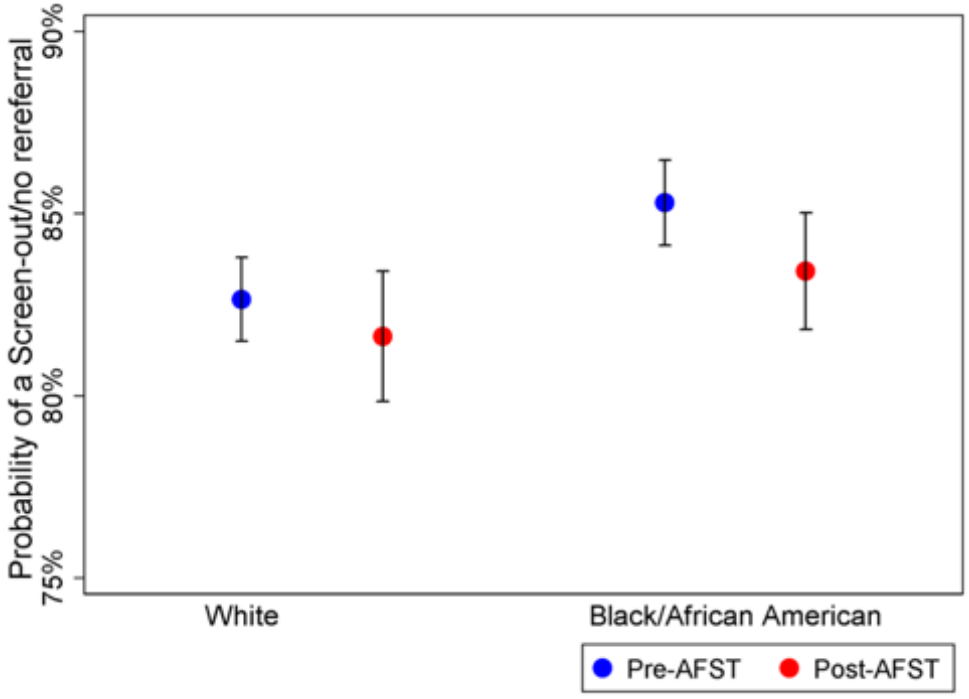


FIGURE 6A: **Workload, ITSA Analysis**

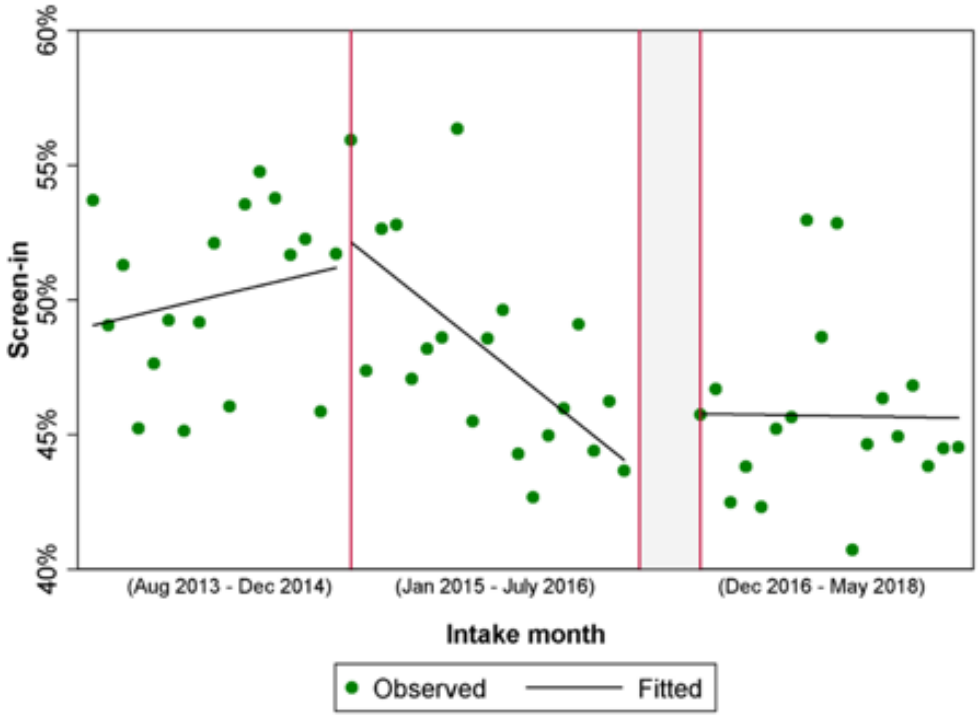


FIGURE 6B: **Workload, ITSA Analysis, by age-group**

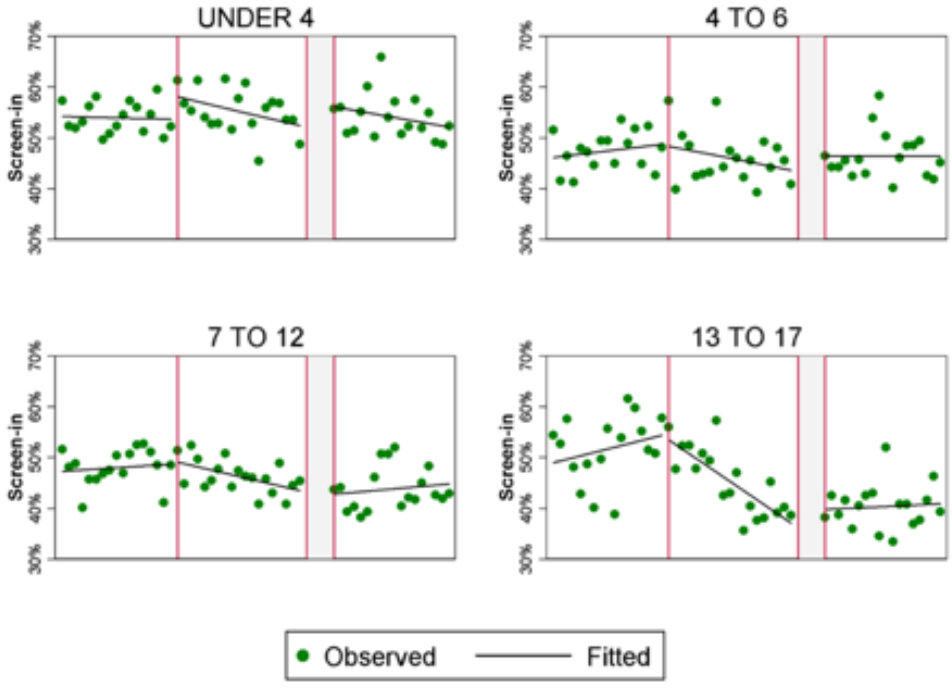


FIGURE 6C: **Workload, ITSA Analysis, by race**

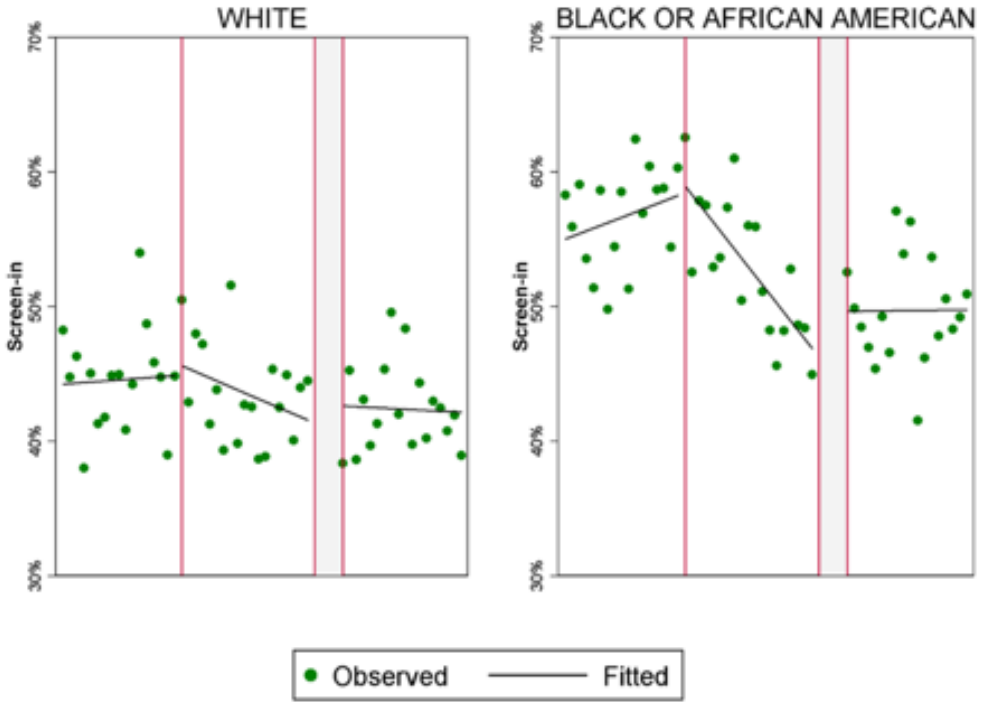


FIGURE 7A: **Workload, adjusted analysis**

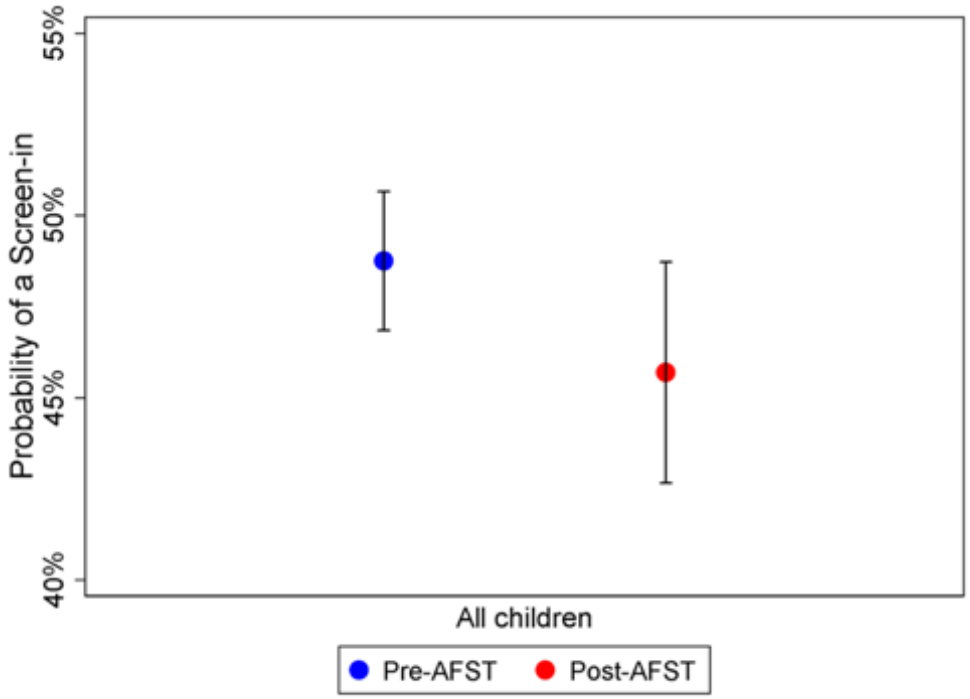


FIGURE 7B: **Workload, adjusted analysis, by age-group**

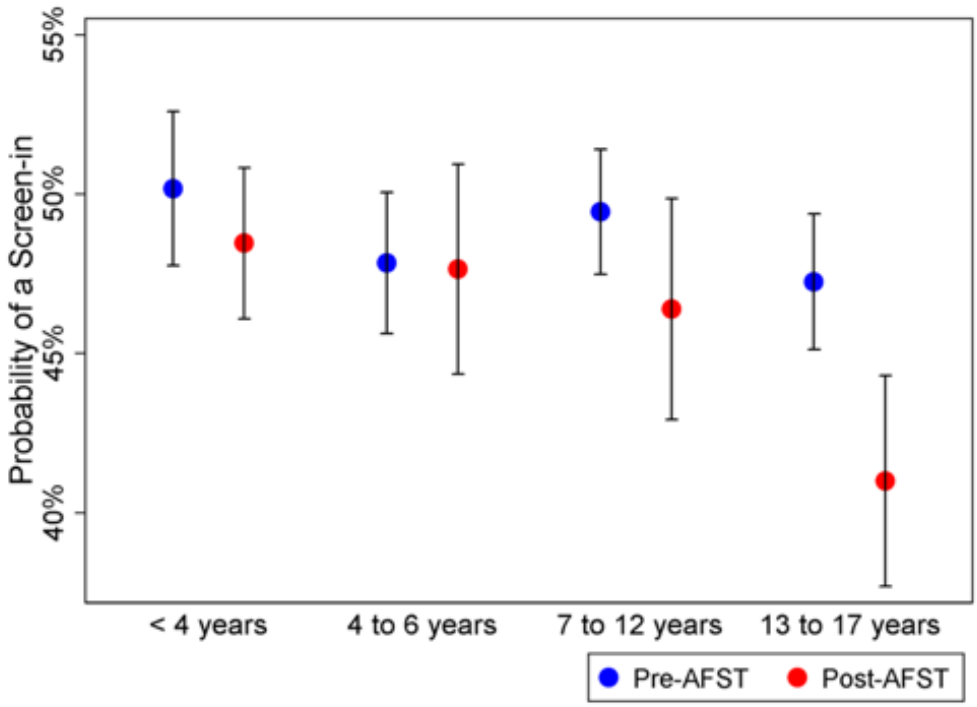


FIGURE 7C: **Workload, adjusted analysis, by race**

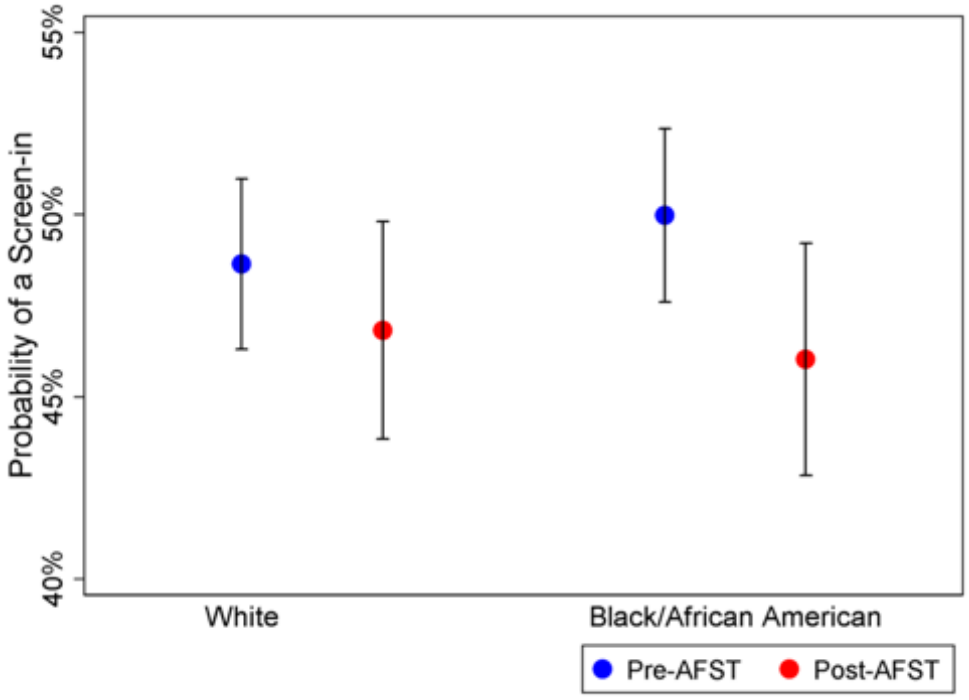


FIGURE 8A: **Consistency of accuracy of screen-in for 11 call screeners, adjusted analysis**

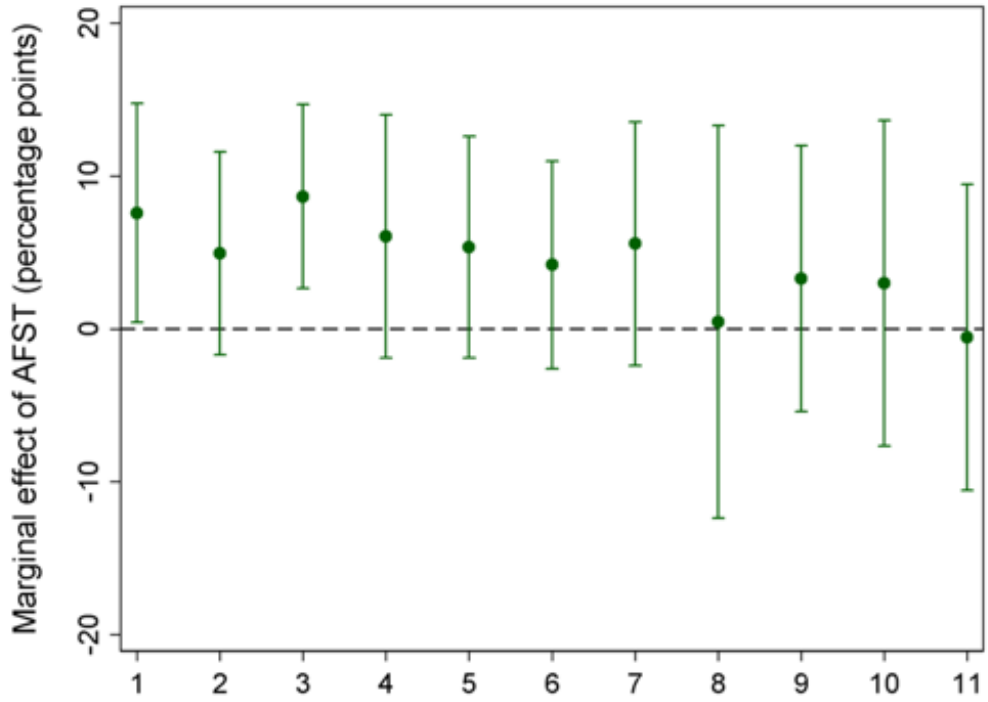


FIGURE 8B: **Consistency of accuracy of screen-in for 11 call screeners, adjusted analysis, by age-group**

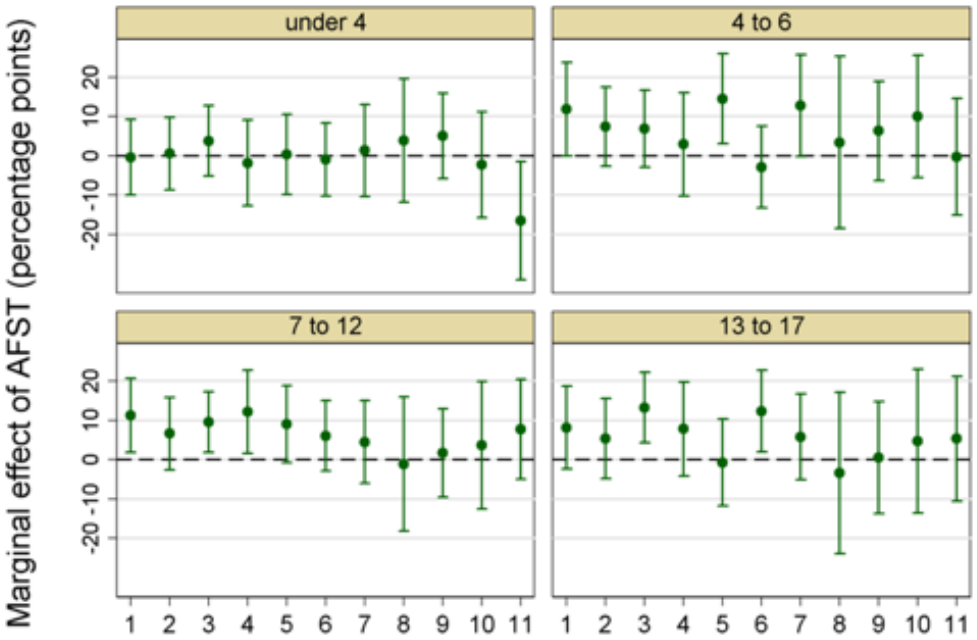


FIGURE 8C: **Consistency of accuracy of screen-in for 11 call screeners, adjusted analysis, by race**

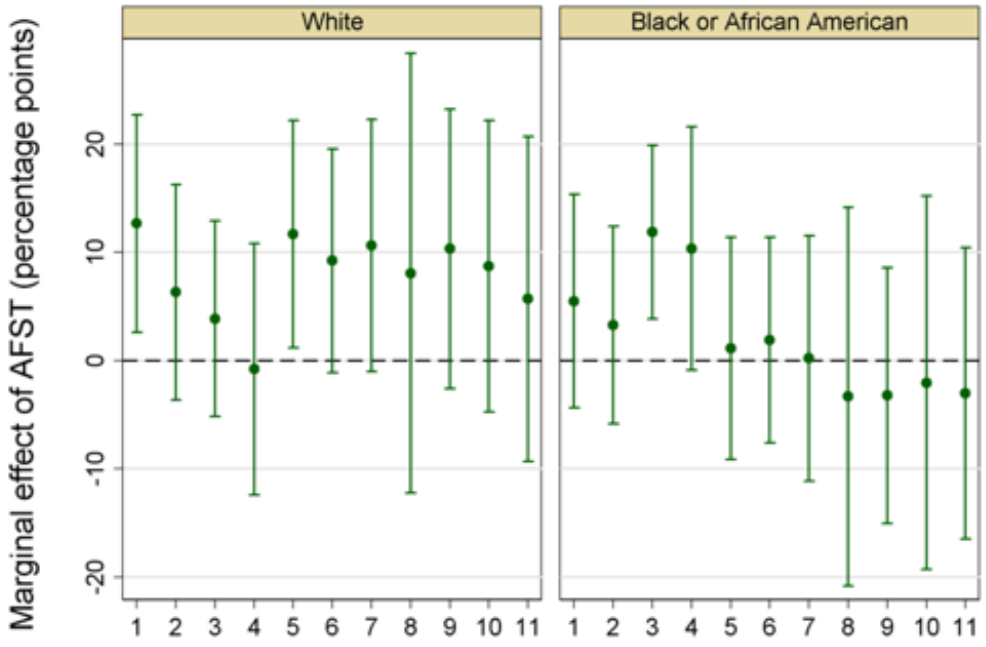


FIGURE 9A: **Consistency of accuracy of screen-out for 11 call screeners, adjusted analysis**

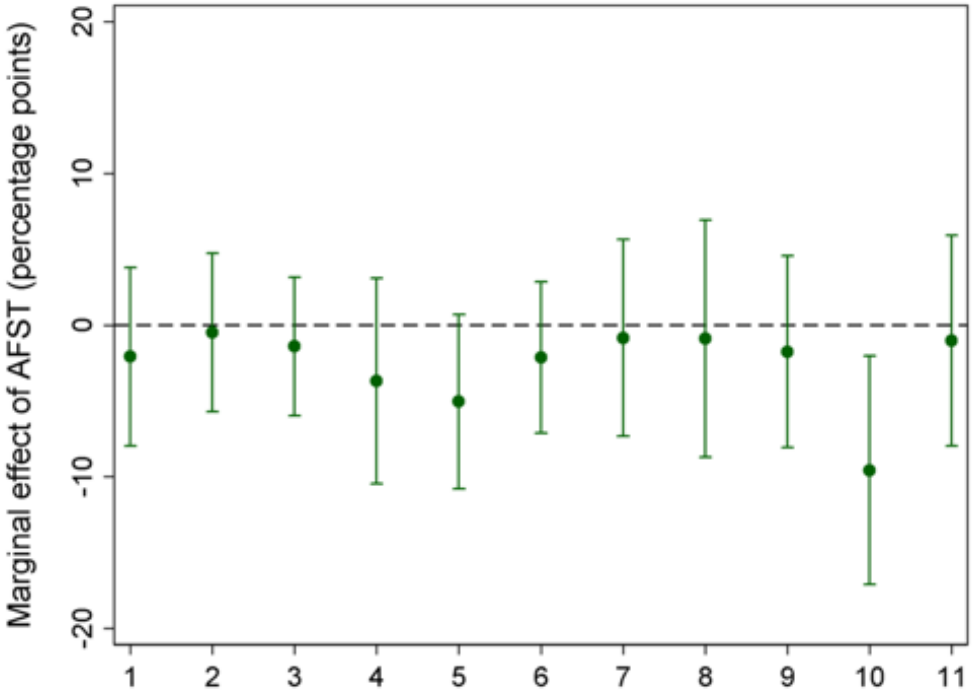


FIGURE 9B: **Consistency of accuracy of screen-out for 11 call screeners, adjusted analysis, by age-group**

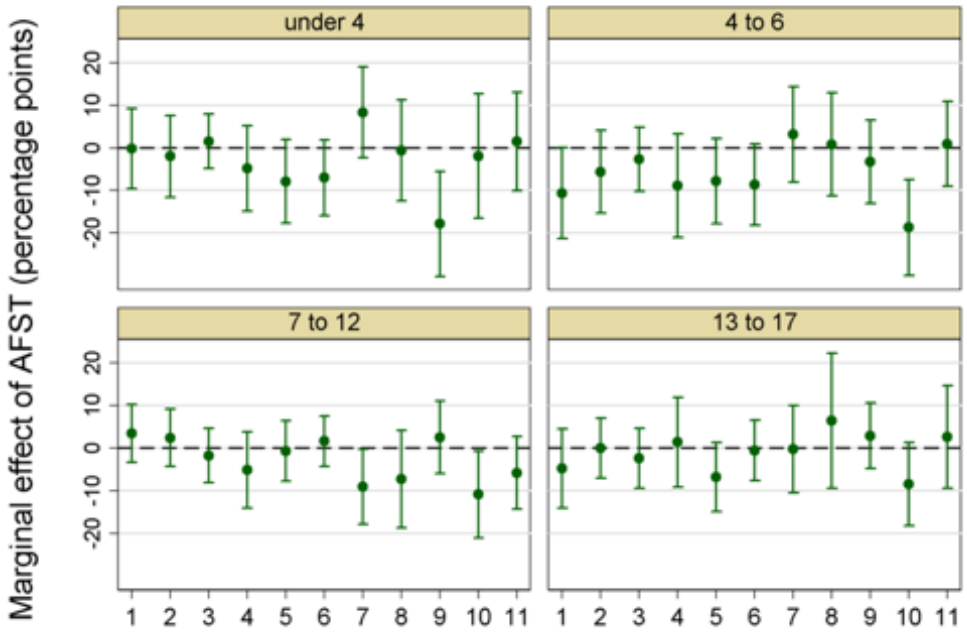


FIGURE 9C: **Consistency of accuracy of screen-out for 11 call screeners, adjusted analysis, by race**

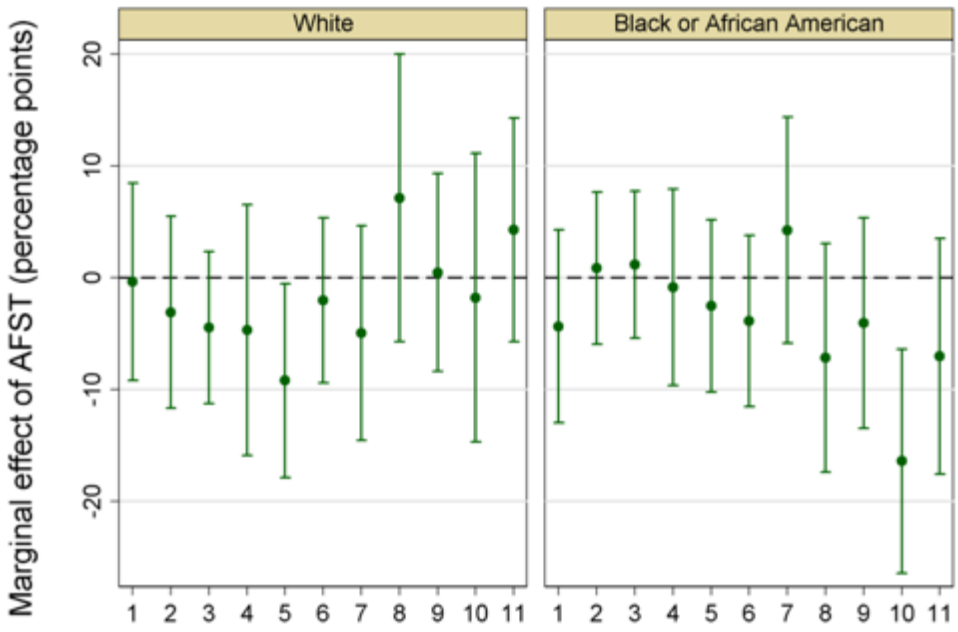


FIGURE 10A: **Consistency of workload for 11 call screeners, adjusted analysis**

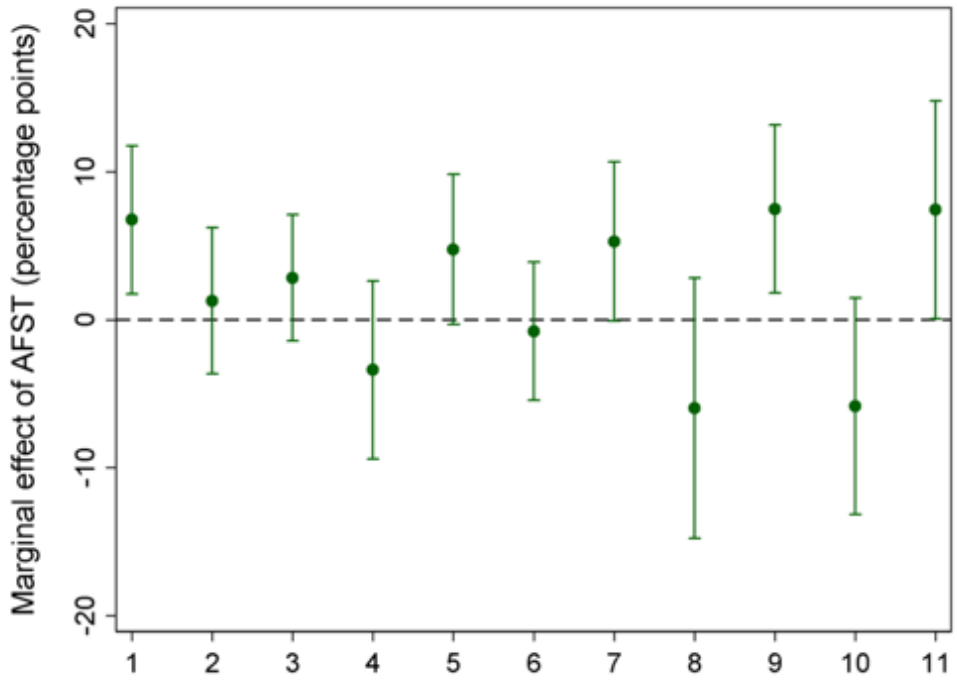


FIGURE 10B: **Consistency of workload for 11 call screeners, adjusted analysis, by age-group**

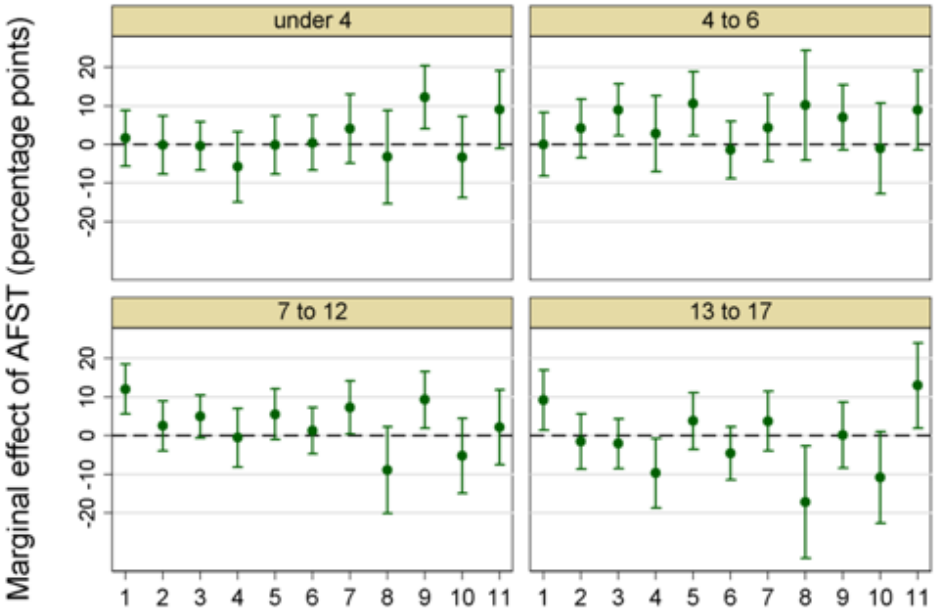
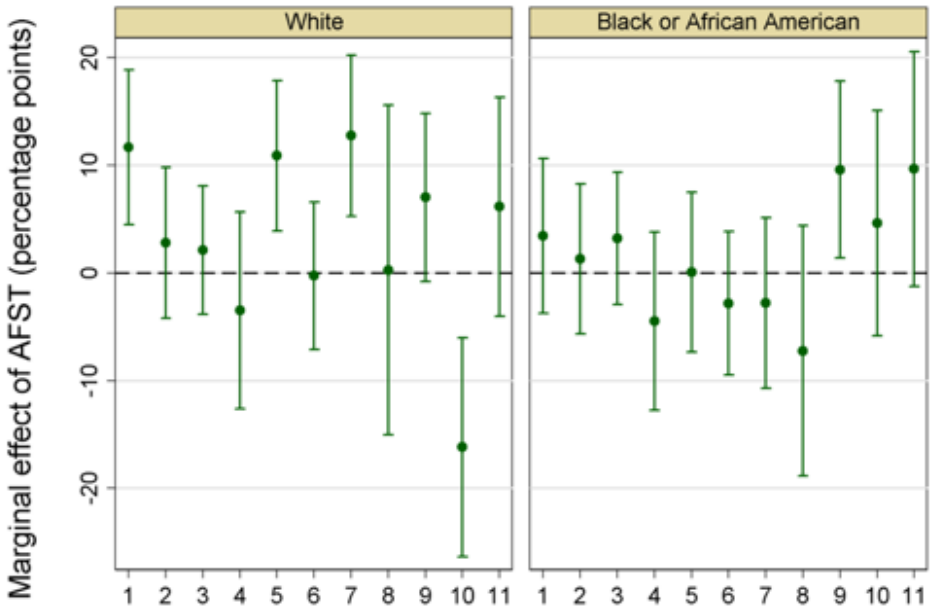


FIGURE 10C: **Consistency of workload for 11 call screeners, adjusted analysis, by race**



APPENDIX A1: ANALYTIC DATASET AND VARIABLE CONSTRUCTION

Construction of the outcome variables

The screen-in variable is constructed from the Referrals data, using the variable "call_scrn_ outcome". A child was coded as "screen-out" if the call-screen-outcome contained the words "screen" and "out" (after accounting for case-sensitivity). A "screen-in" was any case which was not a "screen-out" and which did not have missing information for the call-screen-outcome. Referrals for the entire post-AFST period (December 2016 – May 2018) were included in this outcome.

A **screen-in with further action upon investigation or a screen-in with no further action and a re-referral within a 2-month window** was constructed for children identified as "screen-in" and for whom a service decision was available (i.e. not missing). "Further action" status was given to children with a service decision other than "Do Not Accept for Service", regardless of whether the case was connected to an open or closed case. Referrals for a truncated post-AFST period (December 2016–March 2018) and rereferrals for the entire post-AFST period (December 2016–May 2018) were included in this outcome.

A **screen-out with no re-referral within a 2-month window** was constructed for children identified as "screen-out". For each child, a referral was considered the "index event"" if it was not within 60 days of a previous index referral, or if it was the first time a child entered the dataset. A re-referral was any subsequent call within 60 days of the initial referral date, regardless of outcome or service decision dates. Although we account for re-referrals occurring in the months of April and May 2018, we do not include index events occurring as of April 2018. Notably, while index referrals were for GPS calls, re-referral could be for either CPS or GPS calls. Referrals for a truncated post-AFST period (December 2016 – March 2018) and rereferrals for the entire post-AFST period (December 2016 – May 2018) were included in this outcome.

Exclusions in the analytic dataset

All children in all referrals were included in the primary analytic dataset with the following exceptions. Children > 17 years of age at the time of the referral were excluded (although we account for 18-year-old children in re-referral calls). Children in any CPS referral were excluded from the analytic dataset, as there is no variation in screening-decision for these children. Any referral which had a call screen outcome (variable call_scrn_outcome) coded as "Accept: Actively working with this family" was excluded, as were those with call screen outcome coded as "Assessment Completed on Active Family". The latter two exclusions were at the recommendation of Allegheny County's analysts to perform analyses on data consistent with Allegheny's in-house analyses.

Construction of control variables

Child characteristics include age in years (grouped into categories as under 4, 4–6 years, 7–12 years, 13–17 years), legal sex category (male, female, or undetermined), race category (Black / African American, white, other, unable to determine). A child was considered Black / African

Appendix (continued)

American if their race was coded as such, or any combination of another race and Black / African American and white if the child was coded as "white", and not a mixed race. All other children fall into either the other category (race was specified, but was not black/African American or white, or unable to determine).

Household characteristics include the composition of household members or the number of other people in the referral who fall between specified age ranges (e.g. <1, 1–5, 6–12, 13–17, adult parents, other adults), the mean age in years of adults in a referral (18 to <30; 30 to <50, 50 to <66, 66+). Household characteristics also included a proxy measure for socioeconomic status. This measure was designed to be consistent with the measure used as an input to the AFST. Specifically, the Zip Code in which the household was located was coded in terms of the fraction of residents falling below the federal poverty line based on the American Community Survey (2008-2012). The constructed socioeconomic status variable has five categories: living in areas where 1) 0 to <10%; 2) 10 to <15%; or 3) 15 to <20%; 4) 20% to <25%; or 5) 25%+ of households fall below the federal poverty line. As an indicator of the risk that any referred child faces, we use a maximum risk score category (low, medium, high, mandatory risk) for the household. Maximum risk score is based on the maximum of the binned risk scores for the placement and the re-referral score, based on cutoffs as determined by Allegheny. The risk score used as a control in regression analyses was not the AFST risk score shown to the call screeners in the Post-AFST Period, but rather the risk score, exactly comparable to that constructed for the Pre-AFST Period.

APPENDIX A2: NOTES ON INTERRUPTED TIME-SERIES (ITSA)

ITSA is estimated as an autoregressive model, to account for the form of correlation between observations. For example, observations which occur within a closer timeframe may be more correlated than observations further apart in time. This type of pattern could reflect secular trends or seasonal patterns. Traditionally, there are two general approaches to account for autocorrelation in ITSA, the autoregressive integrated moving-average (ARIMA) models and ordinary least-squares (OLS), with adjustments for autocorrelation. We utilize the itsa command in Stata (v14) which relies on OLS, due to its more flexible and more broadly applicable nature (1–3). We assume that the error term follows an autoregressive process:

$$e_t = pe_{t-n} + u_t$$

Where p is the correlation between error terms that are adjacent in time and the remaining disturbances, u_t, are independent.

We can specify the maximum number of lags in Stata, as part of the ITSA command and test for the correctness of this specification using actest which performs the Cumby-Huizinga general specification test of serial correlation.

Causal inference based on ITSA requires several assumptions:

Assumption 1: Outcomes (levels/trends) remain unchanged in the absences of the program.

Assumption 2: Relative to rapid rate of change in outcomes attributed to the abrupt implementation of the policy of interest, all unobserved time-varying variables change slowly, such that their impact on outcomes would be distinguishable.

Assumption 3: There are no other policies/changes that occur at or around the same time "as the AFST implementation that would impact outcomes substantially.

Assumption 4: Full implementation of the AFST occurs at a discrete point in time.

Assumption 5: The AFST did not materially alter the collection of data on outcomes or covariates or the quality of the data collected.

REFERENCES

1. Box GEP, and Jenkins GM. Time series analysis : forecasting and control. Rev. ed. San Francisco; London: Holden-Day; 1976. xxi, 575 p. p.
2. Linden A. Conducting interrupted time-series analysis for single- and multiple-group comparisons. The Stata Journal. 2015;15(2):280-500.
3. Velicer WF, and Harrop J. The reliability and accuracy of time series model identification. Evaluation Review. 1983;7:551 - 60.

Appendix (continued)

APPENDIX TABLES

TABLE A1A: Accuracy of Screen-in, ITSA analysis, all children, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	63.92%	0.000	61.34%	66.49%
Pre-2014 Policy	Trend	-0.58	0.002	-0.93	-0.23
2014 Policy	Change in level	1.38	0.670	-5.13	7.90
Post 2014 policy, pre-AFST	Change in trend	0.53	0.021	0.09	0.98
AFST implementation	Change in level	6.27	0.001	2.75	9.79
Post-AFST	Change in trend	-0.05	0.835	-0.50	0.40
Total trend in screen-in rates pre-AFST		-0.05	0.771	-0.38	0.29
Total trend in screen-in rate post-AFST		-0.10	0.476	-0.36	0.17

TABLE A1B: Accuracy of Screen-in, ITSA analysis, < 4 years old, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	63.86%	0.000	58.98%	68.75%
Pre-2014 Policy	Trend	-0.50	0.050	-1.00	0.00
2014 Policy	Change in level	-2.77	0.452	-10.11	4.58
Post 2014 policy, pre-AFST	Change in trend	0.73	0.031	0.07	1.39
AFST implementation	Change in level	0.18	0.956	-6.33	6.69
Post-AFST	Change in trend	-0.05	0.890	-0.76	0.66
Total trend in screen-in rates pre-AFST		0.23	0.287	-0.20	0.66
Total trend in screen-in rate post-AFST		0.18	0.526	-0.39	0.75

Note: change in trend is expressed in percentage points/month.

Appendix (continued)

TABLE A1C: Accuracy of Screen-in, ITSA analysis, 4 to 6 years old, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	67.80%	0.000	64.10%	71.50%
Pre-2014 Policy	Trend	-0.97	0.000	-1.40	-0.54
2014 Policy	Change in level	7.33	0.025	0.96	13.70
Post 2014 policy, pre-AFST	Change in trend	0.72	0.006	0.22	1.23
AFST implementation	Change in level	6.55	0.005	2.10	11.01
Post-AFST	Change in trend	0.14	0.587	-0.38	0.67
Total trend in screen-in rates pre-AFST		-0.25	0.077	-0.52	0.03
Total trend in screen-in rate post-AFST		-0.10	0.647	-0.55	0.35

Note: change in trend is expressed in percentage points/month.

TABLE A1D: Accuracy of Screen-in, ITSA analysis, 7 to 12 years old, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	60.47%	0.000	55.70%	65.24%
Pre-2014 Policy	Trend	-0.37	0.210	-0.96	0.22
2014 Policy	Change in level	1.29	0.757	-7.04	9.62
Post 2014 policy, pre-AFST	Change in trend	0.19	0.580	-0.50	0.88
AFST implementation	Change in level	8.77	0.001	3.64	13.89
Post-AFST	Change in trend	0.07	0.855	-0.67	0.80
Total trend in screen-in rates pre-AFST		-0.18	0.319	-0.54	0.18
Total trend in screen-in rate post-AFST		-0.11	0.720	-0.75	0.52

Note: change in trend is expressed in percentage points/month.

Appendix (continued)

TABLE A1E: Accuracy of Screen-in, ITSA analysis, 13 to 17 years old, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	65.64%	0.000	61.16%	70.13%
Pre-2014 Policy	Trend	-0.64	0.000	-0.98	-0.31
2014 Policy	Change in level	1.25	0.662	-4.48	6.98
Post 2014 policy, pre-AFST	Change in trend	0.65	0.024	0.09	1.21
AFST implementation	Change in level	9.56	0.001	3.92	15.20
Post-AFST	Change in trend	-0.41	0.205	-1.06	0.23
Total trend in screen-in rates pre-AFST		0.01	0.982	-0.44	0.45
Total trend in screen-in rate post-AFST		-0.41	0.087	-0.88	0.06

Note: change in trend is expressed in percentage points/month.

TABLE A1F: Accuracy of Screen-in, ITSA analysis, White, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	47.28%	0.000	41.03%	53.52%
Pre-2014 Policy	Trend	-0.57	0.061	-1.17	0.03
2014 Policy	Change in level	2.33	0.535	-5.17	9.82
Post 2014 policy, pre-AFST	Change in trend	0.57	0.129	-0.17	1.32
AFST implementation	Change in level	10.12	0.006	3.13	17.12
Post-AFST	Change in trend	-0.27	0.414	-0.92	0.39
Total trend in screen-in rates pre-AFST		0.00	0.991	-0.44	0.45
Total trend in screen-in rate post-AFST		-0.27	0.272	-0.75	0.22

Note: hange in trend is expressed in percentage points/month.

Appendix (continued)

TABLE A1G: Accuracy of Screen-in, ITSA analysis, Black/African American, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	59.46%	0.000	53.37%	65.56%
Pre-2014 Policy	Trend	-0.72	0.018	-1.31	-0.13
2014 Policy	Change in level	2.91	0.468	-5.09	10.90
Post 2014 policy, pre-AFST	Change in trend	0.49	0.180	-0.23	1.21
AFST implementation	Change in level	7.71	0.002	3.01	12.40
Post-AFST	Change in trend	-0.36	0.257	-0.99	0.27
Total trend in screen-in rates pre-AFST		-0.23	0.273	-0.65	0.19
Total trend in screen-in rate post-AFST		-0.59	0.016	-1.06	-0.11

Note: change in trend is expressed in percentage points/month.

TABLE A2A: Accuracy of screen-in, adjusted analysis, all children, 6-month re-referral window

	PREDICTED PROBABILITY OF A SCREEN-IN WITH FURTHER ACTION	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST	54.96%	0.000	53.85%	56.07%
Post-AFST	58.78%	0.000	57.36%	60.20%
DIFF (Post - Pre)	3.82%	0.000	2.15%	5.49%

TABLE A2B: Accuracy of screen-in, adjusted analysis, by age group, 6-month re-referral window

	PREDICTED PROBABILITY OF A SCREEN-IN WITH FURTHER ACTION	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST				
< 4 years	54.55%	0.000	52.98%	56.12%
4–6 years	55.70%	0.000	53.85%	57.55%
7–12 years	54.99%	0.000	53.40%	56.57%
13–17 years	54.86%	0.000	53.15%	56.58%
Post-AFST				
< 4 years	56.22%	0.000	54.03%	58.42%
4–6 years	58.79%	0.000	56.59%	60.98%
7–12 years	60.25%	0.000	58.71%	61.79%
13–17 years	59.57%	0.000	57.04%	62.10%
Difference Post–Pre				
< 4 years	1.67%	0.166	-0.69%	4.04%
4–6 years	3.09%	0.019	0.51%	5.68%
7–12 years	5.27%	0.000	3.00%	7.54%
13–17 years	4.71%	0.007	1.29%	8.13%

Appendix (continued)

TABLE A2C: **Accuracy of screen-in, adjusted analysis, by race, 6-month re-referral window**

	PREDICTED PROBABILITY OF A SCREEN-IN WITH FURTHER ACTION	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST				
White	55.72%	0.000	54.20%	57.25%
Black/African American	56.58%	0.000	54.96%	58.20%
Post-AFST				
White	61.34%	0.000	59.10%	63.58%
Black/African American	58.73%	0.000	56.62%	60.85%
Difference Post - Pre				
White	5.62%	0.000	3.53%	7.70%
Black/African American	2.15%	0.168	-0.91%	5.21%
Difference Black - White				
Pre-AFST	0.86%	0.449	-1.37%	3.08%
Post-AFST	-2.61%	0.105	-5.75%	0.54%

TABLE A3A: **Accuracy of screen-out, ITSA analysis, all children, 6-month re-referral window**

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	72.58%	0.000	68.84%	76.32%
Pre-2014 Policy	Trend	-0.01	0.953	-0.36	0.34
2014 Policy	Change in level	1.27	0.528	-2.75	5.29
Post 2014 policy, pre-AFST	Change in trend	0.10	0.653	-0.33	0.52
AFST implementation	Change in level	-2.19	0.386	-7.25	2.86
Post-AFST	Change in trend	-0.38	0.197	-0.96	0.20
Total trend in screen-in rates pre-AFST		0.09	0.454	-0.15	0.32
Total trend in screen-in rate post-AFST		-0.29	0.266	-0.82	0.23

Note: change in trend is expressed in percentage points/month.

Appendix (continued)

TABLE A3B: **Accuracy of screen-out, ITSA analysis, < 4 years old, 6-month re-referral window**

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	75.68%	0.000	70.16%	81.21%
Pre-2014 Policy	Trend	-0.16	0.530	-0.65	0.34
2014 Policy	Change in level	0.54	0.867	-5.93	7.01
Post 2014 policy, pre-AFST	Change in trend	0.25	0.505	-0.50	1.01
AFST implementation	Change in level	-1.75	0.746	-12.56	9.06
Post-AFST	Change in trend	-0.27	0.665	-1.54	0.99
Total trend in screen-in rates pre-AFST		0.10	0.736	-0.47	0.66
Total trend in screen-in rate post-AFST		-0.18	0.752	-1.31	0.95

Note: change in trend is expressed in percentage points/month.

TABLE A3C: **Accuracy of screen-out, ITSA analysis, 4 to 6 years old, 6-month re-referral window**

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	70.53%	0.000	66.23%	74.83%
Pre-2014 Policy	Trend	0.18	0.302	-0.17	0.54
2014 Policy	Change in level	0.28	0.906	-4.46	5.02
Post 2014 policy, pre-AFST	Change in trend	-0.07	0.803	-0.61	0.47
AFST implementation	Change in level	-6.28	0.051	-12.59	0.03
Post-AFST	Change in trend	0.19	0.643	-0.63	1.01
Total trend in screen-in rates pre-AFST		0.12	0.559	-0.29	0.52
Total trend in screen-in rate post-AFST		0.31	0.391	-0.41	1.02

Note: change in trend is expressed in percentage points/month.

Appendix (continued)

TABLE A3D: Accuracy of screen-out, ITSA analysis, 7 to 12 years old, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	69.46%	0.000	65.04%	73.87%
Pre-2014 Policy	Trend	0.19	0.352	-0.21	0.59
2014 Policy	Change in level	-0.30	0.894	-4.89	4.28
Post 2014 policy, pre-AFST	Change in trend	0.00	0.998	-0.45	0.45
AFST implementation	Change in level	-1.68	0.570	-7.58	4.23
Post-AFST	Change in trend	-0.83	0.030	-1.57	-0.09
Total trend in screen-in rates pre-AFST		0.19	0.074	-0.02	0.40
Total trend in screen-in rate post-AFST		-0.64	0.077	-1.35	0.07

Note: change in trend is expressed in percentage points/month.

TABLE A3E: Accuracy of screen-out, ITSA analysis, 13 to 17 years old, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	75.84%	0.000	71.38%	80.31%
Pre-2014 Policy	Trend	-0.36	0.114	-0.80	0.09
2014 Policy	Change in level	5.95	0.054	-0.10	12.00
Post 2014 policy, pre-AFST	Change in trend	0.25	0.366	-0.30	0.79
AFST implementation	Change in level	-0.34	0.856	-4.05	3.38
Post-AFST	Change in trend	-0.16	0.500	-0.64	0.32
Total trend in screen-in rates pre-AFST		-0.11	0.489	-0.42	0.21
Total trend in screen-in rate post-AFST		-0.27	0.139	-0.63	0.09

Note: change in trend is expressed in percentage points/month.

Appendix (continued)

TABLE A3F: Accuracy of screen-out, ITSA analysis, White, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	71.67%	0.000	66.68%	76.65%
Pre-2014 Policy	Trend	0.09	0.704	-0.40	0.59
2014 Policy	Change in level	-3.93	0.169	-9.61	1.74
Post 2014 policy, pre-AFST	Change in trend	0.32	0.254	-0.24	0.88
AFST implementation	Change in level	-0.80	0.747	-5.79	4.19
Post-AFST	Change in trend	-1.18	0.000	-1.72	-0.65
Total trend in screen-in rates pre-AFST		0.41	0.003	0.15	0.68
Total trend in screen-in rate post-AFST		-0.77	0.002	-1.23	-0.31

Note: change in trend is expressed in percentage points/month.

TABLE A3G: Accuracy of screen-out, ITSA analysis, Black/African American, 6-month re-referral window

		STARTING RATE (%) OR CHANGE (PERCENTAGE POINTS)	P > |T|	[95% CI]	
Start point (August 2013)	Level	71.98%	0.000	65.87%	78.10%
Pre-2014 Policy	Trend	-0.17	0.470	-0.65	0.30
2014 Policy	Change in level	5.57	0.067	-0.41	11.56
Post 2014 policy, pre-AFST	Change in trend	-0.06	0.864	-0.72	0.61
AFST implementation	Change in level	-3.61	0.348	-11.30	4.07
Post-AFST	Change in trend	0.40	0.365	-0.48	1.27
Total trend in screen-in rates pre-AFST		-0.23	0.322	-0.69	0.23
Total trend in screen-in rate post-AFST		0.17	0.650	-0.58	0.91

Note: change in trend is expressed in percentage points/month.

TABLE A4A: Accuracy of screen-out, adjusted analysis, all children, 6-month re-referral window

	PREDICTED PROBABILITY OF A SCREEN-OUT WITH NO REREFERRAL	P-VALUE	[95% C.I.]	
			LOWER	UPPER
Pre-AFST	73.90%	0.000	73.40%	74.40%
Post-AFST	72.31%	0.000	71.07%	73.54%
DIFF (Post - Pre)	-1.59%	0.014	-2.86%	-0.32%

Appendix (continued)

TABLE A4B: **Accuracy of screen-out, adjusted analysis, by age group, 6-month re-referral window**

	PREDICTED PROBABILITY OF A SCREEN-OUT WITH NO REREFERRAL	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
Pre-AFST				
< 4 years	73.00%	0.000	71.08%	74.92%
4–6 years	74.33%	0.000	73.12%	75.54%
7–12 years	74.10%	0.000	73.10%	75.09%
13–17 years	73.97%	0.000	72.29%	75.65%
Post-AFST				
< 4 years	72.45%	0.000	69.35%	75.55%
4–6 years	71.76%	0.000	69.46%	74.05%
7–12 years	72.22%	0.000	70.18%	74.27%
13–17 years	72.66%	0.000	70.66%	74.67%
Difference Post–Pre				
< 4 years	-0.55%	0.746	-3.86%	2.77%
4–6 years	-2.57%	0.071	-5.36%	0.22%
7–12 years	-1.88%	0.160	-4.49%	0.74%
13–17 years	-1.31%	0.363	-4.13%	1.51%

TABLE A4C: **Accuracy of screen-out, adjusted analysis, by race, 6-month re-referral window**

	PREDICTED PROBABILITY OF A SCREEN-OUT WITH NO REREFERRAL	P-VALUE	[95% C.I.] LOWER	[95% C.I.] UPPER
Pre-AFST				
White	70.55%	0.000	69.30%	71.81%
Black/African American	74.08%	0.000	72.92%	75.24%
Post-AFST				
White	70.45%	0.000	68.71%	72.18%
Black/African American	70.52%	0.000	68.62%	72.42%
Difference Post–Pre				
White	-0.11%	0.929	-2.41%	2.20%
Black/African American	-3.56%	0.001	-5.67%	-1.44%
Difference Black–White				
Pre-AFST	3.53%	0.001	1.39%	5.66%
Post-AFST	0.08%	0.952	-2.41%	2.57%

Appendix (continued)

TABLE APPENDIX A5: **Regression results for further action or no further action and re-referral within 60 days, conditional on screen-in (Outcome 1: accuracy of screen-in)**

	(1)	(2)	(3)
VARIABLES	POLICY ONLY	POLICY INTERACTED WITH RACE	POLICY INTERACTED WITH AGE-GROUP
Post-AFST	0.12** [0.03–0.21]	0.29*** [0.16–0.42]	-0.02 [-0.15–0.10]
Post-AFST interacted with race group			
Post-AFST x Black/African American		-0.30*** [-0.47–-0.12]	
Post-AFST interacted with age-group			
Post-AFST x age 4 to 6 years			0.15* [-0.01–0.30]
Post- AFST x age 7 to 12 years			0.21*** [0.06–0.35]
Post-AFST x age 13 to 17 years			0.21** [0.04–0.38]
Race (comparator is White)			
Black/African American	-0.03 [-0.13–0.07]	0.11* [-0.02–0.24]	-0.03 [-0.13–0.07]
Age-group (age < 4 is comparator)			
age 4 to 6 years	0.10** [0.02–0.17]	0.09** [0.02–0.17]	0.02 [-0.08–0.13]
age 7 to 12 years	0.13*** [0.06–0.20]	0.13*** [0.06–0.19]	0.03 [-0.07–0.13]
age 13 to 17 years	0.11*** [0.03–0.19]	0.11*** [0.03–0.19]	0.01 [-0.11–0.12]
Legal sex (comparator is female)			
Male	-0.04 [-0.09–0.02]	-0.04 [-0.09–0.02]	-0.04 [-0.09–0.02]
HH composition counts			
< 1	-0.23*** [-0.35–-0.12]	-0.22*** [-0.34–-0.11]	-0.23*** [-0.35–-0.12]
1 to 5 years	0.07** [0.01–0.13]	0.07** [0.01–0.13]	0.07** [0.01–0.13]
6 to 12 years	0.04* [-0.01–0.09]	0.05* [-0.01–0.10]	0.04* [-0.01–0.09]
13 to 17 years	-0.08** [-0.15–-0.02]	-0.08** [-0.15–-0.02]	-0.08** [-0.15–-0.02]
Parents	-0.00 [-0.06–0.05]	-0.00 [-0.05–0.05]	-0.00 [-0.06–0.05]
Other adults	-0.03 [-0.09–0.03]	-0.03 [-0.09–0.03]	-0.03 [-0.09–0.03]

Appendix (continued)

VARIABLES	(1) POLICY ONLY	(2) POLICY INTERACTED WITH RACE	(3) POLICY INTERACTED WITH AGE-GROUP
Mean age of all adults in referral (comparator is no adult age reported)			
18–29 years	-0.30 [-0.73–0.13]	-0.30 [-0.74–0.13]	-0.30 [-0.73–0.13]
30–49 years	-0.34 [-0.76–0.08]	-0.34 [-0.77–0.08]	-0.34 [-0.76–0.09]
50–65 years	-0.44* [-0.89–0.01]	-0.44* [-0.90–0.01]	-0.44* [-0.89–0.01]
66/max years	-0.72* [-1.46–0.03]	-0.74* [-1.48–0.00]	-0.71* [-1.46–0.04]
Household poverty zip code bins (comparator is no zip code listed)			
Poorest	-1.25*** [-1.56–-0.94]	-1.26*** [-1.57–-0.94]	-1.25*** [-1.56–-0.93]
Poor	-1.22*** [-1.53–-0.92]	-1.22*** [-1.53–-0.92]	-1.22*** [-1.52–-0.91]
Mid	-1.09*** [-1.42–-0.76]	-1.09*** [-1.42–-0.76]	-1.09*** [-1.41–-0.76]
Wealthier	-1.15*** [-1.46–-0.84]	-1.15*** [-1.46–-0.84]	-1.14*** [-1.46–-0.83]
Wealthiest	-1.12*** [-1.44–-0.81]	-1.13*** [-1.44–-0.81]	-1.12*** [-1.43–-0.81]
Risk score (historical plus projected, comparator is no risk score)			
Low	-2.46*** [-3.41–-1.51]	-2.45*** [-3.39–-1.51]	-2.46*** [-3.41–-1.50]
Middle	-1.68*** [-2.62–-0.74]	-1.67*** [-2.60–-0.74]	-1.68*** [-2.62–-0.74]
High	-0.96** [-1.89–-0.03]	-0.95** [-1.88–-0.03]	-0.96** [-1.90–-0.02]
Mandatory	-0.04 [-0.98–0.89]	-0.03 [-0.96–0.89]	-0.04 [-0.98–0.90]
Observations	**26,010**	**26,010**	**26,010**

Appendix (continued)

TABLE APPENDIX A6: **Regression results for no re-referral within 60 days, conditional on screen-out (Outcome 2: accuracy of screen-out)**

	(1)	(2)	(3)
VARIABLES	POLICY ONLY	POLICY INTERACTED WITH RACE	POLICY INTERACTED WITH AGE GROUP
Post-AFST	-0.09* [-0.19–0.01]	-0.07 [-0.21–0.07]	-0.10 [-0.32–0.11]
Post-AFST interacted with race group			
Post-AFST x Black/African American		-0.07 [-0.29–0.15]	
Post-AFST interacted with age-group			
Post-AFST x age 4 to 6 years			-0.07 [-0.300.16]
Post- AFST x age 7 to 12 years			0.00 [-0.29–0.30]
Post-AFST x age 13 to 17 years			0.09 [-0.15–0.32]
Race (comparator is White)			
Black/African American	0.16*** [0.04–0.28]	0.20*** [0.06–0.34]	0.16*** [0.04–0.28]
Other race	0.14* [-0.02–0.30]		0.14* [-0.02–0.30]
Unable to determine race	0.97*** [0.62–1.33]		0.98*** [0.62–.33]
Age-group (age < 4 is comparator)			
age 4 to 6 years	-0.07 [-0.17–0.03]	-0.05 [-0.15–0.05]	-0.04 [-0.17–0.09]
age 7 to 12 years	-0.05 [-0.15–0.05]	-0.03 [-0.13–0.06]	-0.05 [-0.22–0.12]
age 13 to 17 years	-0.08 [-0.20–0.04]	-0.06 [-0.18–0.06]	-0.13* [-0.27–0.01]
Legal sex (comparator is female)			
Male	0.03 [-0.06–0.12]	0.03 [-0.06–0.11]	0.03 [-0.06–0.12]
HH composition counts			
< 1	0.01 [-0.17–0.18]	-0.02 [-0.21–0.17]	0.01 [-0.17–0.19]
1 to 5 years	-0.01 [-0.09–0.07]	-0.03 [-0.11–0.05]	-0.01 [-0.09–0.07]
6 to 12 years	-0.07*** [-0.12–-0.02]	-0.07** [-0.13–-0.01]	-0.07*** [-0.12–-0.02]
13 to 17 years	0.02 [-0.05–0.08]	0.01 [-0.05–0.07]	0.02 [-0.05–0.08]

Appendix (continued)

	(1)	(2)	(3)
VARIABLES	POLICY ONLY	POLICY INTERACTED WITH RACE	POLICY INTERACTED WITH AGE GROUP
Parents	-0.03 [-0.08–0.03]	-0.02 [-0.08–0.03]	-0.03 [-0.08–0.03]
Other adults	-0.02 [-0.08–0.04]	-0.02 [-0.08–0.04]	-0.02 [-0.08–0.04]
Mean age of all adults in referral (comparator is no adult age reported)			
18 - 29 years	-0.83*** [-1.23–-0.44]	-0.82*** [-1.24–-0.40]	-0.84*** [-1.23–-0.44]
30 - 49 years	-0.76*** [-1.12–-0.39]	-0.73*** [-1.12–-0.34]	-0.76*** [-1.12–-0.39]
50 - 65 years	-0.61*** [-1.00–-0.22]	-0.61*** [-1.02–-0.19]	-0.61*** [-1.00–-0.23]
66/max years	0.02 [-0.68–0.72]	-0.09 [-0.80–0.61]	0.02 [-0.68–0.72]
Household poverty zip code bins (comparator is no zip code listed)			
Poorest	-0.22 [-0.53–0.08]	-0.17 [-0.48–0.14]	-0.22 [-0.53–0.08]
Poor	-0.30** [-0.60–-0.00]	-0.25 [-0.56–0.06]	-0.30** [-0.60–-0.00]
Mid	-0.34** [-0.60–-0.08]	-0.30** [-0.57–-0.02]	-0.34** [-0.60–-0.08]
Wealthier	-0.39*** [-0.64–-0.13]	-0.32** [-0.58–-0.06]	-0.39*** [-0.64–-0.13]
Wealthiest	-0.36*** [-0.61–-0.11]	-0.29** [-0.56–-0.03]	-0.36*** [-0.61–-0.11]
Risk score (historical plus projected, comparator is no risk score)			
Low	0.41 [-0.33–1.15]	0.37 [-0.35–1.09]	0.42 [-0.32–1.15]
Middle	0.10 [-0.65–0.85]	0.08 [-0.66–0.82]	0.11 [-0.64–0.86]
High	-0.39 [-1.13–0.36]	-0.38 [-1.11–0.35]	-0.38 [-1.12–0.36]
Mandatory	-0.38 [-1.07–0.31]	-0.38 [-1.06–0.30]	-0.38 [-1.07–0.31]
Observations	28,957	25,740	28,957

All regression results are based on a GLM model, with standard errors clustered at the screener-level

Appendix (continued)

TABLE APPENDIX A7: Regression results for screen-in (Outcome 3: workload)

	(1)	(2)	(3)
VARIABLES	POLICY ONLY	POLICY INTERACTED WITH RACE	POLICY INTERACTED WITH AGE GROUP
Post-AFST	-0.13 [-0.30–0.03]	-0.06 [-0.24–0.13]	-0.06 [-0.19–0.08]
Post-AFST interacted with race group			
Post-AFST x Black/African American		-0.13* [-0.27–0.01]	
Post-AFST interacted with age-group			
Post-AFST x age 4 to 6 years			0.08 [-0.04–0.19]
Post- AFST x age 7 to 12 years			-0.07 [-0.20–0.06]
Post-AFST x age 13 to 17 years			-0.25*** [-0.40–-0.10]
Race (comparator is White)			
Black/African American	0.01 [-0.06–0.09]	0.07 [-0.05–0.20]	0.01 [-0.06–0.08]
Other race	0.06 [-0.04–0.15]		0.06 [-0.03–0.16]
Unable to determine race	0.12*** [0.03–0.22]		0.12** [0.02–0.21]
Age-group (age < 4 is comparator)			
age 4 to 6 years	-0.07** [-0.14–-0.01]	-0.07** [-0.13–-0.01]	-0.11*** [-0.18–-0.04]
age 7 to 12 years	-0.06** [-0.11–-0.00]	-0.05* [-0.11–0.01]	-0.02 [-0.10–0.05]
age 13 to 17 years	-0.25*** [-0.32–-0.18]	-0.26*** [-0.32–-0.19]	-0.13** [-0.23–-0.02]
Legal sex (comparator is female)			
Male	-0.02 [-0.06–0.01]	-0.03* [-0.06–0.00]	-0.02 [-0.06–0.01]
HH composition counts			
< 1	0.65*** [0.55–0.76]	0.66*** [0.54–0.78]	0.65*** [0.55–0.76]
1 to 5 years	0.10*** [0.05–0.15]	0.10*** [0.04–0.15]	0.10*** [0.05–0.15]
6 to 12 years	0.12*** [0.08–0.17]	0.12*** [0.08–0.17]	0.12*** [0.08–0.17]
13 to 17 years	-0.03 [-0.08–0.02]	-0.03 [-0.08–0.02]	-0.03 [-0.08–0.02]
Parents	-0.04 [-0.08–0.01]	-0.03 [-0.08–0.02]	-0.04 [-0.08–0.01]

Appendix (continued)

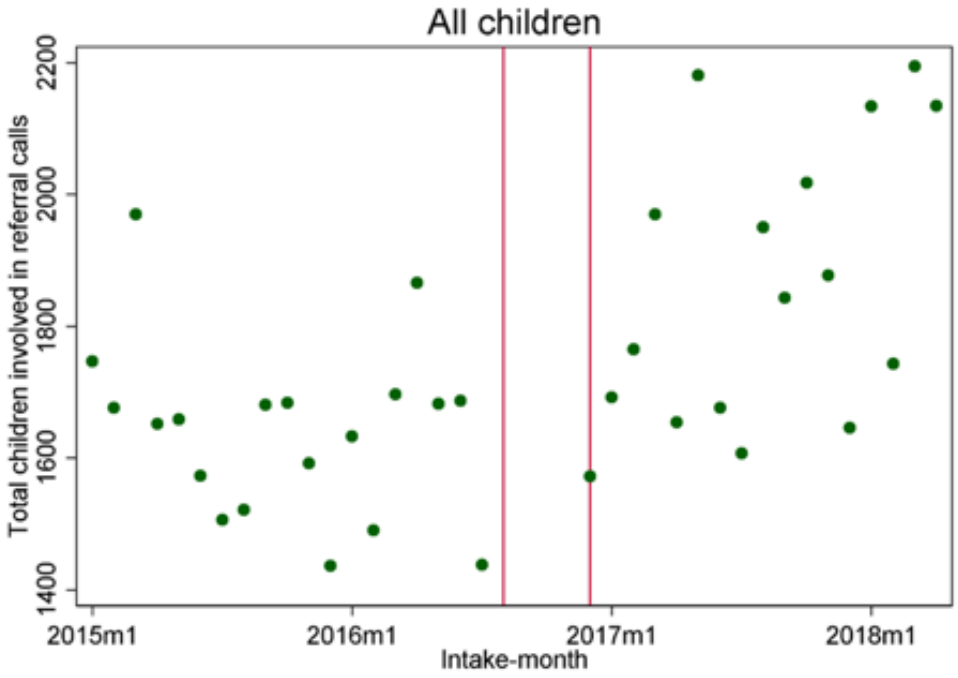
	(1)	(2)	(3)
VARIABLES	POLICY ONLY	POLICY INTERACTED WITH RACE	POLICY INTERACTED WITH AGE GROUP
Other adults	0.09*** [0.06–0.13]	0.10*** [0.06–0.14]	0.09*** [0.06–0.13]
Mean age of all adults in referral (comparator is no adult age reported)			
18–29 years	1.08*** [0.90–1.26]	1.02*** [0.84–1.21]	1.08*** [0.90–1.26]
30–49 years	1.08*** [0.90–1.27]	1.01*** [0.82–1.20]	1.09*** [0.90–1.27]
50–65 years	1.15*** [0.90–1.40]	1.09*** [0.82–1.36]	1.15*** [0.90–1.40]
66/max years	1.05*** [0.56–1.55]	1.12*** [0.64–1.59]	1.05*** [0.56–1.55]
Household poverty zip code bins (comparator is no zip code listed)			
Poorest	0.57** [0.04–1.10]	0.55** [0.00–1.09]	0.57** [0.04–1.10]
Poor	0.66** [0.14–1.18]	0.62** [0.09–1.16]	0.66** [0.14–1.18]
Mid	0.65*** [0.17–1.14]	0.59** [0.09–1.09]	0.65*** [0.17–1.14]
Wealthier	0.73*** [0.22–1.24]	0.69** [0.16–1.21]	0.72*** [0.22–1.23]
Wealthiest	0.66*** [0.17–1.16]	0.61** [0.09–.12]	0.66*** [0.16–1.15]
Risk score (historical plus projected, comparator is no risk score)			
Low	-0.05 [-0.57–0.47]	-0.05 [-0.58–0.48]	-0.06 [-0.58–0.46]
Middle	0.53* [-0.01–1.06]	0.51* [-0.03–1.05]	0.52* [-0.02–1.06]
High	0.93*** [0.38–1.47]	0.92*** [0.38–1.47]	0.92*** [0.37–1.46]
Mandatory	2.19*** [1.62–2.75]	2.18*** [1.61–2.75]	2.18*** [1.61–2.74]
Observations	60,287	54,388	60,287

All regression results are based on a GLM model, with standard errors clustered at the screener-level

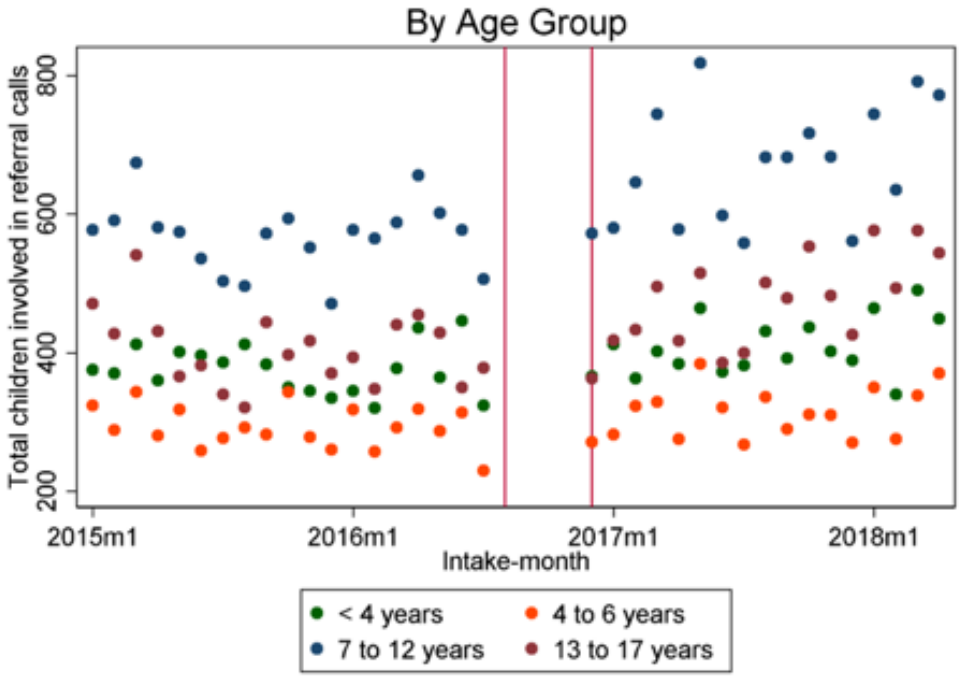
Appendix (continued)

APPENDIX FIGURES

APPENDIX FIGURE 1A: Total children in referral calls, by month

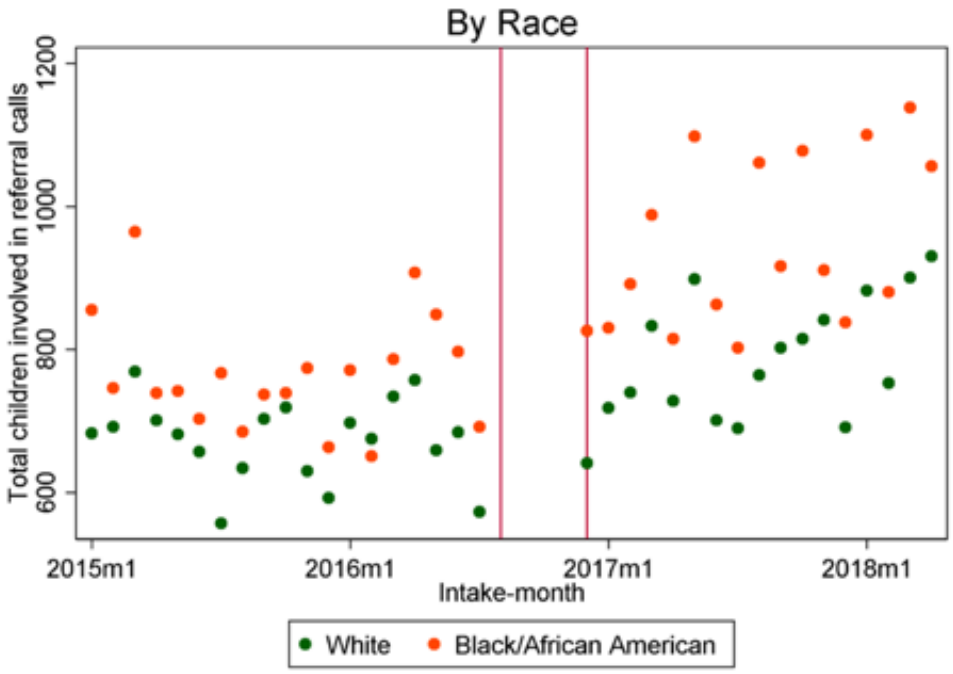


APPENDIX FIGURE 1B: Total children in referral calls, by month and age-group

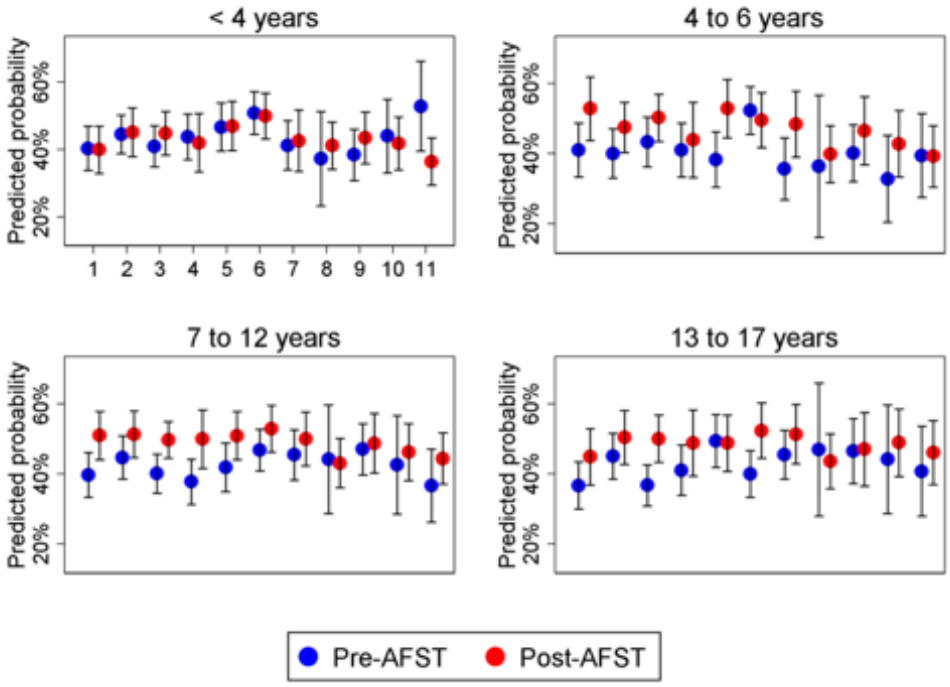


Appendix (continued)

APPENDIX FIGURE 1C: **Total children in referral calls, by month and race**

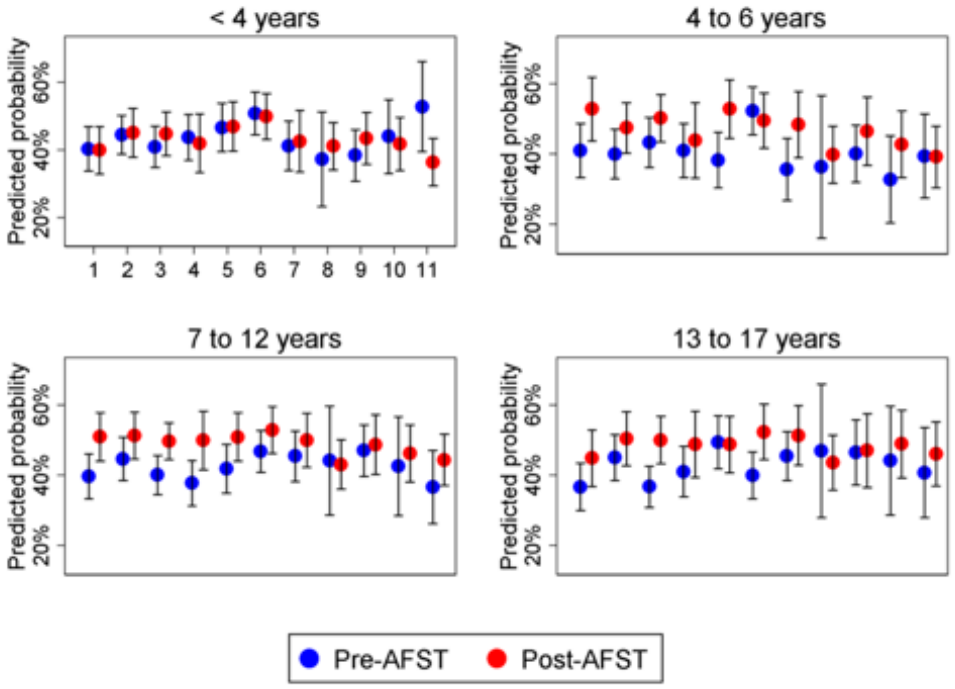


APPENDIX FIGURE 2A: **Predicted probability of accuracy of screen-in, consistency across 11 call screeners, adjusted analysis**

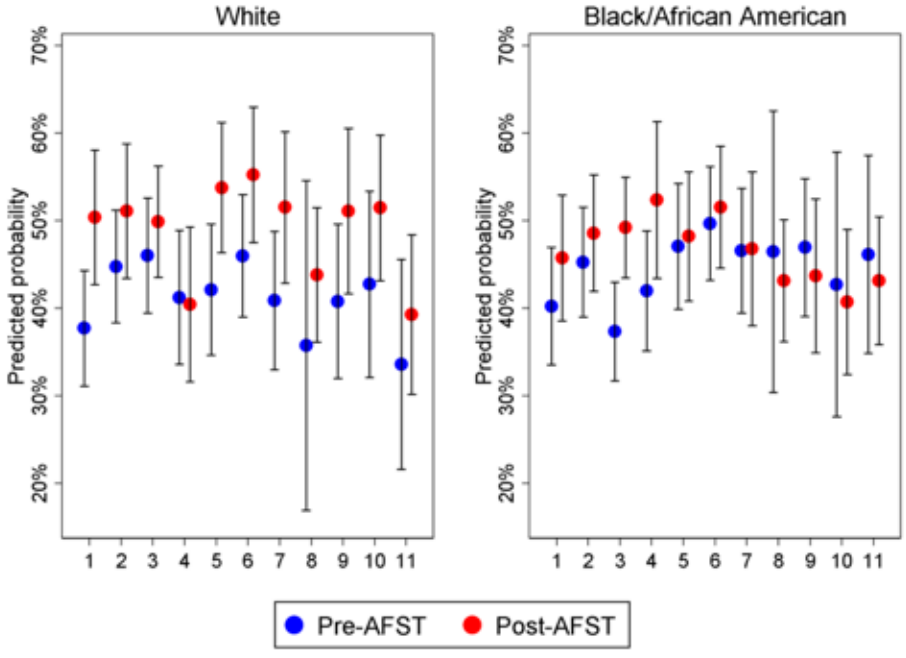


Appendix (continued)

APPENDIX FIGURE 2B: **Predicted probability of accuracy of screen-in, consistency across 11 call screeners, adjusted analysis, by age-group**

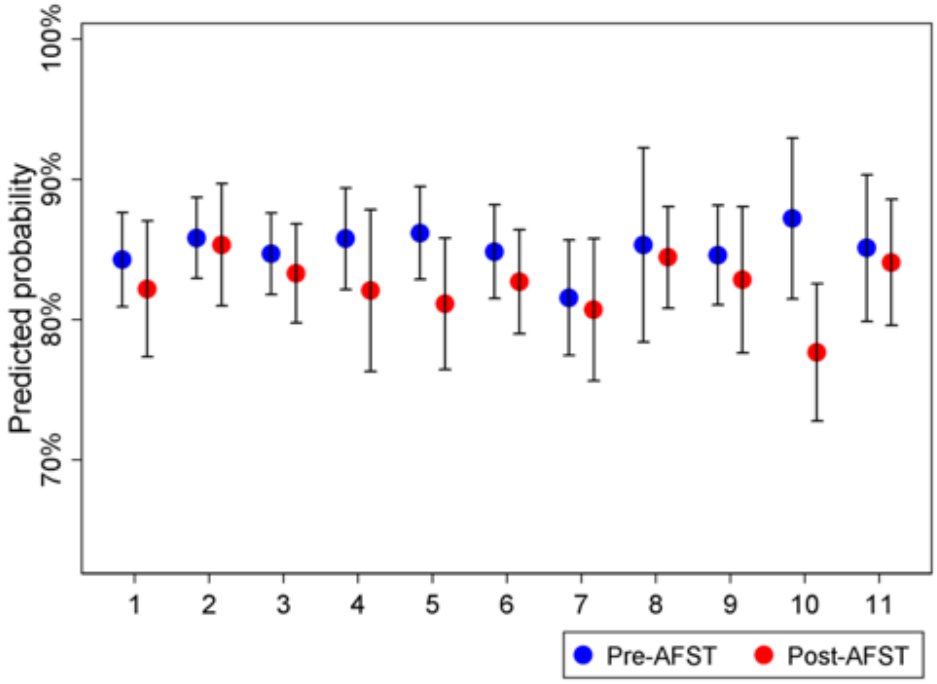


APPENDIX FIGURE 2C: **Predicted probability of accuracy of screen-in, consistency across 11 call screeners, adjusted analysis, by race**

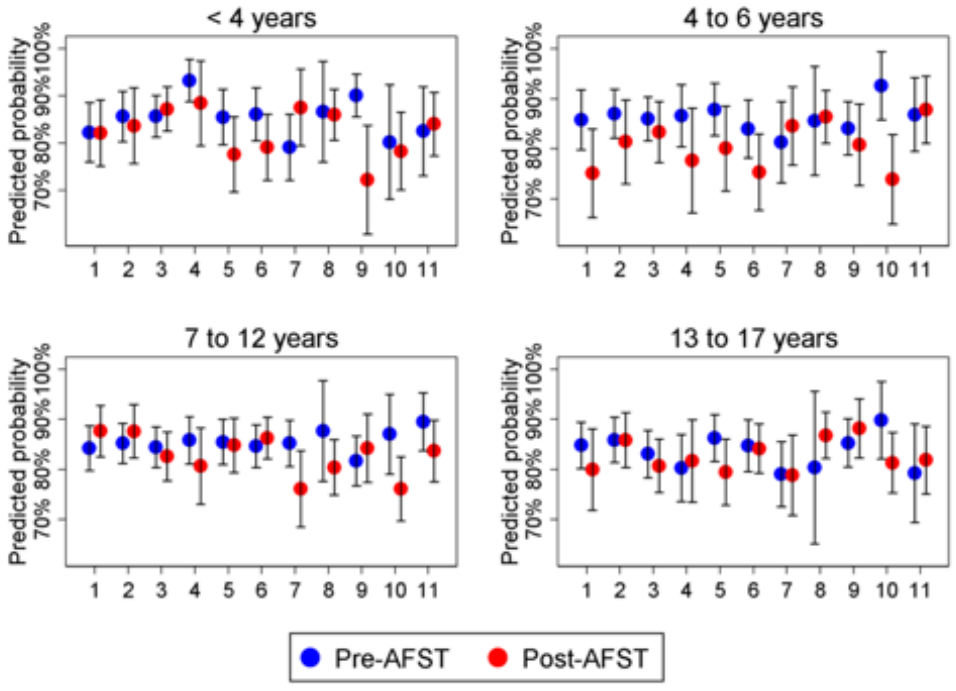


Appendix (continued)

APPENDIX FIGURE 3A: **Predicted probability of accuracy of screen-out, consistency across 11 call screeners, adjusted analysis**

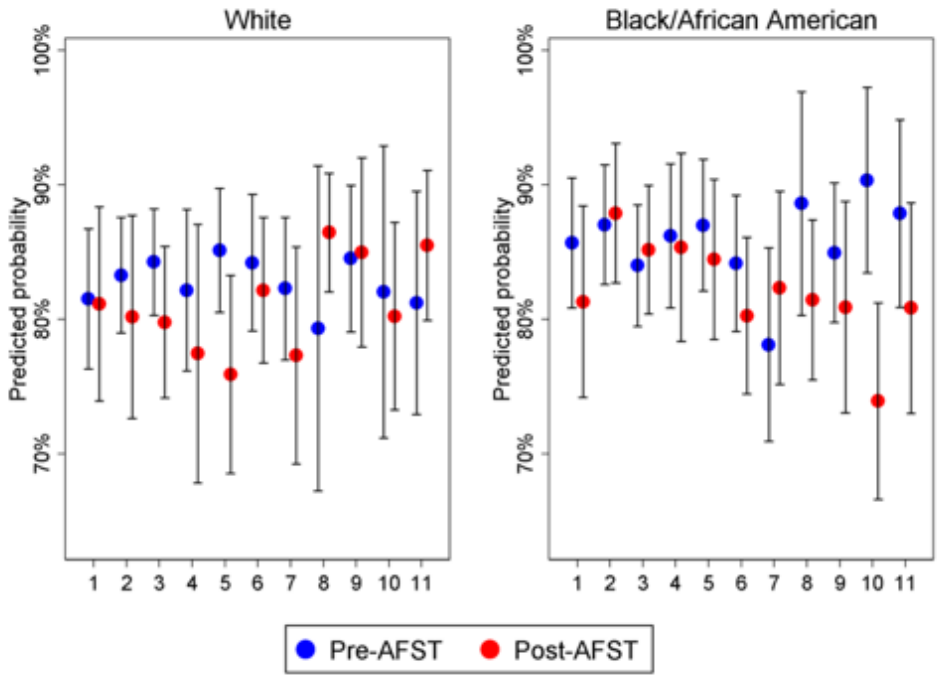


APPENDIX FIGURE 3B: **Predicted probability of accuracy of screen-out, consistency across 11 call screeners, adjusted analysis, by age-group**

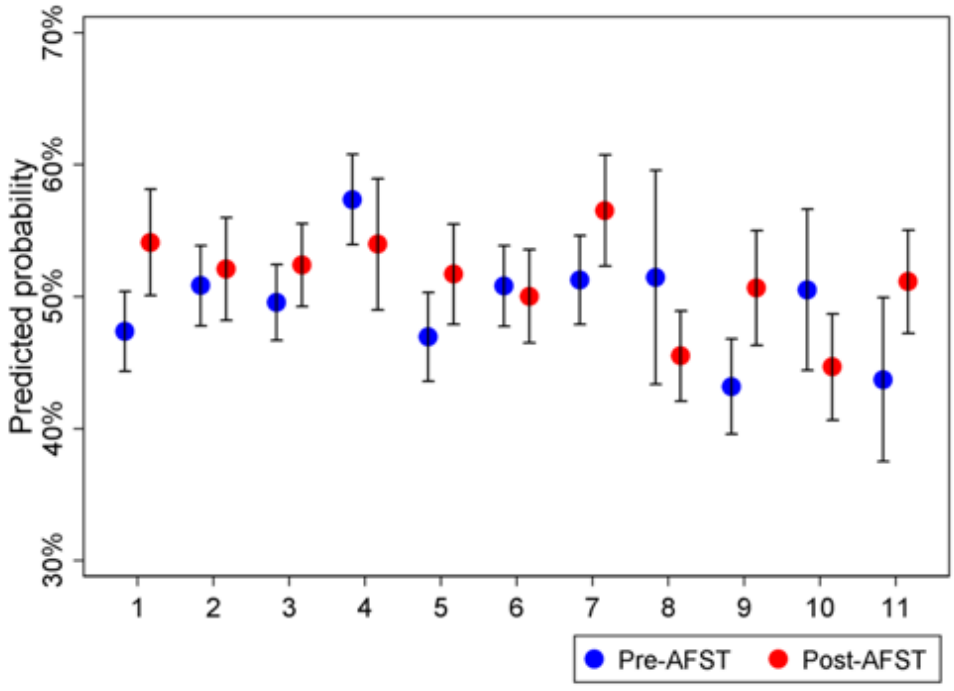


Appendix (continued)

APPENDIX FIGURE 3C: **Predicted probability of accuracy of screen-out, consistency across 11 call screeners, adjusted analysis, by race**

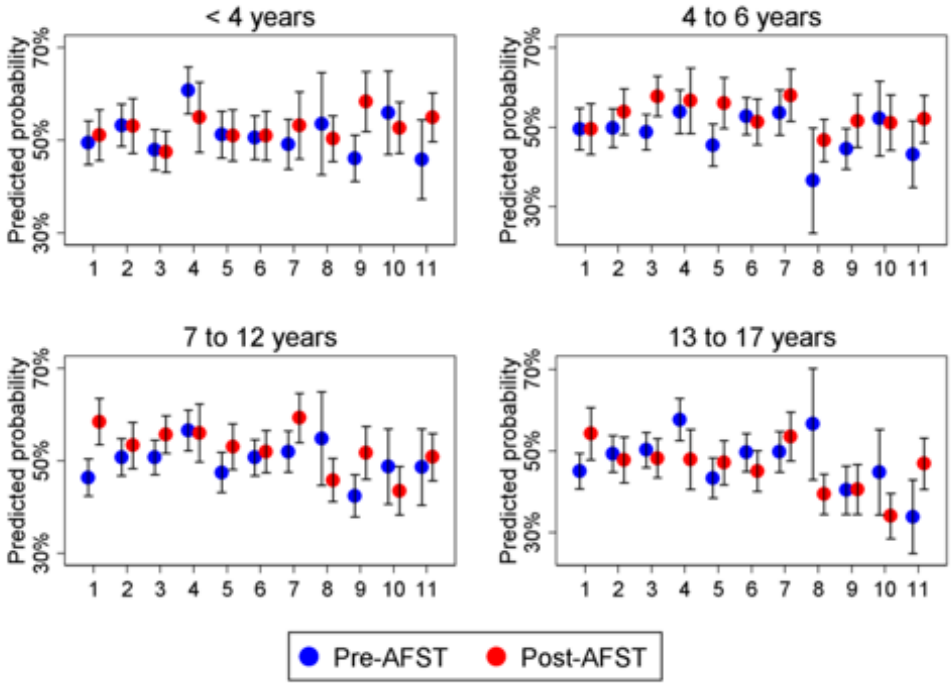


APPENDIX FIGURE 4A: **Predicted probability of workload, consistency across 11 call screeners, adjusted analysis**

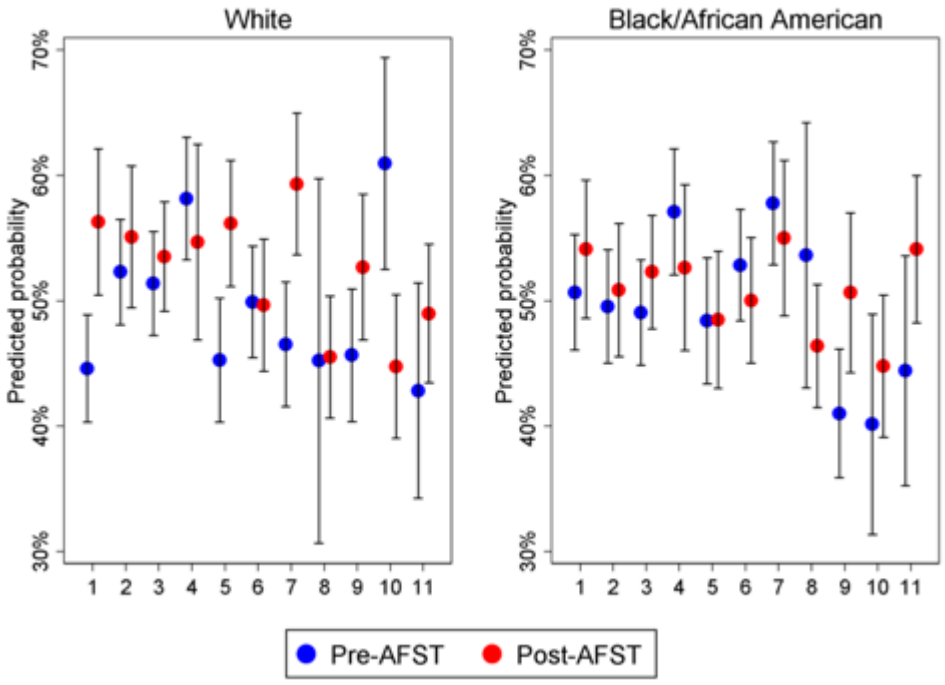


Appendix (continued)

APPENDIX FIGURE 4B: **Predicted probability of workload, consistency across 11 call screeners, adjusted analysis, by age-group**



APPENDIX FIGURE 4C: **Predicted probability of workload, consistency across 11 call screeners, adjusted analysis, by race**



SECTION 7

Allegheny Family Screening Tool: Methodology, Version 2

prepared by Rhema Vaithianathan, PhD (Center for Social Data Analytics, Auckland University of Technology), Emily Kulick (Center for Social Data Analytics), Emily Putnam-Hornstein, PhD (Children's Data Network, University of Southern California), Diana Benavides Prado (Center for Social Data Analytics)

CONTENTS

Background *2*

Major Changes to the AFST since Methodology V1 *2*

- Target Outcomes *2*
- Predictors *3*
- Policy *4*
- Modeling Methodology *6*

External Validation of AFST V2 *7*

Conclusion *9*

References *9*

APPENDIX A: Exploration of Modeling Methodologies for AFST V2 *10*

- Logistic regression (LR) method *10*
- LASSO regression method *10*
- Random Forest *10*
- XG Boost *11*
- Modeling Results *11*
- Discussion of Model Choices *15*

APPENDIX B: Weighted Variables in AFST V2 *16*

APPENDIX C: AFST V2 Visualizations *21*

APPENDIX D: Hospital Injury Classifications *22*

INTRODUCTION

This methodology report describes changes to the Allegheny Family Screening Tool (AFST), building upon and updating the original methodology report, Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions (March 2017). The March 2017 report and accompanying documents include background information on the development of the AFST as well as an ethical analysis, impact evaluation and frequently-asked questions. As such, they provide context for this report about changes that have been made to the AFST since the original methodology report was written.

BACKGROUND

In August 2016, Allegheny County introduced a predictive risk model to support decision-making at the time that child abuse and/or neglect allegations are received. Version 1 (V1) of the AFST decision-support tool was in use from August 2016 through November 2018. Since then, a number of modifications have been made to the tool as part of the County's commitment to updating the model and related policies as source systems and variables are updated or policies are revisited. Modifications implemented in Version 2 (V2) of the AFST include changes to specific predictor fields used in the model itself, the modeling methodology, and County policies concerning the tool's use.

This Methodology V2 report provides information about changes made to the tool between the time the first report was written (April 2017) through April 2019. This report upholds Allegheny County's ongoing commitment to transparency by continuing to inform the community about changes to the tool and the County's policies. As this is a status report, details will likely change over time as the County continues to evaluate the impact of the tool and improve its accuracy.

MAJOR CHANGES TO THE AFST SINCE METHODOLOGY V1

Target Outcomes

AFST V1 consisted of two models. The first model, called the placement model, was trained to predict whether, within the two years following a referral, a child would experience a safety issue so significant that they would need to be removed from their home and placed in an out-of-home setting. The second model, called the re-referral model, was trained to predict whether within that same time period, a child who was initially referred and screened out would be re-referred as an alleged victim of maltreatment. Only a single score, the one that was the highest of the placement and re-referral models across all children on the referral, was shared with the call screener. For example, if there were two children on a referral and the older child scored 12 on the placement model and 15 on the re-referral model, and the younger child scored 7 on the placement model and 11 on the re-referral model, the score shared with the call screener would be 15.

The re-referral model (which predicted whether a child who was referred and screened out would be a re-referred within two years) was not as strongly linked to the primary outcome of concern, serious abuse and neglect. One of the reasons that the re-referral model did not have

strong face validity is because high scores on that model could reflect children embroiled in custody disputes or other situations where there are frequent calls about the same issue. Additionally, initial incoming referral rates also represent the most racially disproportionate step of the referral pathway, and so a model predicting future referrals tends to overrepresent black children relative to white. Finally, the nature and characteristics of calls with higher scores using the re-referral model were resonating less strongly with screening staff as cases appropriate for investigation. An external validation that examined children's assigned risk scores against their medical encounters for injuries also suggested that the scores from the re-referral model did not create value above and beyond the placement model. In AFST V2, we have therefore restricted the model to predicting safety issues that are so significant that they lead to a court-ordered out-of-home placement outcome.

Predictors

Both V1 and V2 of the AFST use existing administrative data concerning children and adults named in a maltreatment referral to automatically generate a risk score. These integrated data are available to Allegheny County child protection staff through the County's data warehouse and reflect records originating from a wide range of sources. In the two years since AFST V1 was implemented, the characteristics of records and information in the data warehouse have changed as a result of changes made to fields in source data systems. This means some data included in the first release of the AFST may no longer be available or is now available in a different form, while other information is newly available. Changes in the source data systems and predictors used to build the AFST are outlined below.

County data sources *used in **both** the AFST V1 and AFST V2*:

- Child welfare records
- Jail records
- Juvenile probation records
- Behavioral health records

County data sources *used in AFST V2 that were **not** used in AFST V1*:

- Birth records

County data sources ***not** used in AFST V2 that were used in AFST V1*:

- Public benefit records (e.g., Temporary Aid to Needy Families [TANF], Supplemental Nutrition Assistance Program [SNAP])

In some cases, while a data source continued to be used to generate predictors in the AFST V2, the specific fields changed.

Despite the wide array of information about the history of referred individuals available in Allegheny County's integrated data warehouse, the need for call screeners and their supervisors

to distill a large volume of information while making quick decisions meant that call screeners historically often relied heavily on the allegation (i.e., the nature of the maltreatment that was being alleged) as a main determinant of screening decisions. AFST V1 did not use allegations as predictors, which might have reduced its face validity with screening staff. AFST V2 includes allegations as additional predictor fields.

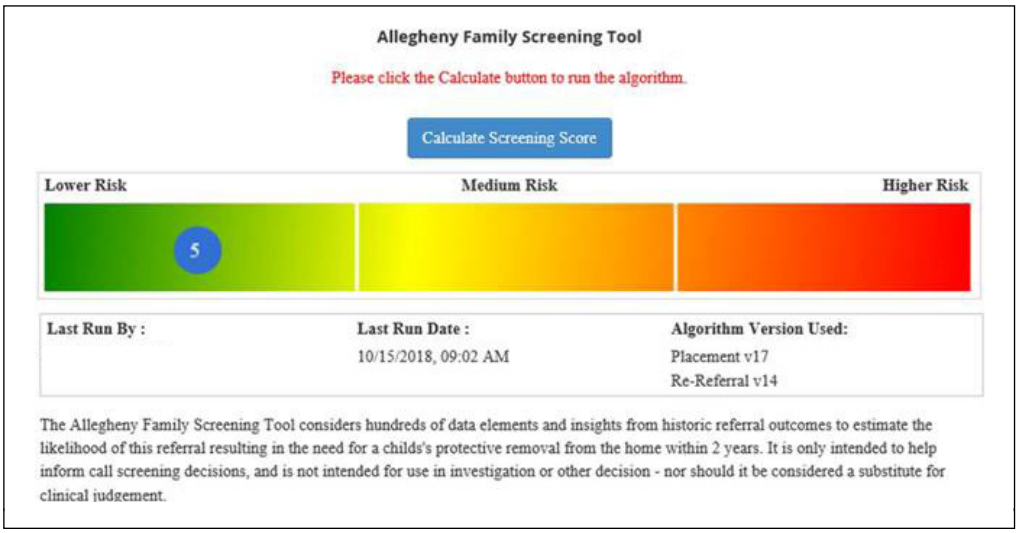
Public benefits data were excluded from V2 as these had changed over time and no longer aligned with the data used to develop V1. Additionally, a majority of the behavioral health fields used in AFST V1 were excluded in V2. In recent years, systematic changes occurred in how behavioral health diagnoses were defined and categorized. These changes meant that the behavioral health classifications in the research data used to build the model did not align with definitions currently "feeding" the algorithms. There was no information available that would allow these classifications to be harmonized across the time periods, and the team is working to restructure the behavioral health fields to reincorporate them into the model. The variables will likely focus on service type and severity, with additional predictors to identify if there were any prior services under each diagnostic category. The behavioral health variables that remain in the V2 model reflect aggregated indicators for whether each individual on the referral received any prior behavioral health service, as well as the number of days since the last behavioral health service.

A full list of predictor fields used in AFST V2 can be found in **Appendix B: Weighted Variables in AFST V2**.

Policy

AFST V2 is being implemented with a new visualization that signals to call screening staff that this is a new and improved model. Additional screen shots of the visualizations can be found in **Appendix C: AFST V2 Visualizations**.

FIGURE 1: **AFST V2 visualization provided to call screening staff**



In addition to the new design, newly developed high-risk and low-risk protocols have been implemented. These protocols make use of "nudges," which default the highest-risk cases to be screened in and require supervisors to explicitly override the decision with written justification if they feel it should not be investigated; a similar default-based nudge with override capability was later added to the lowest-risk cases. The visualization displays the high- and low-risk protocol if the referral meets the criteria described in Table 1, below. If the referral does not meet the high- or low-risk protocol, then call screeners see the underlying risk score. The score for each referral continues to be the maximum score received among all children or victims on a referral. As noted above, the maximum score is now derived solely from the placement model, rather than both the placement and re-referral models.

TABLE 1: **Definitions and protocols for high- and low-risk referrals**

	DEFINITION	PROTOCOL	VISUALIZATION	PERCENTAGE OF ALL REFERRALS THAT FALL IN CATEGORY
High-Risk Protocol	Maximum score in a referral of greater than 17 and a victim child (or other child) age 16 years or younger	The referral is designated to be screened-in for investigation; however, supervisory discretion allows screen-out (override documentation required).	The following text is displayed: "High-Risk Protocol, High-Risk and Children Under Age 16 on Referral"	24%
Low-Risk Protocol	Maximum referral score in a referral of less than 11 and ALL victims and children are at least 12 years of age	Screen-out without investigation is recommended.	The following text is displayed: "Low-Risk Protocol, Low-Risk and All Children Age 12+ on Referral" and "recommended screen out".	4%
Other	All other referrals not defined as high-risk or low-risk	Full discretion and no policy recommendation	The categorical score is displayed on a horizontal bar with a gradient of green (1) to red (20)	72%

Modeling Methodology

The AFST V1 was developed using logistic regression; Methodology V1 utilized an Area Under the Receiver Operator Curve (AUC) to measure the probability that a (randomly chosen) referral that was a true positive had a higher risk score than a randomly chosen referral that was a true negative. A probability of higher than 50 percent indicates that the risk score was useful in guiding the screening decision. The Methodology V1 report found an AUC of between 76.9 percent and 78.3 percent.[1] As discussed in our research paper (Chouldechova et al., 2018), however, this reported AUC was over-stated because our split of records into training and testing sets failed to fully address sibling dynamics. Specifically, while referrals had been correctly split so that no unique child appeared in both sets, siblings with the same parent could have been inappropriately split between the test and research data sets. While this does not impair the performance of the previously deployed AFST V1, it does mean that original AUC was overstated.

1 See **Table 4** of Methodology V1 report.

For AFST V2, we explored a range of additional modeling methodologies to improve the AUC, including LASSO, XG-BOOST, Random Forest, and SVM. We discuss that process in detail in **Appendix A: Exploration of Modeling Methodologies for AFST V2**. In deciding which methodology we should adopt, we looked at 1) overall performance and accuracy for the specific high-risk group that serves as the focus of the County's policy, and 2) equivalent levels of accuracy for black children vs. non-black children.

We also gave due consideration to pragmatic questions of implementation and ongoing quality assurance. Given the large number of databases that are being linked in the AFST V2, quality checks and ongoing model maintenance are critical.

We ultimately decided to implement the LASSO model, and the remainder of this methodology report details the performance of that model. To assess whether the overall performance of the LASSO model across children of different racial and ethnic groups was similar, we computed the AUC by race. The overall AUC for the selected LASSO model is 75.97 percent,[2] the AUC for black children is 74.42[3] percent and the AUC for non-black children is 77.35[4] percent, suggesting that the tool was slightly better at predicting outcomes for non-black children than for black children.

2 95% Confidence interval (c.i.): 74.81%–77.13%

3 95% c.i.: 72.84%–75.99%

4 95% c.i,: 75.59%–79.11%

EXTERNAL VALIDATION OF AFST V2

External validation of the model is important to determine if the AFST V2 model, trained to predict the likelihood of a future child welfare out-of-home placement, is sensitive to more generalized and objective measures of child harm. Because true maltreatment rates are very difficult, if not impossible, to determine, we are left predicting measures of child maltreatment defined by the child protection system. As such, there are valid concerns that the AFST model, and other models trained to predict system outcomes like out-of-home placement, may be predicting the risk of institutionalized or system response rather than the true underlying risk of adverse events.

To address these concerns, we completed external validations of AFST V1 using medical records and critical events data. We have replicated those validations for AFST V2, as described below.

External Validation: Hospital Data

To externally validate the AFST V1 model using hospitalization records, we generated a probabilistic linkage between the County's maltreatment referral data and data from UPMC Children's Hospital of Pittsburgh. UPMC proved an ideal source of external data as it is the hospital that the majority of children in Allegheny County use. This means we had near universal medical encounter data (versus means-tested data) for children in the research dataset.

In our initial external validation, we documented that children who were identified in the highest risk groups by AFST V1 were the same children observed to have more generalized risk of relevant hospital events (see pages 19-23 of Methodology V1 for details on how the data was linked and what trends were observed for AFST V1).

We replicated this hospital validation for AFST V2, using the same linked dataset.

We examined hospital encounters (by cause) using four different approaches:

1. *Highest risk score and an injury encounter*: We looked at all unique children in our data, classified their risk based on the highest risk score assigned for any referral, and coded all associated injury encounters, regardless of whether the injury occurred before or after the child abuse and neglect referral.
2. *Randomly selected risk score and an injury encounter*: We looked at all unique children in the data, randomly selected a referral they were involved in and their risk score at that referral,

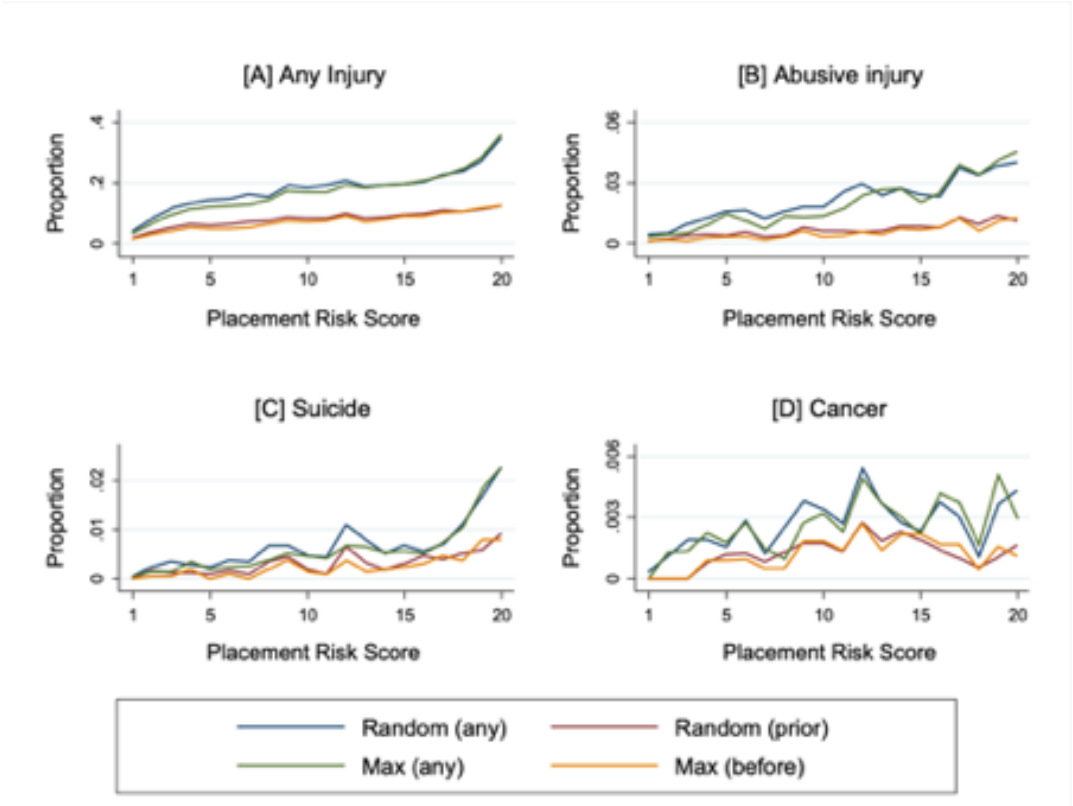
and coded their associated injury encounters, regardless of when the injury occurred relative to the selected child abuse and neglect referral.

3. *Highest risk score before an injury encounter*: We looked at all unique children in the data and coded the child's risk level based on the highest risk score assigned, but before a specific injury encounter.
4. *Randomly selected risk score before an injury encounter*: We looked at all unique children in the data, randomly selected a referral and associated risk score for each child, and coded a medical encounter as having occurred only if the selected referral date was before the injury encounter.

Figure 2 shows the pattern of medical (i.e., emergency department and hospital) encounters against each of the above different approaches. **Figures A** to **C** show a positive correlation between the AFST V2 risk scores and medical encounters for injury, abusive injuries and suicide. We also examined the association between cancer and the risk score as a "placebo" test; we do not see a strong correlation between cancer and risk scores, suggesting that the AFST is accurately identifying children at risk of abuse-related injuries only.

For more detail on how hospitalized injuries were classified see **Appendix D: Hospital Injury Classifications.**

FIGURE 2: **Children's medical encounters and risk scores**



In **Table 2**, we report the odds-ratios for each type of medical encounter following a high-risk referral (as defined in **Table 1**). Note that the odds-ratio for non-black children is larger than for the black children, meaning that non-black children's risk scores at referral were more strongly correlated with later medical encounters.

TABLE 2: Odds-ratio of medical encounter after referral (high-risk vs. non-high-risk)

	ALL CHILDREN (N=82,211)	BLACK (N=36,302)	NON-BLACK (N=45,909)
Injury	1.73*** [1.67, 1.80]	1.41*** [1.35, 1.48]	1.89*** [1.79, 2.00]
Abusive Injury	1.46*** [1.34, 1.59]	1.23*** [1.10, 1.37]	1.60*** [1.41, 1.83]
Suicide	1.71*** [1.48, 1.97]	1.30* [1.05, 1.60]	2.23*** [1.83, 2.72]
Cancer	1.23 [0.95, 1.61]	.90 [.61, 1.32]	1.68** [1.16, 2.43]

*Note: 95% confidence interval under odds-ratio. *=p<.1;**=p<.05;***=p<.01.*

CONCLUSION

This report is part of an ongoing commitment to providing both Allegheny County and broader stakeholders with regular updates on how the AFST is evolving over time. We believe that the changes we have made improve the utility of the tool and increase the accuracy of screening decisions.

Evaluations of the impact of the model by independent evaluators are also underway and will be published as they become available.

REFERENCES

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM.

Chouldechova, Alexandra, et al. "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions." Conference on Fairness, Accountability and Transparency. 2018.

Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal, Complex Systems*. 2006;1695(5):1-9.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

APPENDIX A: EXPLORATION OF MODELING METHODOLOGIES FOR AFST V2

A total of 82,211 unique child-referral observations were extracted for children referred for alleged neglect or abuse in Allegheny County between April 1, 2010 and July 31, 2014. Each observation reflected a unique child-referral record. Some children had more than one referral for a total of 46,507 unique children represented in the data. Outcomes for each child-referral record were observed until the end of the study window, July 31, 2016. To develop the predictive risk model, records were restricted to 45,801 observations in which the child was screened-in for an in-person investigation.

Each child-referral observation was attached to a set of 451 predictive variables describing the characteristics of the child, his/her family, the overall referral, and the alleged perpetrator of abuse. These variables included demographics of the family and alleged perpetrator, allegations associated with the referral, child and mother characteristics at the time of birth, as well as history of interactions with the child welfare system and with other social services such as jail, juvenile probation and mental health. The universe of screened-in referrals were partitioned into a 70/30 training (n=32,224) and validation set (n=13,577).

We used a graph-based method to partition the data into these two sets (Csardi G, Nepusz T., 2006). The method grouped all the children associated with a given referral into either the training or test partition. Because this method can lead to a lack of balance between the test and training partition based on the number of children on the call, balance in the count of children named on the referral was tested with a t-test to compare the average count between test and training set.

The model was trained to predict out-of-home placement within two years of the screened-in referral. Scores were generated at the child-referral level such that each score represents five percent of the referrals. For example, the child-referrals that score a 20 (the highest possible score) fall within the highest five percent of all child-referrals with respect to their predicted probability that the child will be placed in out-of-home care within two years of the scored referral.

Logistic regression method (LR)

This method was used to build an LR model on the training partition of the dataset and was used as the baseline for comparisons to other modeling alternatives.

LASSO regression method (LASSO)

The LASSO model (Tibshirani, 1996) was trained on the training partition using 10-fold cross-validation, with these folds selected randomly. The cross-validated model was trained to optimize for the AUC. The model selected 126 variables as weighted predictors of the target outcome along with the intercept term.

Random Forest (RF)

The Random Forest model (Breiman, 2000) was trained on the training partition with 500 trees and entropy as the splitting criterion. These parameters have been shown to provide the best results in terms of train and test performance in experiments with the Allegheny County dataset.

Appendix A (continued)

XG Boost (XGB)

The XGBoost model (Chen and Guestrin, 2016) was trained on the training partition with the following parameters: 1000 trees, learning rate of 0.01, maximum depth of individual regression estimators of 14, regularization lambda of 80, regularization alpha of 1e-05, minimum number of examples in a node of 1, a subsample ratio of columns per tree of 0.8, a ratio of number of examples of the negative class with respect to the positive class of 4.43, and a subsample percentage of 0.9. These parameters have been shown to provide the best results in terms of train and test performance in experiments with the Allegheny County dataset.

Modeling Results

Figure 3 shows the Receiver Operator Curve for the four modeling methods. As is clear from this Figure, LASSO, RF and XG Boost all perform similarly in terms of general predictive power: LASSO achieves an AUC of 75.97 (95% c.i. 74.81 – 77.16), RF achieves an AUC of 76.34 (95% c.i. 75.18 – 77.5), XGBoost achieves an AUC of 75.83 (95% c.i. 74.67 – 77.0). Logistic Regression achieves an AUC of 64.04 (95% c.i. 62.65 – 65.43), which is significantly lower than the other methods.

FIGURE 3: **Receiver Operator Curve for AFST V2 (test data only)**[5]

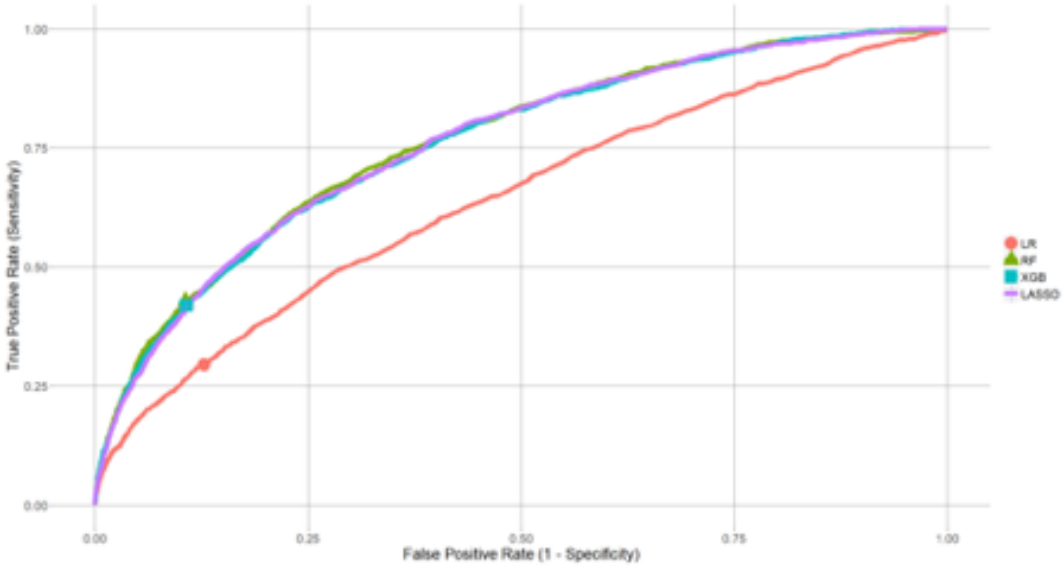


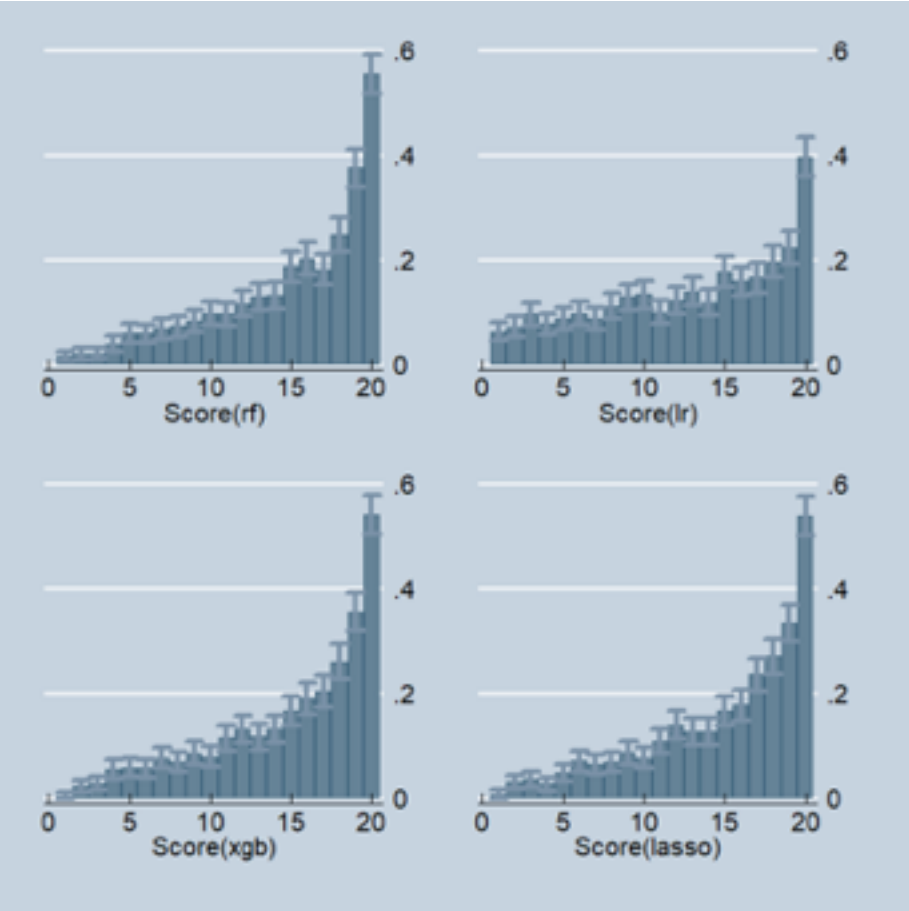
Figure 4 shows the outcomes by risk score for each of the models (for referrals in the validation partition only). The vertical axis shows the percentage of child-referrals that received a specific score (as noted in the horizontal axis) and in which a child was placed within two years in the test data and the 95 percent confidence intervals. Since Random Forest, XGBoost and LASSO produced the same AUCs, it is not surprising that they have very similar outcome rates by score.

5 LR is a logistic regression model; RF is a random forest model trained with 500 trees; XGBoost is an XGBoost model trained with 1000 trees, a learning rate of 0.01, a subsample ratio of columns for each tree of 0.8, a maximum depth of 14, a minimum child tree node of 1, a regularization alpha of 1e-05, a regularization lambda of 80, a class weight for imbalance of 4.433, a subsample ratio of the training instances of 0.9, all these parameters selected by grid-search.

Appendix A (continued)

For example, if we look at only those referrals where a child scored a 20, around 55 percent of those referrals will end up placed within two years – and that rate is the same across all models except for Logistic Regression.

FIGURE 4: **Rates of Placement Outcomes for Four Modeling Strategies (test data only)**



We also looked at fatalities and near fatalities to test whether there were any significant differences in the correlation between the scores and whether a child eventually experiences a fatality or near fatality. To do this, we estimated a logit regression where the dependent variable was equal to one if the child ended up having a fatality or near fatality that met the criteria for review under the provisions of Legislation Act 33 of Pennsylvania's Child Protective Services Law (CPSL) and zero otherwise. We restricted attention to children with a fatality event that occurred more than 50 days after the referral. Table 3, below, reports the results of this regression. The estimated marginal effects of a one unit increase in the predictive risk modeling (PRM) score (e.g., from 5 to 6) on the probability that the child will be a victim of a fatality or near fatality more than 50 days after the referral ranges from 0.074 per 1,000 to 0.059 per 1,000 for the

Appendix A (continued)

various models. All estimated effects are statistically significantly different from zero—but not statistically different from each other.

TABLE 3: **Marginal effect of a 1-unit increase in risk score on probability of a fatality or near fatality more than 50 days after the referral**

MODELING CHOICE	MARGINAL EFFECT (AVERAGE PER 1,000 AND 95% C.I)
LR	0.061 (0.025, 0.097)
LASSO	0.074 (0.040, 0.108)
RF	0.070 (0.035, 0.104)
XG Boost	0.059 (0.022, 0.095)

While the AUCs, outcome plots and mortality regressions provide general information about the accuracy of the algorithm across the range of scores, the more important metric for Allegheny County, given their protocols (as described above in Table 1), is to consider how well it serves to discriminate between high- and low-risk children.

Table 4 shows the positive predictive value (PPV) and true positive rate (TPR) for the four models with respect to the high- and low-risk protocols. The table is at the referral level and uses only the test data (that is, the referrals that were not used to build the models). The top part of the table shows the results for referrals which would have been flagged as high-risk by the tool, i.e., a referral where a child's score is greater than 17 and there is at least one child or victim on the referral who is aged 16 years or younger. The third row of the table shows the percentage of referrals that would be scored as high-risk by the protocol. Because the models identify different families as scoring greater than 17, the percentage of referrals that are identified as "high-risk" depends on the model. All models will flag around 25 percent of referrals as high-risk. The average placement rates for these referrals are between 35.4 percent for the Logistic Regression (LR) and 44.8 percent for the Random Forest. These rates are calculated at the referral level. For example, if the LASSO model were used to identify high-risk referrals, then 25 percent of referrals would be flagged; around 43 percent of all referrals would have at least one child in the referral who was placed within two years. These referrals would account for 55 percent of all referrals where at least one child is placed.

The low-risk protocol flags children as low-risk if the corresponding referral scores 10 or under, and all victims and children in the referral are more than 12 years old (i.e., intake date is after their 12th birthday). These referrals only account for between 2.9 and 9.2 percent of all referrals.

Appendix A (continued)

TABLE 4: **Comparison of Modeling Approaches for High- and Low-Risk Referrals**

	LR	LASSO	RF	XG BOOST
High-Risk Flag *(highest score on referral is greater than 17 and there is at least one child or victim on the referral who is aged 16 years or younger)*				
Proportion of referrals that receive the flag	23.41%	23.8%	25.1%	26.2%
Proportion of referrals flagged high-risk where child ends up placed within 2 years (PPV)	35.4%	47.6%	47.6%	46.2%
Proportion of all referrals where child ends up being placed, who are flagged	39.3%	53.7%	56.6%	57.4%
Low -Risk Flag *(highest score on referral is 10 or under, and all victims and children in the referral are more than 12 years old).*				
Proportion of referrals that receive the flag	9.2%	4.1%	2.8%	2.9%
Proportion of referrals flagged low- risk where child ends up placed within 2 years (PPV)	16.4%	7.6%	5.9 %	4.4%
Proportion of all referrals where child ends up being placed, who are flagged low risk	7.2%	1.5%	0.8%	0.6%

Table 5 shows the share of black children who are identified as high-risk and compares the performance across the models. Across all models, the share of black children flagged in high-risk referrals would be between 28 percent and 37 percent black (in the test sample). Rows 2 and 3 show the relative risk of being placed for black children vs. non-black children. This shows that conditional on race, the models are not miscalibrated in the sense that the relative risk of placement for black children is similar to that for non-black children. The relative risks are most similar for Lasso.

TABLE 5: **Comparison of Modeling Approaches for High-Risk by Race**

	LR	LASSO	RF	XG BOOST
Proportion of children flagged as high- risk who are black	28.1%	36.8%	36.1%	35.2%
Relative Risk of being placed if flagged as High- Risk and black vs. not flagged as High-Risk and black (95% c.i.)	1.95 [1.71, 2.22]	3.10 [2.73, 3.53]	2.72 [2.39, 3.08]	2.54 [2.24, 2.89]
Relative Risk of being placed if flagged as High-Risk and non-black vs. not flagged as High-Risk and non-black (95% c.i.)	1.70 [1.44, 2.02]	3.20 [2.70, 3.73]	3.30 [2.81, 3.88]	3.33 [2.84, 3.91]

Appendix A (continued)

Discussion of Model Choices

In deciding which methodology we should adopt, we looked at 1) overall performance and accuracy for the specific high-risk group that serves as the focus of the County's policy, and 2) equivalent levels of accuracy for black children vs. non-black children.

We also gave due consideration to pragmatic questions of implementation and ongoing quality assurance. Given the large number of databases that are being linked in the AFST V2, quality checks and ongoing model maintenance (e.g., to ensure that there is no feature drift) are critical.

We also analyzed whether the modeling methods resulted in differences in association between the fatalities/near fatalities and the AFST scores. We found that the scores generated by all models show positive correlation with the probability that a child was involved in an Act 33 fatality or near fatality more than 50 days after the score.

LASSO and Logistic Regression approaches, which consist of a simple set of weights, are easier to implement, while Random Forest and XG Boost, consisting of a sequence of linked trees, are hardest because of the difficulties with de-bugging the complex deployed algorithm.

The slight difference in PPVs by race suggests that non-black children are being given too high a score compared to black children. This phenomenon (that we first noted in Chouldecheva (2018)) is similar across all methods. However, when we consider the relative risk (conditional on race) with respect to the high risk protocols being implemented by the County, the models are choosing similarly risky groups.

APPENDIX B: WEIGHTED VARIABLES IN AFST V2

The weights of the model are available upon request from the Allegheny County Department of Human Services.

Definition of suffixes:

vict_othr	All other victim children named in this referral (other than the focal victim child who is being risk scored)
vict_self	The focal victim child being risk scored
prnt	The parent/guardian
perp	The alleged perpetrator. Please note, an individual on the referral could be included in multiple roles (e.g., an individual that is both the parent of the child and the alleged perpetrator).
chld	Other children named in the referral, but who are not identified as the victim

Weighted Variables LASSO:[6]

VARIABLE	DESCRIPTION
INFANT_VIC_NULL	=1 if the victim child <1 year of age at current referral; 0 otherwise
TOD_VIC_NULL	=1 if victim child is btw 1<=age<3; 0 otherwise
SC1_VIC_NULL	=1 if victim child btw 6<=age<9; 0 otherwise
VIC_AGE_SC2_NULL	=1 if victim child btw 9<=age<13; 0 otherwise
TEEN_VIC_NULL	=1 if victim child btw 13<=age<18; 0 otherwise
VIC_1_NULL	=1 if there's a single victim child in the referral; 0 otherwise
AGE_AT_RFRL_MISS_VICT_SELF	=1 if focal child has no age or invalid age; 0 otherwise
CHLD_3_NULL	=1 if there are 3 children involved in the referral who are not identified as victims of the referral; 0 otherwise
CHLD_AGE_INF	=1 if counts of the number of other involved children that are less than 1 year old at the time of referral; 0 otherwise
CHLD_VICTIM_VICT_SELF	=1 if focal child is specifically "Alleged Victim Child" (as opposed to just "Child"); 0 otherwise
FEMALE_NULL	= 1 if victim is female; 0 otherwise
BIO_DAD_NULL	=1 if victim in this referral has a bio dad identified in the relationship table; 0 otherwise
BIO_MOM_NULL	= 1 if victim in this referral has a bio mom identified in the relationship table; 0 otherwise
PERP_0_NULL	=1 if there is no perpetrator in the referral; 0 otherwise
PERP_3_NULL	=1 if there are 3 perpetrators in the referral; 0 otherwise
PERP_AGE_5564_NULL	counts of the number of perpetrators that are 55<=age<65
PERP_AGE_65_NULL	counts of the number of perpetrators that are more than 65
PERP_FEMALES_NULL	counts of the number of perpetrators that were female
PRNT_0_NULL	if there is no person identified as a parent in the referral
PRNT_AGE_19_NULL	counts of the number of parents that are 13<=age<20

6 The full set of variables is calculated for each child on the referral. The variable value is zero if the underlying data required to calculate the variable is missing. In many of the variable categories, an additional variable to indicate if data was missing was included.

Appendix B (continued)

VARIABLE	DESCRIPTION
PRNT_AGE_2024_NULL	counts of the number of parents that are 20<=age<25
PRNT_AGE_4554_NULL	counts of the number of parents that are 45<=age<55
PRNT_AGE_65_NULL	counts of the number of parents that are more than 65
PRNT_OVER2_NULL	if there are more than 2 individuals named on the referral identified as parents
IN_AJD_CHLD	= 1 if the child's MCI ID was created before the referral date; 0 otherwise
IN_AJD_OTH	= 1 if the person's MCI ID was created before the referral date; 0 otherwise
IN_AJD_VICT_SELF	= 1 if the focal child's MCI ID was created before the referral date; 0 otherwise
IN_HOUSEHOLD_NULL	= 1 if the victim is living in the mom's household; 0 otherwise (using InHousehold flag)
REF_PAST365_COUNT_OTH	aggregated no. of referrals in the past 365 days for all individuals involved with role of other (0 if missing)
REF_PAST365_COUNT_VICT_SELF	aggregated no. of referrals in the past 365 days for the focal child (0 if missing)
REF_PAST548_COUNT_VICT_SELF	aggregated no. of referrals in the past 548 days for the focal child (0 if missing)
REF_PAST90_COUNT_CHLD	aggregated no. of referrals in the past 90 days for all individuals involved with role of child (0 if missing)
REF_PAST90_COUNT_VICT_SELF	aggregated no. of referrals in the past 90 days for the focal child (0 if missing)
REFER_TIME_DAY_NULL	=1 if the intake time for the current referral is in the AM; 0 otherwise
PREVIOUS_RFRL_PERP	=1 if the perpetrator has prior referrals; 0 otherwise
PREVIOUS_RFRL_VICT_SELF	=1 if the focal child has prior referrals
FNDG_PAST365_COUNT_CHLD	aggregated no. of founded allegations in the past 365 days for all individuals involved with role of child
FNDG_PAST90_COUNT_CHLD	aggregated no. of founded allegations in the past 90 days for all individuals involved with role of child
SER_PAST180_COUNT_CHLD	aggregated no. of case openings in the past 180 days for all individuals involved with role of child
SER_PAST180_COUNT_PERP	aggregated no. of case openings in the past 180 days for all individuals involved with role of perpetrator
SER_PAST365_COUNT_CHLD	aggregated no. of case openings in the past 365 days for all individuals involved with role of child
SER_PAST548_COUNT_CHLD	aggregated no. of case openings in the past 548 days for all individuals involved with role of child
SER_PAST548_COUNT_OTH	aggregated no. of case openings in the past 548 days for all individuals involved with role of other
SER_PAST548_COUNT_PRNT	aggregated no. of case openings in the past 548 days for all individuals involved with role of parent
SER_PAST548_COUNT_VICT_OTHR	aggregated no. of case openings in the past 548 days for all individuals involved with role of other

Appendix B (continued)

VARIABLE	DESCRIPTION
SER_PAST548_COUNT_VICT_SELF	aggregated no. of case openings in the past 548 days for all individuals involved with role of victim
PLSM_NOW_NULL	= 1 if the referral was received during a placement episode; 0 otherwise
PLSM_PAST180_COUNT_NULL	victim's no. of placement episodes during the last 180 days
PLSM_PAST180_DUMMY_NULL	=1 if the victim was in placement during the last 180 days; 0 otherwise
PLSM_PAST365_COUNT_NULL	victim's no. of placement episodes during the last 365 days
PLSM_PAST365_DUMMY_NULL	=1 if the victim was in placement during the last 365 days; 0 otherwise
PLSM_PAST548_COUNT_NULL	victim's no. of placement episodes during the last 548 days
ALG_PR_12MONTHS_CNT_VICT_SELF	Count of number of total duplicated allegations (regardless of Allegation High Level Category) reported for child in prior 365 days to current referral.
ALGABS_CHLDBHVR_VICT_SELF	=1 if the focal child has an allegation in the Child Behaviors category on this referral; 0 otherwise
ALGABS_CRGVSUBABUSE_VICT_SELF	=1 if the focal child has an allegation in the Caregiver Substance Abuse category on this referral; 0 otherwise
ALGABS_IMMRISK_VICT_SELF	=1 if the focal child has an allegation in the Imminent Risks category on this referral; 0 otherwise
ALGABS_INADHOME_VICT_SELF	=1 if the focal child has an allegation in the No/Inadequate Home category on this referral; 0 otherwise
ALGABS_NEGLECT_VICT_SELF	=1 if the focal child has an allegation in the Neglect category on this referral; 0 otherwise
ALGABS_OTHER_VICT_SELF	=1 if the focal child has an allegation in the Other category on this referral; 0 otherwise
ALGABS_OTHREFSRC_VICT_SELF	=1 if the focal child has an allegation in the Other Referral Source category on this referral; 0 otherwise
ALGABS_PHYALT_VICT_SELF	=1 if the focal child has an allegation in the Physical Altercation category on this referral; 0 otherwise
ALGABS_PHYMALTRTMNT_VICT_SELF	=1 if the focal child has an allegation in the Physical Maltreatment category on this referral; 0 otherwise
ALGABS_PRNTCHLDCNFL_VICT_SELF	=1 if the focal child has an allegation in the Parent/Child Conflict category on this referral; 0 otherwise
ALGABS_SEXABUSE_VICT_SELF	=1 if the focal child has an allegation in the Sexual Abuse or Exploitation category on this referral; 0 otherwise
ALGABS_SEXCNTCTCHLD_VICT_SELF	=1 if the focal child has an allegation in the Sexual Contact Between Children category on this referral; 0 otherwise
ALGABS_TRUANCY_VICT_SELF	=1 if the focal child has an allegation in the Truancy category on this referral; 0 otherwise
ALGABS_UNWILLPRVDCR_VICT_SELF	=1 if the focal child has an allegation in the Unwilling or Unable to Provide Care category on this referral; 0 otherwise
ALGABSP_CHLDBHVR_VICT_SELF	total no. of prior referrals where the focal child had an allegation in the Child Behaviors category

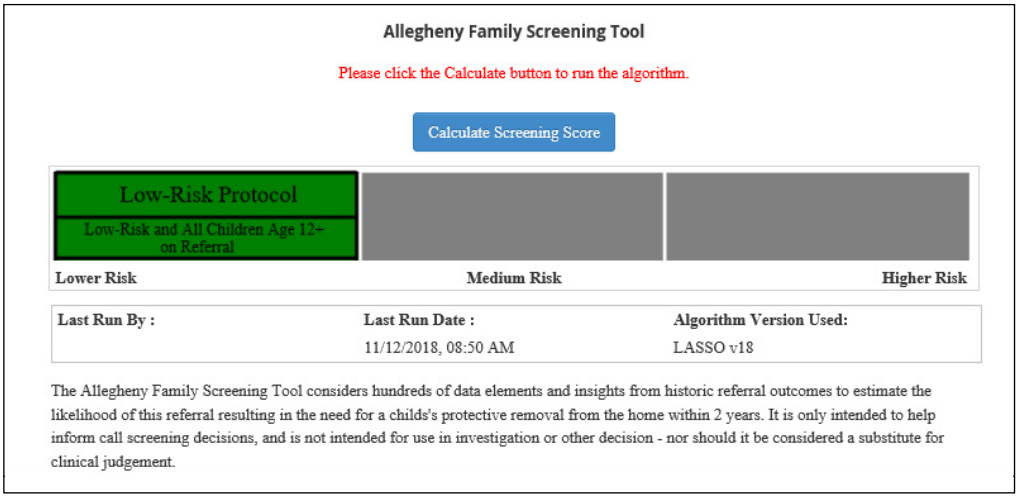
Appendix B (continued)

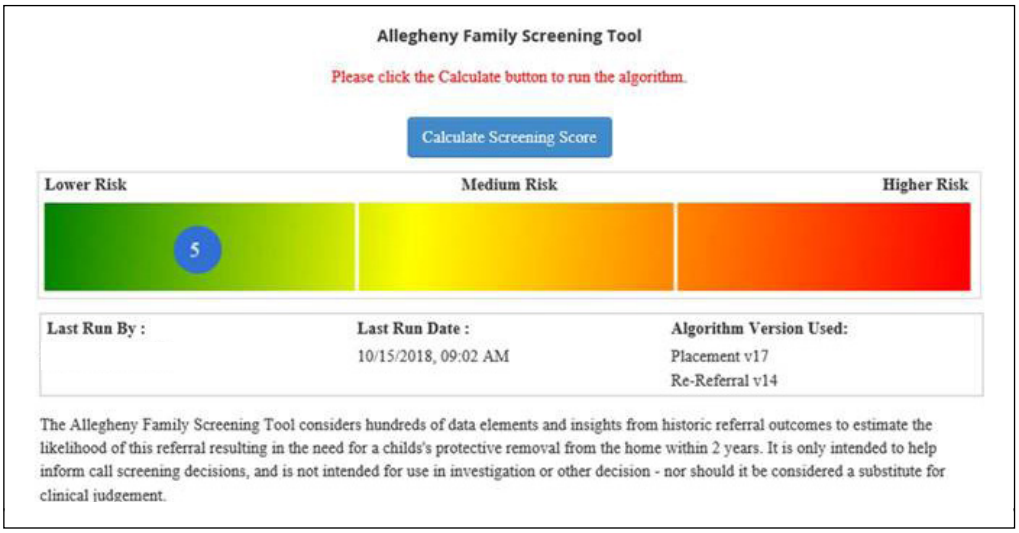
VARIABLE	DESCRIPTION
ALGABSP_CRGVSUBABUSE_VICT_SELF	total no. of prior referrals where the focal child had an allegation in the Caregiver Substance Abuse category
ALGABSP_INADPHYSCARE_VICT_SELF	total no. of prior referrals where the focal child had an allegation in the Inadequate Physical Care category
ALGABSP_MEDNEGLECT_VICT_SELF	total no. of prior referrals where the focal child had an allegation in the Medical Neglect category
ALGABSP_OTHREFSRC_VICT_SELF	total no. of prior referrals where the focal child had an allegation in the Other Referral Source category
ALGABSP_PHYALT_VICT_SELF	total no. of prior referrals where the focal child had an allegation in the Physical Altercation category
ALGABSP_PRNTCHLDCNFL_VICT_SELF	total no. of prior referrals where the focal child had an allegation in the Parent/Child Conflict category
BC_FEMALE_OR_MISS_VICT_SELF	=1 if gender of child on birth certificate record is female; 0 otherwise
BD_AGE_18_19_VICT_SELF	=1 if father's age was 18-19 at time of child's birth; 0 otherwise
BD_AGE_20_24_VICT_SELF	=1 if father's age was 20-24 at time of child's birth; 0 otherwise
BD_AGE_25_29_VICT_SELF	=1 if father's age was 25-29 at time of child's birth; 0 otherwise
BD_AGE_40PLUS_VICT_SELF	=1 if father's age was 40 or greater at time of child's birth; 0 otherwise
BD_EDUC_MISS_VICT_SELF	=1 if father's education was "Unknown" or missing. This includes actual "Unknown" values, null values, and any invalid values; 0 otherwise
BM_AGE_30_34_VICT_SELF	=1 if mother's age was 30-34 at time of child's birth; 0 otherwise
BM_AGE_35_39_VICT_SELF	=1 if mother's age was 35-39 at time of child's birth; 0 otherwise
BM_AGE_40PLUS_VICT_SELF	=1 if mother's age was 40 or greater at time of child's birth; 0 otherwise
BM_EDUC_BA_OR_HIGHER_VICT_SELF	=1 if mother's education is "Associate degree", "Bachelor's degree", "Master's degree" OR "Doctorate or Professional degree"; 0 otherwise
BM_EDUC_LESS_HS_VICT_SELF	=1 if mother's education is "8th grade or less" OR "9th–12th grade; No diploma"; 0 otherwise
BM_MARRIED_VICT_SELF	=1 if mother is married; 0 otherwise
BM_PAY_MEDICAID_VICT_SELF	=1 if source of payment for delivery was Medicaid; 0 otherwise
BM_PAY_OTHER_VICT_SELF	=1 if source of payment for delivery was other; 0 otherwise
BM_PAY_PRIVATE_VICT_SELF	=1 if source of payment for delivery was private insurance; 0 otherwise
BM_PR_LV_BIRTHS_4PLS_VICT_SELF	=1 if there were 4 or more previous live births; 0 otherwise
BM_SMKD_3MTH_PRIOR_M_VICT_SELF	=1 if cigarette smoking before pregnancy is missing; 0 otherwise
BM_SMKD_3MTH_PRIOR_VICT_SELF	=1 if cigarette smoking before pregnancy; 0 otherwise
BR_MED_PREG_INF_YES_VICT_SELF	=1 if any of infections present/treated; 0 otherwise
BR_MED_PREG_RF_YES_VICT_SELF	=1 if any of risk factors are present; 0 otherwise

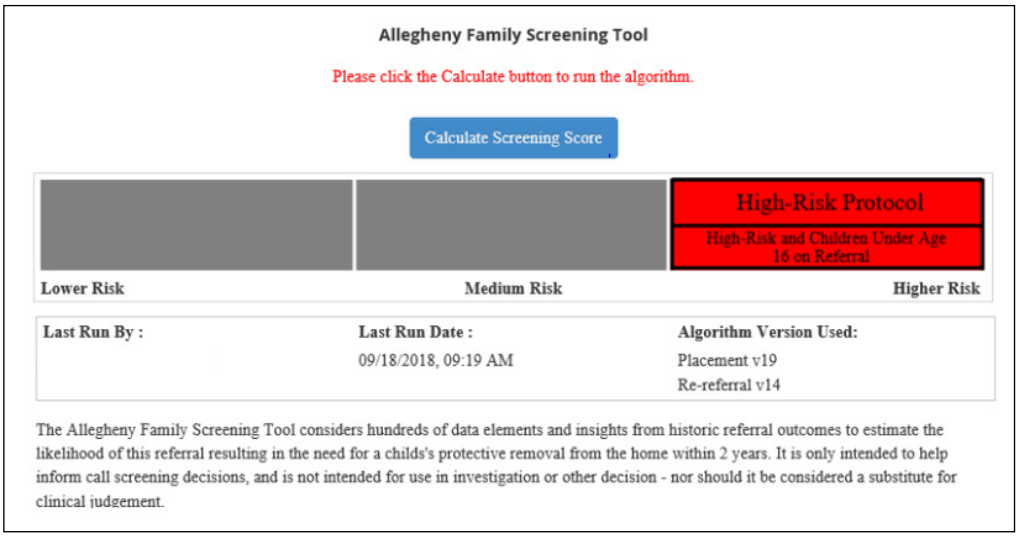
Appendix B (continued)

VARIABLE	DESCRIPTION
POVERTY_0_NULL	=1 if poverty rate = 0; 0 otherwise
POVERTY_30OVER_NULL	=1 if poverty rate is greater than 30; 0 otherwise
POVERTY_UNDER20_NULL	=1 if poverty rate is greater than 10 but under 20; 0 otherwise
POVERYRATE_NULL	=1 if no poverty rate available; 0 otherwise
ACJ_1_PER_PERP	% of months seen in Allegheny County Jail last 1 year
ACJ_1_PER_PRNT	% of months seen in Allegheny County Jail last 1 year
ACJ_1_PER_VICT_OTHR	% of months seen in Allegheny County Jail last 1 year
ACJ_1_PERP	total no. of months in Allegheny County Jail in the last year
ACJ_1_PRNT	total no. of months in Allegheny County Jail in the last year
ACJ_1_VICT_OTHR	total no. of months in Allegheny County Jail in the last year
ACJ_2_VICT_OTHR	total no. of months in Allegheny County Jail in the last 2 years
ACJ_3_PER_PRNT	% of months seen in Allegheny County Jail last 3 years
ACJ_3_PRNT	total no. of months in Allegheny County Jail in the last 3 years
ACJ_EVERIN_PERP	=1 if the person was in Allegheny County Jail before; 0 otherwise
ACJ_EVERIN_VICT_SELF	=1 if the person was in Allegheny County Jail before; 0 otherwise
ACJ_NOW_OTH	=1 if the person was in Allegheny County Jail at the time of the referral; 0 otherwise
ACJ_NOW_VICT_SELF	=1 if the person was in Allegheny County Jail at the time of the referral; 0 otherwise
JPO_1_CHLD	total no. of months in Juvenile Probation in the last year
JPO_1_PER_CHLD	% of months seen in Juvenile Probation in last 1 year
JPO_1_PER_VICT_SELF	% of months seen in Juvenile Probation in last 1 year
JPO_1_VICT_SELF	total no. of months in Juvenile Probation in the last year
JPO_3_PER_VICT_OTHR	% of months seen in Juvenile Probation in last 3 years
JPO_3_VICT_OTHR	total no. of months in Juvenile Probation in the last 3 years
JPO_EVERIN_OTH	=1 if the person was in Juvenile Probation before; 0 otherwise
JPO_EVERIN_PERP	=1 if the person was in Juvenile Probation before; 0 otherwise
JPO_EVERIN_PRNT	=1 if the person was in Juvenile Probation before; 0 otherwise
JPO_EVERIN_VICT_SELF	=1 if the person was in Juvenile Probation before; 0 otherwise
JPO_NOW_PERP	=1 if the person was in Juvenile Probation at the time of the referral; 0 otherwise
JPO_NOW_VICT_SELF	=1 if the person was in Juvenile Probation at the time of the referral; 0 otherwise
NO_BH_PERP	=1 if no behavioral health history for this person; 0 otherwise
NO_BH_PRNT	=1 if no behavioral health history for this person; 0 otherwise
NO_BH_VICT_SELF	=1 if no behavioral health history for this person; 0 otherwise

APPENDIX C: AFST V2 VISUALIZATIONS







APPENDIX D: HOSPITAL INJURY CLASSIFICATIONS

Hospital Event Injury Type and ICD-9 Codes

INJURY TYPE	ICD9 CODES
Injury from physical activity	E0000-E030; E927-E9282
Injury from transportation	E8000-E848; E9290-E9291
Accidental poisoning drugs/pharms	E8500-E8699; E9292
Injury from medical procedure	E8700-E8799
Accidental fall	E8800-E8889; E9293
Injury from smoke/fire	E8900-E899
Accident climatic or natural disaster	E9000-E903; E9294-E9295
Accident due to abandonment/neglect	E9040-E9049
Toxic reaction from animal or plant	E9050-E9069
Accidental drowning	E9100-E9109
Accidental obstruction respiratory	E911-E9139
Accident struck by object/person	E914-E9269; E9283-E9289; E9298-E9299
Adverse effect therapeutic drug use	E9300-E9499
Self-inflicted injury	E9500-E959
Physical assault	E9600-E978
Injury on accident or purpose	E9800-E989

SECTION 7

Frequently-Asked Questions

by the Allegheny County Department of Human Services

CONTENTS

Introduction *3*

Background *4*

What is the Allegheny Family Screening Tool (AFST) and how does it work? *4*

Who are the key partners and how were they selected? *4*

Has the local community been involved in the decision to use the AFST? *4*

How will the AFST be evaluated? *5*

AFST Version 1 *5*

What was the total cost of developing the AFST? *5*

What data does the AFST use? *5*

Doesn't the AFST just predict child welfare system decision-making? *5*

Does the AFST use race as a factor? *6*

Does the AFST use prior allegations of maltreatment as a factor? *6*

How accurate is the AFST? *6*

Has the AFST been validated? *7*

What did the research tell us about existing practice? *7*

What happens when there is missing/duplicate information? *7*

Is the AFST score assigned to a child/family permanently? *7*

What safeguards are in place to make sure the AFST is working appropriately? *8*

Will the County improve the AFST over time? *8*

How does the AFST compare to other approaches? *8*

Practice *8*

How many referrals come into the call screening center on an annual basis? *8*

What is the number of call screeners on staff? *8*

What is the average length of time devoted to each screening call? *9*

Who gets an AFST score and how? *9*

Are there some children for whom an AFST score can't be generated? *9*

Who has access to the AFST score? *9*

Does a certain AFST score make screening-in (for an investigation) mandatory? *9*

Will caseworkers be afraid to 'defy the score?' *10*

How do the AFST and the County minimize the risk of stigma? *10*

Are AFST scores higher for black children? *10*

Can the AFST help to reduce unwarranted variation in decision making? *10*

Does involvement in services always increase the AFST score? *11*

Outcomes *11*

Does a "mandatory screen-in" score always mandate an investigation? *11*

Has the AFST increased the number of investigations? *11*

What are the screen-in rates by category? *11*

Have more families been accepted for service since implementation of the AFST? *11*

What is the likelihood that an investigation leads to a placement? *12*

Process Evaluation *12*

What data collection methods did HZA use in its process evaluation? *12*

How well did staff feel the training prepared them to use the AFST? *13*

What aspect of the training was found to be most helpful? *13*

How well do call screeners understand the AFST? *13*

Are call screeners confident in the AFST's ability to accurately assess the risk of a future referral or out-of-home placement? *13*

Have there been any technical issues related to implementation of the AFST? *13*

Did DHS effectively engage and communicate with external stakeholders about the development of the AFST? *13*

How easy is it to navigate/use the AFST? *14*

How useful is the graphic display of the score (in the form of a thermometer)? *15*

Do call screening staff conduct a more thorough data search (either in ClientView or in child welfare's Key Information and Demographics System) when the AFST is high? *15*

What concerns do call screeners have about the AFST? *15*

Do call screeners anticipate that the AFST will have an impact on practice? *15*

Is the AFST creating a more data-driven culture at DHS? *15*

Are call screeners using the AFST to inform their recommendations? *15*

What recommendations emerged from the process evaluation? *16*

Impact Evaluation *16*

Where can I find the full impact evaluation report? *16*

Where can I find a summary of the evaluation? *16*

Who conducted the impact evaluation? *16*

What evaluation methods were used? *16*

What time period does the evaluation cover? *17*

What were the main evaluation findings? *17*

Will the County continue the evaluation? *18*

AFST Version 2 *18*

Why were changes made to the original model, operating from August 2016 to November 2018? *18*

What changes were made to the methodology for AFST Version 2? *19*

What modeling methodology is used in AFST V2? *19*

Was the model validated? *20*

Were there accuracy improvements in AFST V2? *20*

Implementation Lessons *20*

Do the process and impact evaluations cover everything you've learned since you started building and using the AFST? *20*

What are some of the technical lessons learned during AFST implementation? *20*

What are some of the lessons learned (and still evolving) about the use of the AFST in practice? *22*

What are some reflections around the policies associated with AFST? *24*

Did these technical, practice and other challenges impact the results of the evaluation? *25*

Has the policy landscape around the implementation of predictive risk modeling changed since DHS began this work? *25*

You have reported outcomes for the first year of implementation—can you provide results for the full period under AFST Version 1? *25*

APPENDIX A: AFST Screening Score Data *26*

INTRODUCTION

In August 2016, the Allegheny County Department of Human Services (DHS) implemented the Allegheny Family Screening Tool (AFST), a predictive risk modeling tool designed to improve child welfare call screening decisions. The AFST was the result of a two-year process of exploration about how existing data could be used more effectively to improve decision-making at the time of a child welfare referral. The original model (Version 1) was utilized from August 2016 through November 2018. An updated model (Version 2) is now being used. For more information about the AFST, see here.

The process began in 2014 with a Request for Proposals and selection of a team from Auckland University of Technology led by Rhema Vaithianathan and including Emily Putnam-Hornstein from University of Southern California, Irene de Haan from the University of Auckland, Marianne Bitler from University of California – Irvine and Tim Maloney and Nan Jiang from Auckland University of Technology. Prior to implementation, the model was subjected to an ethical review by Tim Dare of the University of Auckland and Eileen Gambrill of the University of California-Berkeley. Upon the conclusion of this review, to which DHS prepared a response, the County proceeded with implementation. Concurrent with this process was the issuance of a second Request for Proposals, at the end of 2015, for an impact and process evaluation of the model. Contracts were awarded to Stanford University (impact evaluation) and Hornby Zeller Associates (process evaluation).

A report on the development of the AFST,[1] prepared by Rhema Vaithianathan, PhD; Nan Jiang, PhD; Tim Maloney, PhD; Parma Nand, PhD; and Emily Putnam-Hornstein, PhD, was published in April 2017 and a report on the development of the AFST Version 2 was published in April 2019. The following Frequently-Asked Questions are presented as a quick reference for those interested in highlights from these publications as well as the evaluations and should be considered within the context of the full publications. Page numbers are provided throughout the document, indicating where the reader may find more detailed information.

1 Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening

BACKGROUND

What is the Allegheny Family Screening Tool (AFST) and how does it work?

The AFST was developed to support one key decision in the child welfare process: whether or not to screen-in a referral for investigation.

To generate the AFST scores, the AFST uses more than 100 predictive factors for each child on the referral. In V1 of the AFST, these factors were then weighted through a logistic regression model to calculate two AFST scores (ranging from 1–20) for each child: the risk of placement within two years if the referral is screened-in and the risk of re-referral within two years if the referral is screened-out.[2] Call screeners and supervisors see the maximum AFST score from the referral. For example, if there are two children on the referral and one has a maximum risk score of 12 and the other has a maximum risk score of 16, the call screener will see a score of 16.

2 This methodology was altered in V2 of the AFST; see page 18 of this FAQs document for more information about V2.

It should be noted that while in some settings machines have been used to make decisions that were previously made by humans, this is not the case for the AFST. It was never intended or suggested that the algorithm would replace human decision-making. Rather, the AFST should help to inform, train and improve the decisions made by the child welfare staff.

Who are the key partners and how were they selected?

The Allegheny County Department of Human Services (DHS) issued a Request for Proposals (RFP) in 2014, to design and implement a system of decision-support tools and predictive analytics for human services.[3]

3 Decision Support Tools and Predictive Analytics in Human Services RFP

We received 15 proposals in response to the RFP. After review by an evaluation committee, researchers from Auckland University of Technology (AUT), University of Southern California (USC), University of California-Berkeley and University of Auckland were awarded the contract and conducted the work. The research team was led by Rhema Vaithianathan (AUT).

Has the local community been involved in the decision to use the AFST?

Community engagement has been a priority for the County throughout the project. The County sought input from the community through various meetings, including six project-specific meetings. Three were held at early stages of the project to collect feedback from key external stakeholders and funders. DHS then held three open community meetings where over 30 stakeholder groups (including the Courts and the ACLU) were invited to discuss the work to date, implementation timeline and results. Additionally, DHS shared project updates with existing community networks including the Children's Cabinet and the Children, Youth and Families Advisory Board, and through the DHS Speaker Series. Feedback from these community meetings has influenced the project throughout its development.

How will the AFST be evaluated?

An RFP for two independent evaluations of the AFST (process and impact) was issued in 2015.[4] Hornby Zeller Associates was selected to conduct an process evaluation and Stanford University was selected to conduct an impact evaluation. The process evaluation is available here. The impact evaluation focused on whether the AFST increased the accuracy of decisions, reduced unwarranted variation in decision-making and reduced disparities, and also examined overall referral rates and workload. A summary of the impact evaluation can be viewed in Section 5 and the full impact evaluation can be viewed in Section 6.

4 Evaluation of a Predictive Risk Modeling Tool for Improving the Decisions of Child Welfare Workers RFP

AFST VERSION 1

What was the total cost of developing the AFST?

The total cost was $1,185,424, as detailed below:

VENDOR	SERVICE	TOTAL
Auckland University of Technology	Methodology and Model Design	$500,000
Deloitte	Technology	$280,000
Stanford University	Impact Evaluation	$310,000
Hornby Zeller Associates	Process Evaluation	$95,424
TOTAL		**$1,185,424**

What data does the AFST use?

The AFST uses information from DHS's integrated data system that links administrative data from 21 sources including child protective services, publicly funded mental health and drug and alcohol services, and bookings in the County jail. Please **see page 11** of the methodology and implementation report for additional information on the data used. See the section about AFST Version 2 for information about changes that have been made to data sources since implementation.

Doesn't the AFST just predict child welfare system decision-making?

A challenge is to identify outcomes to predict that are truly independent of the system and not too rare to be predicted.

The first adverse outcome predicted by the AFST is placement within two years of screen-in. Because placements are determined by a judge, and all parties (parents, children and County) are represented by attorneys, a placement outcome is reasonably independent of the County child welfare system.

The second adverse outcome that the AFST predicts — re-referral after an initial referral has been screened-out — is independent of the County child welfare system because referrals come from the community. In AFST Version 2, we eliminated the second outcome. See the FAQs section related to Version 2 for information about this change.

Does the AFST use race as a factor?

No. The County made the decision not to include race as a factor in the AFST because including race does not improve the accuracy of the score. This doesn't mean, however, that other variables in the tool aren't correlated with race. There are other predictors that are correlated with race due to potentially institutionalized racial bias (e.g., criminal justice history) that would imply that race is still a factor. For this reason, continued monitoring of application of the model with regard to racial disparities should be undertaken.

Please **see page 29** of the methodology and implementation report for additional information on the impact of race as a predictor and *Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County.*

Does the AFST use prior allegations of maltreatment as a factor?

Yes, because historical data tell us that previous reports of maltreatment, substantiated or not, have predictive power (there is no factor included in the model that does not have significant predictive power). However, Title 23 Sec. 6337 of the PA Consolidated Statutes and the Pennsylvania Department of Human Services provide guidance as to the length of time that allegation reports remain in KIDS (the child welfare case management system), one of the sources queried by the algorithm. Once a report is expunged, the algorithm is no longer able to access it and it is therefore not included in the algorithm. Expungement timelines range from one year and 120 days (for unfounded reports) to five years and 120 days after receipt of the report or closure of services (or until the subject child is 23) for founded reports.

How accurate is the AFST?

Measuring the accuracy of predictive tools is not simple; however, at rollout, the accuracy of the AFST for predicting whether a child would be placed in care within two years after being referred and screened-in for investigation was 70 percent (if measured by area under the curve (AUC)[5].

The new model is better than digital mammography in asymptomatic women.

Please **see page 15** of the methodology and implementation report for additional information on model performance and AFST Version 2 for updated information on model performance.

Has the AFST been validated?

In addition to assessing the accuracy of the AFST in predicting placement and re-referral, the research team also conducted an external validation looking at the likelihood of hospital events (emergency department visits and inpatient admissions). Findings show that over a broad range of injury types there is a positive correlation between the placement scores generated by the AFST at referral and the rate of hospital events.

For example, those children with a placement risk score of 20 (the highest possible score) have a hospital event rate for self-inflicted injury or suicide of 0.65 percent compared to 0.03 percent

5 This figure is an update of a previously higher reported figure in the FAQs that over-stated the AUC because of some technical issues related to the way in which the data was split. For more technical details, please see Chouldechova, Alexandra, et al. "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions." Conference on Fairness, Accountability and Transparency. 2018."

for those with a placement risk score of 1 (the lowest possible score). That is, a child who scores a 20 at referral is 21 times more likely to be hospitalized for a self-inflicted injury than a child who scores a 1.

Please **see page 19** of the methodology and implementation report for additional information on the hospital validation study. Additional information is available on **page 7** of the Methodology, Version 2.

What did the research tell us about existing practice?

Prior to introduction of the AFST, call screeners could access and use historical and cross-sector administrative data related to individuals associated with a report of child abuse or neglect through Client View, a front-end application to the integrated data system. Call screeners were required to review all relevant information related to a referral and provide it to the call screening supervisor so that a screen-in/screen-out decision could be made. However, it was challenging for call screeners to efficiently access, review and make meaning of all available records. The AFST provides a consistent way to access and weight the available information to predict the risk of future adverse events for each child on the referral.

Researchers found that existing practice had screened out one in four children who the model would screen-in due to their score. For these children, who the model scored as highest risk, 9 in 10 were re-referred (if screened out) and half were placed in foster care (if screened in) within two years. Forty-eight percent of the lowest-risk cases were screened-in with only one percent of these referrals leading to placement within two years.

What happens when there is missing/duplicate information?

The AFST leverages a probabilistic matching algorithm to catch as many duplicate IDs as possible. This method, however, does not capture all duplicate IDs for the same person and, thus, it is possible for an AFST score to exclude data held on a second ID. Efforts to minimize duplicate client records are ongoing.

Is the AFST score assigned to a child/family permanently?

No, because the AFST score will change as underlying data change. The County will retain AFST scores for quality assurance and evaluation purposes.

What safeguards are in place to make sure the AFST is working appropriately?

Immediately before the AFST was put into operation, researchers validated the scores generated by the DHS Data Warehouse (for individuals in historical, de-identified data) by generating scores for the same individuals in the research environment, to ensure that the Data Warehouse was accurately running the AFST. Since implementation, County child welfare leadership has been reviewing monthly quality assurance reports to monitor the performance of the AFST.

AFST scores are securely stored and cannot be manually altered by call screeners. However, as an additional quality assurance check, DHS has added functionality to the AFST that allows workers to report feedback on scores that seem wrong/surprising to them.

The independent impact evaluation and process evaluation highlighted some issues, as did the experience of call screeners and supervisors.

Will the County improve the AFST over time?

The AFST has already been rebuilt once by the research team since it came into use in August 2016, taking learnings from practice and using those to optimize how the AFST scores are generated. In 2018, the County built Version 2 of the model, which included improvements identified by process and impact evaluations. See FAQs related to Version 2 in this document and the Methodology Version 2 report for details about the updates.

How does the AFST compare to other approaches?

The AFST has a similar purpose to other decision-support tools like the Structured Decision Making tool (SDM), but the AFST creates a score without the reliance on manual data input that is required for SDM. For the highest category of risk, the AFST outperformed the SDM model.

Please **see page 24** of the methodology and implementation report for additional information on comparing the model to SDM (including a validation study [Dankers and Johnson, 2014]), and rule-based/threshold approaches.

PRACTICE

How many referrals come into the call screening center on an annual basis?

In 2017, the call screening center received 15,768 referrals, of which 11,751 were GPS allegations.

What is the number of call screeners on staff?

As of April 2019, there were 23 call screener positions. The number of screeners working at a given time depends on the day, ranging from 4 on weekend evenings to 15 on weekday afternoons.

What is the average length of time devoted to each screening call?

A typical referral takes 30 to 60 minutes to process.

Who gets an AFST score and how?

All children involved in an allegation of maltreatment,[6] regardless of whether they are described as the victim or not, will be included in the AFST score; that is, all children living in the same household or added to the case by the call screener. When an allegation of maltreatment is received and the call screener enters details into the child welfare case management system (KIDS), a click will automatically generate the AFST score. Call screeners and call screening

6 The AFST is intended to assist in decision-making for CPS referrals; any allegation meeting CPS criteria is immediately investigated (state-mandate).

supervisors are required to generate the AFST score prior to finalizing a screening decision.

Are there some children for whom an AFST score can't be generated?

Yes, those not known to the system and those for whom not enough data are held in the Data Warehouse. The County has determined that the AFST will only be used to screen for risk when data that goes beyond demography (e.g., age, gender, address) are held for one or more person associated with the allegation. If only demographic data are held for all individuals, then the allegation will be assessed using the existing approach (no AFST score will be generated). As of April 2017, approximately 10 percent of incoming referrals were not generating an AFST score.

Who has access to the AFST score?

Only the call screener and call screening supervisor have access to the AFST score. If and when a referral moves to the investigation stage, investigations staff cannot access any AFST score. The Courts also do not have access to the AFST score. DHS is considering the value and appropriateness of changing this policy.

Please **see page 26** of the methodology and implementation report for additional information on the implementation of the AFST score.

Does a certain AFST score make screening-in mandatory?

The AFST flags some scores as "mandatory screen-ins."[7] The threshold for the mandatory screen-in was determined solely by the placement score and designed to capture as many of the children at heightened risk of abuse-related fatal or near-fatal injuries (Act 33 Events) as possible. The model includes functionality that allows call screening supervisors to override the "mandatory screen-ins" at their discretion; overrides are documented and reviewed.

Please **see page 26** of the methodology and implementation report for additional information on mandatory screen-ins.

Will caseworkers be afraid to 'defy the score?'

The only caseworkers who make screen-in/screen-out decisions are the call screening supervisors. They consider all information provided by the call screeners, including details shared during the call, by the person alleging abuse or neglect, the score generated by the AFST and recommendations from the call screener.

Screening decisions are not in any way 'dictated' by the AFST. Call screening supervisors have full discretion over call screening decisions, regardless of generated AFST scores, and call screening decisions are not required to align with the AFST score. In the AFST's first full year of operation, just 63 percent of referrals with a "mandatory screen-in" score were actually screened-in for an investigation. Conversely, even the lowest AFST scores had about a 30 percent screen-in rate.

7 The term "mandatory screen-in" is enclosed in quotations to reflect the fact that call-screening supervisors may override the score.

How do the AFST and the County minimize the risk of stigma?

No system can entirely remove the chance of screening-in some of the 'wrong' children, so wrongly stigmatizing them. The ethicists suggest, however, that we must then take a comparative view: Is the proposed tool as good or better than the existing approach, when it comes to minimizing the risk of stigma? Compared to the existing system, the AFST is expected to increase accuracy and consistency of decision-making, which means wrongful stigma is expected to be reduced. The impact evaluation assesses this.

In particular, the County will work to minimize stigmatization by carefully controlling access to AFST scores and providing appropriate training that aims to reduce stigmatization and ensures that call screeners are aware of the possibility of false positives/negatives and understand the risk of confirmation bias.

Are AFST scores higher for black children?

The AFST model does not apply any weights based directly on race. However, race is associated with many of the underlying data used by the model, so it is not surprising that the tool's scores have been slightly higher for black children compared to white children. For example, up until the end of 2017, 47% of black children received a "high"-range score (15–20), compared to 39% of white children. Conversely, 18% of white children have received a "low"-range score (1–9), compared to 10% of black children. Some degree of racial disproportionality has already been identified at child welfare decision points in prior published analyses, including at call screening. Whether or not the AFST has any impact (positively or negatively) on the degree of variation associated with child race is a key focus of the impact evaluation. See Methodology, Version 2 for an update.

Can the AFST help to reduce unwarranted variation in decision making?

Whether or not the AFST reduces unwarranted variation in decision-making (such as by race/gender, or variation between individual decision-makers) is a key focus of the impact evaluation. Results are available in the impact evaluation report (Section 6 of this packet), the impact evaluation summary (Section 5), and the AFST Version 2 FAQs on **page 18** of this document.

Does involvement in services always increase the AFST score?

No. For example, for 45% of families, receiving of public benefits (e.g., SNAP, TANF) is, in fact, protective. That is, for those families, receiving those services was associated with lower scores than for similar families that did not receive those services.

It is important to note that the fact of receiving a benefit (of any kind) is not of itself associated with a positive or negative effect on the AFST score. Moreover, receiving assistance in a particular service area is not, of itself, associated with a positive or negative effect on the score. The effect depends on which individual on the referral received the service, what type of service it was, and the intensity, duration and recency of the service.

OUTCOMES

Does a "mandatory screen-in" score always mandate an investigation?

No. In fact, with AFST V1, more than one-third of children classified as highest risk by the AFST were screened out by the intake manager.

Has the AFST significantly increased the number of investigations?

In absolute terms, the percentage of calls screened in during the first year of the tool has increased by less than a percentage point. Whether this resulting screen-in rate is higher or lower than it would have otherwise been in the absence of the tool is one thing the impact evaluation hopes to more thoroughly investigate.

What are the screen-in rates by category?

For AFST V1, which was in use from August 2016 through November 2018, screen-in rates by category were as follows:

SCORE CATEGORY	PERCENT SCREENED-IN FOR INVESTIGATION
Mandatory	61%
High	47%
Medium	42%
Low	31%
No Score	23%
Total	41%

Have more families been accepted for service since implementation of the AFST?

As a percentage of new General Protective Services referrals screened-in for the investigation, the accept-for-service rate was about 39% for AFST Version 1 (in use from August 2016 through November 2018)—about a five-percentage-point rise from a comparable year of data prior to the tool's implementation. It is important to note that workers investigating a referral are not able to access the referral's score according to the AFST, and investigative practice does not vary in any way based on a referral's score.

What is the likelihood that an investigation leads to a placement?

Under Version 1 of the AFST, about 9% of GPS referrals screened in for investigation led to at least one child being removed in the following 90 days.

PROCESS EVALUATION

What data collection methods did HZA use in its process evaluation?

HZA utilized interviews, surveys and data analysis to complete the process evaluation.

Interviews were conducted prior to implementation of the AFST (in July 2016) and four months after implementation (in December 2016). The July 2016 interviews were conducted with 23 DHS administrators and staff, and were designed to learn about a) their involvement in the implementation of the AFST, b) steps taken to prepare call screening staff to use predictive risk modeling to inform their decision-making, and c) the call screening process as it existed prior to implementation of the AFST. The December 2016 interviews were conducted with DHS stakeholders (child welfare staff, staff from the DHS Office of Analysis, Technology and Planning), as well as representatives from community service providers, advocacy groups, foundations and family court. DHS staff were asked about a) their involvement in implementing the AFST, b) the training they received, and c) how the AFST informs or impacts their work. External stakeholders were asked about a) their awareness of DHS's efforts to implement predictive risk models, b) their hopes for what the AFST would accomplish, and c) the successes and challenges they expected DHS to face.

A ***web-based survey*** was administered to call screeners approximately two months post-implementation (September 2016), and a follow-up survey was administered in February 2017 to account for improvements that had been made to the AFST. Using a series of Yes/No and Likert scale questions, call screeners were asked about the training they received, the functionality of the tool, visualization of the scores and the impact of the tool on their decision making. Several open-ended questions were also asked to gather input on what could be done to improve the use of the tool and the training provided to prepare staff to use it.

Data analysis consisted of 1) quantitative analysis of summary statistics, frequency counts and percentages and 2) qualitative analysis of the common themes and items of importance from the interviews and open-ended survey questions. Using a grounded theory approach, the results of the qualitative analysis described the implementation process from the perspective of the stakeholders.

See page 5 of the HZA evaluation report for more detail on the evaluation methods.

How well did staff feel the training prepared them to use the AFST?

The survey administered to call screeners two months after implementation showed that 82% felt somewhat (38%) or very well (44%) prepared to use the AFST. Only six percent reported being "limitedly" prepared and none reported that they were not at all prepared. No opinion was expressed by 13% of responders. By the time the follow-up survey was administered, 100% of respondents reported being adequately prepared to use the tool.

What aspect of the training was found to be most helpful?

Most helpful components were Information about how predictive analytics was to be applied in Allegheny County (36%), use of case scenarios (29%), overview of predictive risk modeling (21%), and overview of changes to KIDS and policy/practice (7% each).

How well do call screeners understand the AFST?

The follow-up survey included a series of questions designed to gauge screeners' understanding of the AFST. Ninety-four percent both understand what the score is predicting and how it should inform screening decisions. Eighty-nine percent understand the content of the data sources used to produce the score.

Are call screeners confident in the AFST's ability to accurately assess the risk of a future referral or out-of-home placement?

Half of call screeners said they were confident of the AFST's ability to assess risk and 61 percent were confident in the research that went into its development. Lack of confidence in the AFST's ability to predict risk seemed to stem from its inability to take expected improvement or individual circumstances into account; for example, when families are receiving services that are improving their situation.

Have there been any technical issues related to implementation of the AFST?

Nearly three-quarters of call screeners noted that they occasionally encounter a score that seems inaccurate, with an additional 11 percent frequently encountering an inaccurate score. In response, they either notify a supervisor, review and use available data, or contact technology staff.

Two early technical issues related to missing or duplicate Master Client Index numbers, were corrected in November 2016. However, an ongoing issue is that the system is reportedly slow and sometimes times out before generating a score.

Did DHS effectively engage and communicate with external stakeholders about the development of the AFST?

External stakeholders appreciated DHS's efforts to educate and inform them about the purpose, development and implementation of the AFST. They felt positive about the tool, its potential to improve decision making, and DHS's plans for implementation. A desire for ongoing communication was noted.

How easy is it to navigate/use the AFST?

Over 60 percent of respondents found the AFST easy or very easy to use, although this response declined between the initial and follow-up surveys (from 69% to 61%). Slightly more than 30 percent of respondents to both surveys were neutral about this question while six percent of respondents to the follow-up survey found the tool difficult to use.

How useful is the graphic display of the score (in the form of a thermometer)?

Responses to this question were mixed, with 44 percent responding that the thermometer was helpful or somewhat helpful, 38 percent reporting no opinion and 19 percent reporting that it was not helpful or helpful only on a limited basis.

Do call screening staff conduct a more thorough data search (either in ClientView or in child welfare's Key Information and Demographics System) when the AFST is high?

More than 60 percent of survey respondents reported that they "rarely" or "never" conduct an additional search, with full-time screeners more likely to conduct additional searches. Most call screeners did not conduct additional searches because the AFST score is already based on those data or because they had already completed searches in the Data Warehouse earlier in the process.

What concerns do call screeners have about the AFST?

Call screener concerns related mostly to the tool's inability to incorporate human judgement into the score or to recognize information that needs to be updated, thus generating a score that inaccurately portrays a family's actual circumstances.

Do call screeners anticipate that the AFST will have an impact on practice?

Between the first and second surveys, the percentage of those who anticipated no impact decreased from 50 percent to 44 percent. The percentage of those who thought the AFST would strengthen practice remained consistent at 44 percent. There was an increase in the percentage of those who thought the tool would diminish practice (from 6% to 11%).

Is the AFST creating a more data-driven culture at DHS?

Sixty-one percent of respondents to the follow-up survey agreed that the tool is creating a data-driven culture. Considering this finding along with the impact finding (previous question) might indicate that call screeners already thought that DHS's culture was data-driven (i.e., based on good screening practices).

Are call screeners using the AFST to inform their recommendations?

By the time of the follow-up survey, 72 percent of call screeners reported using the tool at least occasionally; only 11 percent always use it, while another 28 percent almost always use it. Whereas this percentage increased slightly from the initial survey (at 69%), the percentage of those who always use the tool decreased and the percentage of those using it occasionally or almost always both increased.

What recommendations emerged from the process evaluation?

HZA made the following recommendations in response to the evaluation results:

1. Maintain transparent communication with internal and external stakeholders.
2. Increase user buy-in.
3. Continue to resolve technical issues as they arise, documenting solutions.
4. Develop implementation benchmarks to foster buy-in and promote use of the tool for decision-making.

See page 19 of the HZA evaluation report for more detail about the recommendations.

IMPACT EVALUATION

Where can I find the full evaluation report?

To read the full technical report, please see: Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office.

Where can I find a summary of the evaluation?

To read, please see: Impact Evaluation Summary.

Who conducted the impact evaluation?

Stanford University was awarded the contract for the impact evaluation through a competitive process (Evaluation of a Predictive Risk Modeling Tool for Improving the Decisions of Child Welfare Workers RFP). The Request for Proposals was issued in December 2015; we received seven proposals and the County made its selection in early 2016. A report describing the results of the impact evaluation was finalized in March 2019. Two peer reviewers provided critical feedback on drafts of the report.

What evaluation methods were used?

Stanford University used a set of methodologically strong, quasi-experimental methods (e.g., interrupted time series analyses, generalized linear models). Quasi-experimental methods refer to a type of evaluation approach used when it is not possible or desirable to implement a randomized controlled trial (RCT). While less robust than a gold-standard RCT, carefully designed quasi-experimental methods are considered the next-best approach to testing program impact. The County decided not to pursue an RCT primarily for practical reasons.[8]

More specifically, evaluators used the following tests:

- **Unadjusted Population Means.** The simplest comparison performed was a comparison of unadjusted means for the Pre- and Post-AFST periods, testing whether they are statistically different from one another using a two-sided t-test of equality of means.
- **Interrupted Time Series Analysis (ITSA).** Changes in the level and trend of monthly rates of each outcome during the Pre- and Post-AFST periods were assessed using an Interrupted Time Series Analysis. In this evaluation, the ITSA measures changes in both the level and slope of each outcome in the Post-AFST months in relation to the Pre-AFST months.

8 The State of Colorado has contracted with Cornell University to conduct an RCT of their implementation of a similar predictive risk model implemented at the child welfare hotline in Douglass County, Colorado.

The ITSA approach captures population-level changes in outcomes and trends after a policy change (in this case, the implementation of the AFST) in comparison to the levels and trends prior to that change.

- **Child-Level Multivariate Regression Analysis.** Finally, the evaluators used multivariate individual-level regression analyses to assess the impact of the AFST on the predicted level of each outcome Pre- and Post-AFST, while adjusting for child and household characteristics. These analyses focus on estimates of the average effect of the AFST, adjusting for evolving case mix over time. The predictive margins presented in the evaluation can be interpreted as the average outcome if all children in the sample were in either the Pre-AFST or the Post-AFST time-frame, holding all other control variables constant.

What time period does the evaluation cover?

The evaluation consists of outcome comparisons for two groups of children: (1) the approximately 31,000 children who were referred for alleged maltreatment during the 18-month period before the AFST was implemented (Pre-AFST: January 1, 2015 through July 31, 2016) and (2) the approximately 34,000 children reported after the AFST was fully implemented (Post-AFST: December 1, 2016 through May 31, 2018). Outcomes for both groups (Pre-AFST vs. Post-AFST) were examined for 15 to 17 months after the initial maltreatment report was received.

What were the main evaluation findings?

1. **Overall, the AFST did not lead to increases in the rate of children screened-in for investigation.** Use of the tool appears to have resulted in a different pool of children screened-in for investigation (including more children who needed intervention supports, see finding 2 below). But from a workload perspective, there was no significant increase in the number or proportion of children investigated among all children referred for maltreatment.
2. **Implementation of the AFST increased the identification of children determined to be in need of further child welfare intervention.** Use of the tool led to an increase in screen-in rates for "higher-risk" children who needed intervention supports. Specifically, there was a statistically significant increase in the proportion of children screened-in who then had a child welfare case opened or, if no case was opened, were re-referred within 60 days. (Please note that investigators and supervisors making case opening decisions remained blind to the score.)
3. **Use of the AFST did not lead to decreases in re-referral rates for children screened-out without investigation.** Re-referral rates among children screened-out stayed the same for children overall, with the exception of children who were 4–6 years of age. This age group was directly affected by County changes to mandatory field screening protocols, which changed mandatory field screenings for referred families with a child under 7 years of age to families with a child under 4 years of age. Unfortunately, for the 4–6 age group there was a slight but statistically significant increase in the likelihood of being re-referred.

4. **The AFST led to reductions in overall case opening disparities between black and white children.** During the Post-AFST period, increases in the identification of higher-risk white children, coupled with slight declines in the rate at which black children were screened-in for investigation, led to reductions in racial disparities. Specifically, there was an increase in the number of white children who had cases opened for services, reducing Pre-AFST case disparities between black and white children.
5. **There was no evidence that the AFST resulted in greater screening consistency within individual call screeners.** Specifically, for the subgroup of 11 call screeners who handled a substantial volume of both Pre-AFST and Post-AFST referrals, attempts were made to assess whether the AFST led to more "within-screener" consistency. Likewise, changes in screening consistency by children's age group and racial group were also assessed. No changes were detected, although it should be noted that there was likely insufficient power to identify anything other than very large shifts.

Will the County continue to fund an independent evaluation?

Yes, Stanford University will continue to follow the outcomes of the AFST in practice, extending the results in time, observing AFST Version 2, and expanding the outcomes reviewed to look at home removals.

AFST VERSION 2

Why were changes made to the original model, operating from August 2016 to November 2018?

DHS was always committed to continuing to improve the model, and we expected to make changes once we had implementation and outcome data. Specifically, the changes were motivated by a number of factors, including:

- Some of the variables (data sources) were unsteady, meaning that they either changed significantly while the model was live and/or they changed from the time period the researchers used to construct the model.
- The re-referral model (which predicted whether a child would be a re-referred within two years) was not as strongly linked to the primary outcome of concern, serious abuse and neglect. Additionally, initial incoming referral rates also represent the most racially disproportionate step of the referral pathway, and so a model predicting future referrals figures to overrepresent black children relative to white. Finally, the nature and characteristics of calls with higher scores using the re-referral model were resonating less strongly with screening staff as cases appropriate for investigation.
- LASSO, the machine learning approach used in the second version, performs better than the logistic regression model used in the original model.

What changes were made to the methodology for AFST Version 2?

Changes were made to the **target outcome**, to the **data sources used in the algorithm**, and to the **policies regarding high-risk and low-risk (and how they are displayed on the visualization)**. Specifically:

- **Target Outcome:** AFST Version 1 (V1) was designed to predict: 1) the likelihood a child would experience abuse or neglect serious enough to be placed in an out-of-home setting within two years of the initial call if the call were screened-in for investigation and 2) the likelihood there would be a re-referral to the hotline within two years if the call were screened-out. Based on feedback from staff and external validation of the model using hospitalization data, we determined that the scores from the re-referral model were not as strongly related to the key outcome of concern, serious abuse and neglect. AFST Version 2 (V2) therefore only predicts the likelihood of out-of-home placement within two years.
- **Data Sources:** In V2, public benefits data were excluded, as were a majority of behavioral health records. Birth records – which Allegheny County began to receive after the building of V1 – were added to the model. Public benefits data were excluded as the current data feeds no longer align to the historic data used to develop V1. Some behavioral health records were eliminated because of temporal variability. In addition, variables regarding the current allegations on the referral were added at the request of call-screening staff.

  Data sources used in V1 of the AFST and continued in V2 include child welfare, jail, and juvenile probation records.

  A complete listing of the variables used in V2 can be found in **Appendix B** of the Methodology V2 report.
- **High-Risk and Low-Risk Policies/Visualization:** The visualization was changed to reflect new high- and low-risk protocols and to provide a visual cue to remind staff that this is a new version (the new visualization can be seen in **Appendix C** of the Methodology V2 report).

In V2, if the maximum referral score is greater than 17 and any child on the referral is younger than 16, the referral is designated to be screened-in for investigation (although supervisory discretion allows an override to screen-out the referral, requiring supporting documentation), and the visualization displays the text "High-Risk Protocol, High Risk and Children Under Age 16 on Referral." If the maximum referral score is under 11 and all children are at least 12, the visualization displays the text "Low-Risk Protocol, Low-Risk and All Children Age 12+ on Referral." The default for referrals identified as low risk is a screen-out unless otherwise deemed necessary; low-risk referrals have to be overridden to be screened in. All other scores are displayed in the visualization and staff has full screening discretion.

What modeling methodology is used in AFST V2?

A number of methodologies were explored for V2, including LASSO, XG-BOOST, Random Forest and SVM logistic regression. To determine which methodology would be used, researchers considered 1) overall performance and accuracy for the high-risk groups; 2) accuracy for black

children versus non-black children; 3) ease of implementation and quality checking; and 4) whether the model showed a positive correlation between the score generated and the probability that the child would be involved in a fatality or near-fatality 50 days or more after the score was generated.

Based on these factors, we chose the LASSO model. A discussion of the performance and external validation of LASSO appears in the Methodology V2 report.

Was the model validated?

An external validation of the model was conducted using Children's Hospital of Pittsburgh data. Encounters were examined (by cause) using four approaches (highest risk score and an injury encounter, randomly selected risk score and an injury encounter, highest risk score before an injury encounter, and randomly selected risk score before an injury encounter). We found a positive correlation between the risk scores and medical encounters for injury, abusive injuries and suicide, showing that the model accurately identifies the children most at risk for relevant hospital events.

Were there accuracy improvements in AFST V2?

There are a number of metrics that can provide information about the accuracy of the algorithm (e.g., area under the receiver operator curve [AUC], outcome plots, mortality regressions, how well the algorithm distinguishes high- and low-risk children, accuracy for black vs. non-black children). Among the methodologies tested, LASSO provided the best balance in increased accuracy, with an overall AUC of 76 percent (74.42% for black children and 77.35% for non-black children) and ability to implement and perform quality assurance checks.

IMPLEMENTATION LESSONS

Do the process and impact evaluations cover everything you've learned since you started building and using the AFST?

No, even the best evaluations can't cover everything. The following FAQs provide additional information and lessons learned during implementation, covering technical, practice and policy reflections.

What are some of the technical lessons learned during AFST implementation?

The technical lessons fall into three categories: efficiency and auditability of variable calculations, unforeseen changes in data availability or content, and complexity of database structures and "real-time" calculations.

1. **Efficiency and auditability of variable calculations**
   The broad array of variables built into the tool, and the initial design for variable calculation and storage made thorough testing and the discovery of all possible calculation defects challenging. It also made the creation of long-term research datasets a burdensome

undertaking; for example, to compile a research modeling dataset with full variables for approximately six years of historic child welfare referrals required almost two weeks of continuous runtime. Initially, the tool called for hundreds (approximately 1000) of distinct variables to be constructed, without sufficient consideration for how they would be utilized by the eventual algorithm(s) producing a score, how they would be used for rebuilding the model (generating a multi-year research data set), and the quality assurance requirements. Tool processing time and design were secondary considerations and created many challenges throughout the initial implementation.

In the interest of enhancing real-time AFST processing speed from the worker perspective, recent efforts have been undertaken to backtrack and decommission obsolete or unused variables from the initial design that are calculated and stored when the tool runs but not actively weighted in any actual algorithms. Being more strategic and selective with initial variable creation in the design stage may have resulted in a leaner and more manageable product structure. A couple of examples of lessons related to efficiency of the model include:

- Many of the variables are repetitive with just slight variations in the data being summarized, however each variable is an independent script in the implementation. In cases where we identify an issue in the calculation, we have to identify all of the variables impacted by the issue and update each one independently. Ideally, there would be more shared/referenced code so that the update would need to be made only once and the changes would be consistent across affected variables.
- The initial design did not capture enough of the intermediate data and calculations. For quality assurance purposes, staff should be able to walk through the logic to get from the source data to the final variable calculations in a clear and transparent way. Investigating potential calculation errors took a significant amount of time due to the way the data model was designed. This process could be much more efficient if the data model was designed with strong consideration for ease of quality assurance.

Based partly on these experiences, major improvements have occurred in how brand-new variables (such as County birth data) were implemented in V2. Performance and processing time were considered at every step in development, validation rules and outlier boundaries were carefully crafted and documented, and variables were only developed if explicitly thought to be important.

2. **Unforeseen changes in data availability or content**

Unlike many analyses conducted, a data mining approach like that used by the AFST doesn't, by definition, draw on clean data sets. In fact, it is more likely the model finds data in less used parts of the system particularly in need of quality assurance. It's probably fair to say that Allegheny County devotes more resources than is typical to monitoring data sources and data quality across its enterprise. That said, the initial quality assurance protocols established were not sufficient to properly monitor the AFST and additional quality assurance had to be developed.

For example, in recent years, structural changes occurred in how behavioral health diagnoses were defined and categorized that did not align with the historic diagnosis variables used in the initial AFST modeling. Because of this, some variables saw significant distribution increases or decreases in incoming prevalence compared to historic data that had been used to determine appropriate weights. In hindsight, the initial tool design should have included some form of automated detection for when incoming data include outliers or appear significantly different than expected, rather than requiring manual detection by analysts monitoring the tool's performance. The team is currently developing an automated quality assurance tool to monitor variable values over time and generate alerts if there are significant changes in a variable that may impact model performance.

3. **Complexity of database structure and "real-time" calculations**
   In developing a tool that aimed to access and utilize real-time, incoming data of varying quality and completeness, constructing variables that were able to properly navigate temporary data staging tables proved to be a new and challenging endeavor. Datasets produced for analyses and for AFST research and modeling inherently had the benefit of full, finalized data entry for a given historic call, without any snapshot mechanism for accurately simulating how complete a data element typically would be in the midst of the early call screening stage. Additionally, in some instances, variables were initially coded to search for data in finalized data tables (where data would eventually be stored later in the process) rather than being directed toward temporary data staging tables where the data would normally exist at the point of call screening. The lesson learned was to spend as much time as possible understanding the exact flow and completeness of data at various processing points.

What are some of the lessons learned (and still evolving) about the use of the AFST in practice?

The most significant, and probably most obvious, lesson is that practice and culture change takes time and that a new tool will have limited immediate impact on culture. As a field, we are slowly evolving from a system that focuses almost exclusively on the allegation of abuse and neglect to one that puts this input in the proper context. In the rebuild of the model, call screening staff requested that the current allegation be included in the model. We had initially decided that this variable should be excluded since the algorithm is best at assessing longer-term risk of abuse and neglect and the call screener alone could assess the current allegation alongside future risk to make a screening determination. We yielded to the requests of the call screening staff to include the variable in the model because it increased their confidence in the score. Nevertheless, our work to change the culture from an allegation-only focus to one with greater understanding of latent risk is just beginning.

Another lesson is that the AFST cannot fix, nor anticipate, other external shocks to the system that might impact practice. This means there must be either very strong communication with frontline managers and/or monitoring of the whole decision-making process. The following

example describes the type of challenge that likely occurs in systems throughout the country and which must be identified and managed if practice is to be consistent.

In late 2017, a combination of factors (staff turnover, staff on medical leave and increased call volume) led to a situation where call screening staff, overwhelmed by the call volume and reduced staff, halted their full business process and began triaging based on referral information. The staff triaged calls they thought serious in one pile, possibly serious in another pile, and likely not serious in a third pile. The problem with making these decisions so early in the process, without the benefit of the full review process or the AFST score, is that cases deemed not serious were later – sometimes a week or two later – determined to be high risk on the AFST. Because we had no monitoring in place to catch this sort of process challenge and because frontline managers did not report the problem, we were lucky to catch the issue at all. Once identified, we considered a variety of solutions and eventually DHS leadership put in place extra supports to allow call screening staff to follow the established protocol, which includes running the score. Today, we have active monitoring and have established tools that allow call screening supervisors to monitor the flow of cases through the decision-making process. The image below displays a week-by-week breakdown of data (12/31/17 through 10/20/18) showing the time that passed between initial referrals and (1) screening decisions and (2) generation of initial AFST scores.

					Time to Screening Decision Approval						Time to Initial ASFT Score					
Week	Start	End	Incoming Referrals Processed	Scr %	Median Days	Average Days	Pct. Over 2 Days	Pct. SOs Over 2 Days	Pct. Over 7 Days	Pct. SOs Over 7 Days	Median Days	Average Days	Pct. Over 2 Days	Pct. SOs Over 2 Days	Pct. Over 7 Days	Pct. SOs Over 7 Days
1	12/31/2017	1/6/2018	249	100%	2.0	3.8	60.6%	85%	16.1%	36%	0.4	1.4	14.5%	20%	4.0%	8%
2	1/7/2018	1/13/2018	312	99%	1.0	3.0	47.4%	74%	11.9%	23%	0.3	0.7	7.4%	14%	0.3%	1%
3	1/14/2018	1/20/2018	292	100%	1.0	3.6	46.9%	66%	15.4%	30%	0.2	1.0	13.7%	19%	0.7%	2%
4	1/21/2018	1/27/2018	356	99%	2.0	5.9	58.1%	87%	18.3%	36%	0.3	2.9	12.4%	24%	1.7%	3%
5	1/28/2018	2/3/2018	349	99%	2.0	4.7	60.5%	92%	18.3%	40%	0.6	2.1	17.5%	36%	2.3%	3%
6	2/4/2018	2/10/2018	307	99%	2.0	3.9	54.4%	87%	13.4%	29%	0.3	1.1	8.1%	17%	2.6%	7%
7	2/11/2018	2/17/2018	331	99%	2.0	6.1	55.3%	78%	26.6%	54%	0.4	3.5	21.8%	31%	12.4%	25%
8	2/18/2018	2/24/2018	308	100%	3.0	6.4	71.4%	94%	32.5%	70%	0.9	3.0	25.0%	39%	13.6%	30%
9	2/25/2018	3/3/2018	337	99%	2.0	4.5	56.7%	87%	19.6%	46%	0.3	1.7	11.3%	22%	6.8%	17%
10	3/4/2018	3/10/2018	377	99%	3.0	5.4	57.3%	81%	29.2%	57%	0.5	2.3	15.6%	29%	13.0%	26%
11	3/11/2018	3/17/2018	344	100%	2.5	5.6	59.3%	93%	28.8%	57%	0.4	2.8	14.8%	29%	12.8%	26%
12	3/18/2018	3/24/2018	323	100%	2.0	3.9	57.3%	78%	21.7%	49%	0.3	1.9	17.3%	37%	12.7%	29%
13	3/25/2018	3/31/2018	286	98%	2.0	4.2	53.5%	82%	18.9%	35%	0.3	1.4	13.3%	23%	3.8%	7%
14	4/1/2018	4/7/2018	330	100%	3.0	4.6	64.5%	84%	27.9%	55%	0.9	2.1	26.4%	37%	10.6%	22%
15	4/8/2018	4/14/2018	307	100%	1.0	5.8	48.9%	84%	24.1%	54%	0.3	3.0	14.3%	27%	9.8%	22%
16	4/15/2018	4/21/2018	349	99%	2.0	5.6	53.9%	83%	27.8%	58%	0.7	3.0	16.9%	32%	11.2%	24%
17	4/22/2018	4/28/2018	385	99%	2.0	7.7	61.3%	83%	29.6%	63%	0.9	4.9	27.0%	42%	15.6%	35%
18	4/29/2018	5/5/2018	367	100%	3.0	5.7	64.9%	89%	31.6%	65%	0.7	2.1	26.2%	37%	14.2%	30%
19	5/6/2018	5/12/2018	348	100%	3.0	1.5	71.8%	88%	30.5%	63%	1.1	4.3	28.7%	43%	15.2%	33%
20	5/13/2018	5/19/2018	334	100%	3.0	3.4	76.6%	96%	32.3%	65%	0.9	4.2	30.8%	48%	14.4%	30%
21	5/20/2018	5/26/2018	369	100%	2.0	1.4	65.3%	87%	25.2%	47%	0.9	-1.9	25.7%	36%	10.6%	20%
22	5/27/2018	6/2/2018	293	100%	2.0	-3.2	64.8%	84%	10.9%	23%	0.4	1.5	22.2%	32%	1.7%	4%
23	6/3/2018	6/9/2018	304	99%	2.0	3.4	51.0%	79%	14.8%	32%	0.4	1.0	11.8%	19%	1.6%	3%
24	6/10/2018	6/16/2018	234	100%	1.0	2.7	40.6%	64%	11.1%	25%	0.2	0.5	6.8%	9%	0.4%	1%
25	6/17/2018	6/23/2018	266	99%	1.0	2.8	46.6%	71%	10.9%	24%	0.2	0.7	9.8%	9%	0.4%	1%
26	6/24/2018	6/30/2018	266	100%	1.0	2.1	39.5%	59%	8.3%	21%	0.2	0.4	8.6%	9%	1.1%	2%
27	7/1/2018	7/7/2018	203	100%	2.0	3.3	62.6%	87%	11.8%	23%	0.2	0.8	13.3%	14%	0.5%	1%
28	7/8/2018	7/14/2018	270	99%	1.0	3.4	43.0%	63%	14.8%	30%	0.2	0.9	10.7%	12%	1.9%	2%
29	7/15/2018	7/21/2018	282	100%	2.0	3.6	59.6%	79%	13.8%	26%	0.3	1.1	17.0%	19%	0.7%	1%
30	7/22/2018	7/28/2018	285	99%	2.0	3.2	53.3%	79%	11.9%	23%	0.5	1.0	9.8%	14%	1.4%	3%
31	7/29/2018	8/4/2018	271	100%	1.0	1.8	34.7%	50%	10.3%	20%	0.2	-0.1	5.5%	5%	1.1%	2%
32	8/5/2018	8/11/2018	256	99%	1.0	1.1	45.7%	68%	16.8%	36%	0.3	0.8	5.9%	9%	1.6%	3%
33	8/12/2018	8/18/2018	264	99%	1.0	0.8	44.7%	58%	13.3%	20%	0.2	0.7	7.6%	8%	0.8%	1%
34	8/19/2018	8/25/2018	226	100%	1.0	2.7	41.2%	57%	10.6%	22%	0.2	0.6	9.7%	12%	0.0%	0%
35	8/26/2018	9/1/2018	339	100%	1.0	3.3	48.1%	64%	13.0%	26%	0.8	1.3	15.0%	18%	1.5%	3%
36	9/2/2018	9/8/2018	310	98%	3.0	4.4	65.8%	90%	18.4%	33%	0.9	1.9	28.7%	44%	1.0%	1%
37	9/9/2018	9/15/2018	304	100%	2.5	4.3	62.5%	91%	16.1%	34%	0.8	1.6	25.0%	34%	1.0%	2%
38	9/16/2018	9/22/2018	368	99%	2.0	3.7	62.8%	75%	16.0%	34%	0.9	1.6	21.7%	22%	1.1%	1%
39	9/23/2018	9/29/2018	344	95%	2.0	3.5	48.5%	78%	12.8%	30%	0.8	2.2	18.6%	29%	5.8%	9%
40	9/30/2018	10/6/2018	391	95%	2.0	3.2	56.0%	79%	14.8%	31%	1.0	2.0	22.8%	34%	4.1%	6%
41	10/7/2018	10/13/2018	311	84%	2.0	3.3	57.9%	90%	12.2%	40%	1.0	2.3	24.7%	47%	8.4%	18%
42	10/14/2018	10/20/2018	252	73%	1.0	1.3	31.4%	62%	6.0%	0%	0.8	1.1	17.9%	9%	0.4%	0%

What are some reflections around the policies associated with the AFST?

Two policy reflections jump to the fore: (1) whether Allegheny County made the right decision to limit the score to call screeners/supervisors and whether this is still the right decision and (2) whether high- and low-risk protocols are sufficient.

(1) Allegheny County leadership took a conservative approach to the use of the AFST, determining that the score was only to be used by call screeners and call screening supervisors, with no exceptions. We've been successful in applying this approach and think it was the right decision. However, now that we are more than two years into the process, we see improvements in call screening decision-making, but the established process still leaves far too many high-risk cases that are either not accepted for services or not triaged properly once accepted for services. In recognition of this reality, beginning in spring of 2019, DHS will explore how the score might be used elsewhere in the child welfare process. We will do this work, as we have in the past, thoughtfully and with engagement with experts and community leaders. As part of this exploration we will consider:

- whether the AFST should be provided to the clinical manager overseeing investigations to help him/her determine the response time and staffing.
- whether to use the score as one additional way to identify cases that undergo our quality assurance reviews (compliance and/or quality reviews).
- whether the score should be available to investigative supervisors to help them ensure due diligence on high-risk cases.
- whether the score could/should replace the state required risk assessment.

Any additional use of the AFST must be weighed carefully to assess the value of its ability to help us protect children and support families versus the risk of providing undue weight to one approach or reinforcing our own system behavior. As in the past, we will have to consider the way in which the system currently makes these determinations and whether the AFST can help improve that process (and the outcomes), acknowledging that such a model will never be perfect.

(2) High- and low-risk protocols: Because of concern that the score would have too much power in decision-making, we implemented "nudges," which defaulted the highest-risk cases to be screened in and required supervisors to explicitly override the decision with written justification if they felt it should not be investigated; a similar default-based nudge with override capability was later added to the lowest-risk cases. These nudges have led to only minimal additional concurrence with the model. We are looking at whether we should take a stronger approach to achieve more concurrence on very high-risk and very low-risk cases (acknowledging that the low-risk protocol has only been in place since November 2018). One particular reason that the high-risk protocol is only followed in about sixty-one percent of GPS cases is because many of these children are older and have allegation reasons that do not feel like abuse/neglect to the call screening staff. The model views

them as high-risk because of the considerable child protective services history and history of other supports, and the validation (using hospital data) confirms that these children face elevated risk of serious injuries, including self-inflicted injuries. Given this information, we should consider how DHS, as an integrated human services department, can divert these youth from the child welfare system (within the child protective services law) into a set of supports better aligned to meet their ongoing service needs. This is an ongoing challenge that requires additional work.

Did these technical, practice and other challenges impact the results of the evaluation?

It's not clear if these challenges impacted the results of the evaluation, but it's possible the results would be more robust and attenuate less absent these challenges. That's why we'll continue to improve quality assurance, monitor our work and continue independent evaluation.

Has the policy landscape around the implementation of predictive risk modeling changed since DHS began this work?

Yes, all of the fields surrounding this issue are in rapid evolution. When we started this work, there was no handbook on how to develop algorithms in the public interest and today there are numerous checklists, guidebooks and research groups established to help governments deploy predictive analytics in human services. The machine learning field is also rapidly evolving as are the official definitions of algorithmic fairness and discrimination in modeling. Allegheny County has attempted to both keep pace with these evolutions and to continue to improve our work based on these advancements. It is likely that we'll look back on the earliest models and see them for their flaws, but it is better to judge them on their improvement over previous practice and for our ability and willingness to continuously examine and improve.

You have reported outcomes for the first year of implementation —can you provide results for the full period under AFST Version 1?

Yes, for the period December 1, 2016–November 29, 2018 (and observed through 3/8/2019 data entry):

	PERCENT SCREENED-IN FOR INVESTIGATION	PERCENT OF THOSE SCREENED IN FOR INVESTIGATION THAT WERE ACCEPTED FOR SERVICE
Mandatory*	61%	45%
High	47%	41%
Medium	42%	37%
Low	31%	35%
No Score	23%	29%
Total	41.4%	39.1%
Pre AFST Comparison**	45.5%	34.3%

** Note that "mandatory" screen-ins may still be screened out at the discretion of the call screener and call screener supervisor.*

***December 1, 2014– November 29, 2015, selected as a comparison period because it is the most recent full calendar year (pre-AFST) with the same seasonal distribution of the observed AFST period above.*

See **Appendix A** for more detailed data.

APPENDIX A: AFST VERSION 1 SCREENING SCORE DATA, 12/1/16 TO 11/29/18

Dec. 1, 2016 - Nov. 29, 2018 GPS Screening Score Statistics

	Family Screening Score		GPS Referrals	Screening Outcomes											Investigation Outcomes			
			Count	Total Scrn-In	Total Scrn-Out	Total w/ Decision	Pct. Scrn-In	Tier Pct.	Field-Screen Unit Assigned	Pct. FS	Tier Pct.	FS Scrn-In	FS Scrn-Out	*Referrals on Active Family*	Investigated to Date*	Accepted for Service*	Pct. Accepted	Tier Pct.
	Mandatory	14%	3457	1402	907	2309	61%	61%	209	6%	6%	30	158	*1102*	1283	580	45%	45%
High Scores	20**		1488	471	561	1032	46%		118	8%		16	97	*440*	461	186	40%	
	19**		1093	371	473	844	44%		109	10%		14	85	*236*	364	142	39%	
	18**	36%	1112	384	525	909	42%	47%	122	11%	10%	16	103	*195*	377	145	38%	41%
	17		1728	722	755	1477	49%		188	11%		16	167	*241*	702	297	42%	
	16		1668	702	764	1466	48%		188	11%		24	152	*191*	684	295	43%	
	15		1551	700	726	1426	49%		181	12%		16	157	*115*	686	291	42%	
Med. Scores	14		1373	592	672	1264	47%		165	12%		11	147	*99*	585	222	38%	
	13		1147	453	611	1064	43%		137	12%		12	115	*69*	449	186	41%	
	12	22%	1133	425	631	1056	40%	42%	147	13%	13%	14	126	*68*	421	137	33%	37%
	11		913	356	499	855	42%		125	14%		9	109	*49*	351	118	34%	
	10		826	281	499	780	36%		114	14%		6	103	*40*	280	103	37%	
Low Scores	9		821	291	494	785	37%		106	13%		7	92	*30*	289	113	39%	
	8		690	200	464	664	30%		89	13%		3	79	*18*	197	74	38%	
	7		556	190	350	540	35%		68	12%		3	63	*14*	190	70	37%	
	6		634	193	424	617	31%		72	11%		5	64	*15*	193	69	36%	
	5	18%	540	127	402	529	24%	31%	59	11%	11%	2	54	*8*	127	37	29%	35%
	4		372	111	254	365	30%		44	12%		5	38	*5*	111	27	24%	
	3		330	89	235	324	27%		24	7%		3	20	*5*	88	29	33%	
	2		164	39	122	161	24%		11	7%		0	10	*2*	37	16	43%	
	1		124	37	87	124	30%		7	6%		0	7	*0*	37	8	22%	
	No Score	10%	2410	527	1790	2317	23%	23%	218	9%	9%	11	198	*82*	523	152	29%	29%
	Total		**24130**	**8663**	**12245**	**20908**	**41.4%**		**2501**	**10.4%**		**223**	**2144**	*3024*	**8435**	**3297**	**39.1%**	
											FS Scr. Rate:	*9.4%*	*90.6%*					

*** Scores of 18, 19, and 20 in this analysis must be driven by the re-referral model, since placement scores of 18+ would instead appear under the "Mandatory" row; this may explain differences in prevalence and screening rates*

**Investigations that end in attaching to an open case are omitted from counts; only decisions on potential "new cases" counted. As a result, "Investigated to Date" will be slightly smaller than the number screened-in for investigation*

Notes: Referral file as of 3/8/2019. Produced by ACDHS-DARE. At point of extract, a few dozen referrals from prior date ranges were stored in database "holding tables" and may not be included. This can indicate data entry lag for presumed screen-outs, but it can also include broken/aborted/accidental referrals that should be omitted anyway.
A data fix implemented on November 29, 2016 changed the tools in ways that increased the prevalence of higher (~18-20) scores; these figures span both before and after the change.
Please not that "Mandatory" scores are defined as referrals that score an 18-20 on the placement model, and thus the rows 18, 19, and 20 above are driven by the "re-referral" model (and cannot have had an 18+ on the placement model).

Dec. 1, 2014 - Nov. 29, 2015 (Pre-AFST) GPS Comparison Sample Period

	GPS Referrals	Screening Outcomes					Investigation Outcomes		
	Count	Total Scrn-In	Total Scrn-Out	Total w/ Decision	Pct. Scrn-In	*Referrals on Active Family*	*Investigated to Date**	*Accepted for Service**	*Accept %*
Total	**9742**	**4148**	**4963**	**9111**	**45.5%**	*627*	**3903**	**1340**	**34.3%**
Avg. Yr Change	2323	183.5	1159.5	1343	-4.1%	885	314.5	308.5	4.8%
Change (%)	23.8%	4.4%	23.4%	14.7%	--	141.1%	8.1%	23.0%	--