Allegheny Family Screening Tool

FREQUENTLY-ASKED QUESTIONS

UPDATED AUGUST 2018

CONTENTS

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. For more information about the AFST, see http://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx.

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). The process evaluation, "Allegheny County Predictive Risk Modeling Tool Implementation: Process Evaluation", is complete and available here. The impact evaluation of the original model is expected by the end of 2018.

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. The following Frequently-Asked Questions are presented as a quick reference for those interested in highlights from this publication as well as the process evaluation 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. These factors are 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. 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.

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.[2]

2 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.[3] Hornby Zeller Associates was selected to conduct an process evaluation and Stanford University was selected to conduct an impact evaluation. The process evaluation is complete and available here. The impact evaluation, expected by the end of 2018, is focusing on whether the AFST increases the accuracy of decisions, reduces unwarranted variation in decision-making and reduces disparities, and will also examine overall referral rates and workload.

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

THE MODEL

What was the total cost of developing the AFST?

The total cost was $1,035,424, as detailed below:

VENDOR	SERVICE	TOTAL
Auckland University of Technology	Methodology and Model Design	$500,000
Deloitte	Technology	$280,000
Stanford University	Impact Evaluation	$160,000
Hornby Zeller Associates	Process Evaluation	$95,424
TOTAL		**$1,035,424**

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.

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.

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))[4]. In our next implementation, expected by the end of 2018, the research team intends to have a better performing model — with an AUC of 80 percent.

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.

[4] 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."

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 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.

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 will add functionality to the AFST that will allow workers to report when a score seems wrong/surprising to them (expected by July 2018). All reported referrals will be reviewed anonymously by the research team.

An independent impact evaluation (which will assess the effectiveness of the AFST compared to the existing approach) and process evaluation (which will evaluate how the AFST is implemented) will alert the County and research team to any concerns about the effectiveness and operation of the AFST. The process evaluation is complete and can be viewed here. The first phase impact evaluation is expected by the end of 2018.

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. The County intends to build a "Version 2" of the AFST that will include improvements identified by process and impact evaluations. Version 2 is expected to be implemented in the fall of 2018.

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 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?

There are currently 34 call screeners and call screening supervisors.

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,[5] 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 supervisors are required to generate the AFST score prior to finalizing a screening decision.

[5] The AFST is intended to assist in decision-making for CPS referrals; any allegation meeting CPS criteria is immediately investigated (state-mandate).

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.

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."[6] 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.

[6] The term "mandatory screen-in" is enclosed in quotations to reflect the fact that call-screening supervisors may override the score.

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.

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 will assess 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 ongoing impact evaluation.

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 ongoing impact evaluation. Results in these areas should be available by the end of 2018.

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, in the first year of operation, more than 30 percent 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 the tool's first full year in use (8/2016–8/2017), screen-in rates follow:

SCORE CATEGORY	PERCENT SCREENED-IN FOR INVESTIGATION
Mandatory	63%
High	49%
Medium	43%
Low	30%
No Score	25%
Total	42.2%
Prior Year	41.7%
Change	+0.5%

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 has risen from about 36% in the year prior to the AFST to about 41% in the first year of the tool's use. 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?

During the first year of the AFST (8/2016–8/2017), 8.4% of CPS referrals screened in for investigation led to at least one child being removed in the following 90 days.

PROCESS EVALUATION

(HZA process evaluation report available here)

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 Data Analysis, Research and Evaluation), 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 pages 3–4 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% 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?

A.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?

A.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 17 of the HZA evaluation report for more detail about the recommendations.