DEVELOPING PREDICTIVE RISK MODELS to Support Child Maltreatment Hotline Screening Decisions

MARCH 2017

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. 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). The process evaluation has been completed and the impact evaluation is expected by the end of 2018.

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 three reports:

1) *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
2) *Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County*, by Tim Dare and Eileen Gambrill
3) *Response to Ethical Analysis* by the Allegheny County Department of Human Services

The process evaluation is available here. Once the impact evaluation is completed, it will also be made available.

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, eventually, 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

Contents

List of Figures

List of Tables

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.

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.

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.

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.

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

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.

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*.

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.

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.

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.

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/

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

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).

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

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/.

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.

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).

Figure 1: Physical Activity Injury

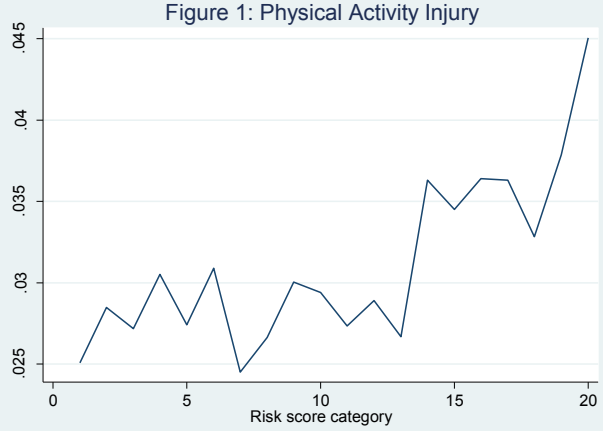


Figure 2: Transportation Injury

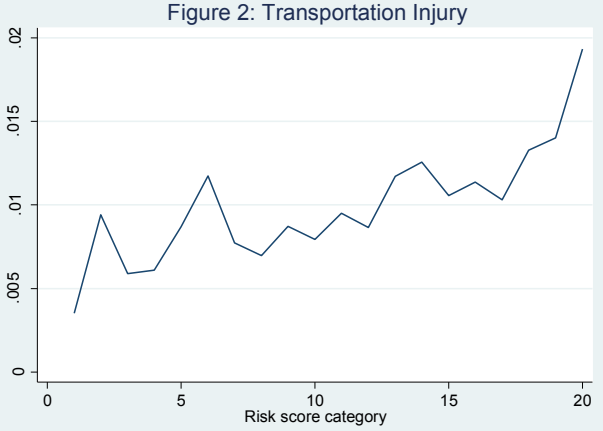


Figure 3: Accidental Fall

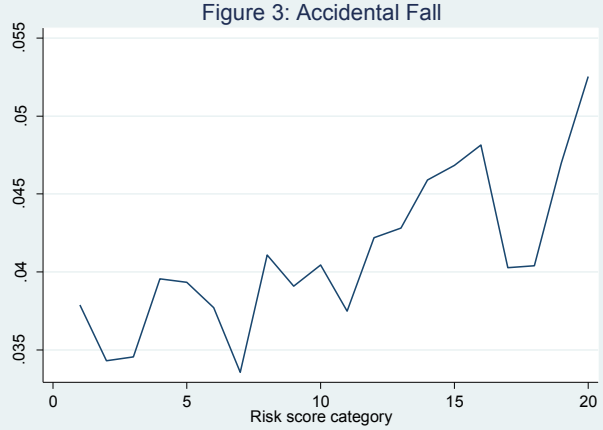


Figure 4: Struck Object or Person

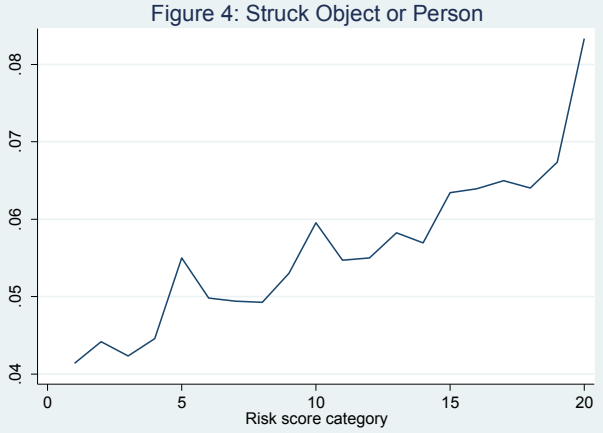
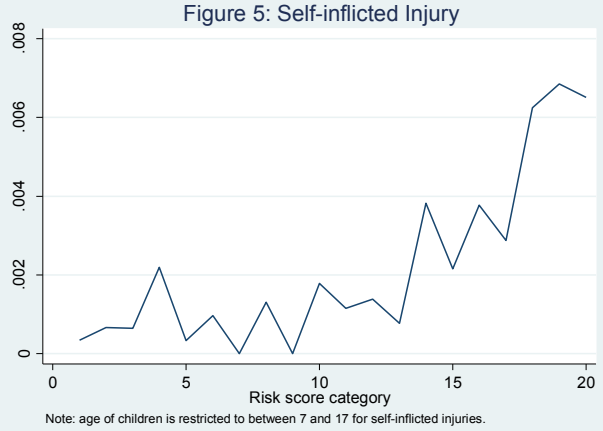
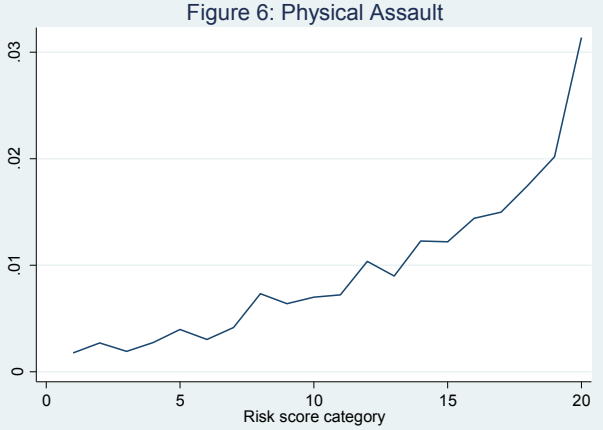


Figure 5: Self-inflicted Injury



Note: age of children is restricted to between 7 and 17 for self-inflicted injuries.

Figure 6: Physical Assault



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.

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.

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

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,

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.

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

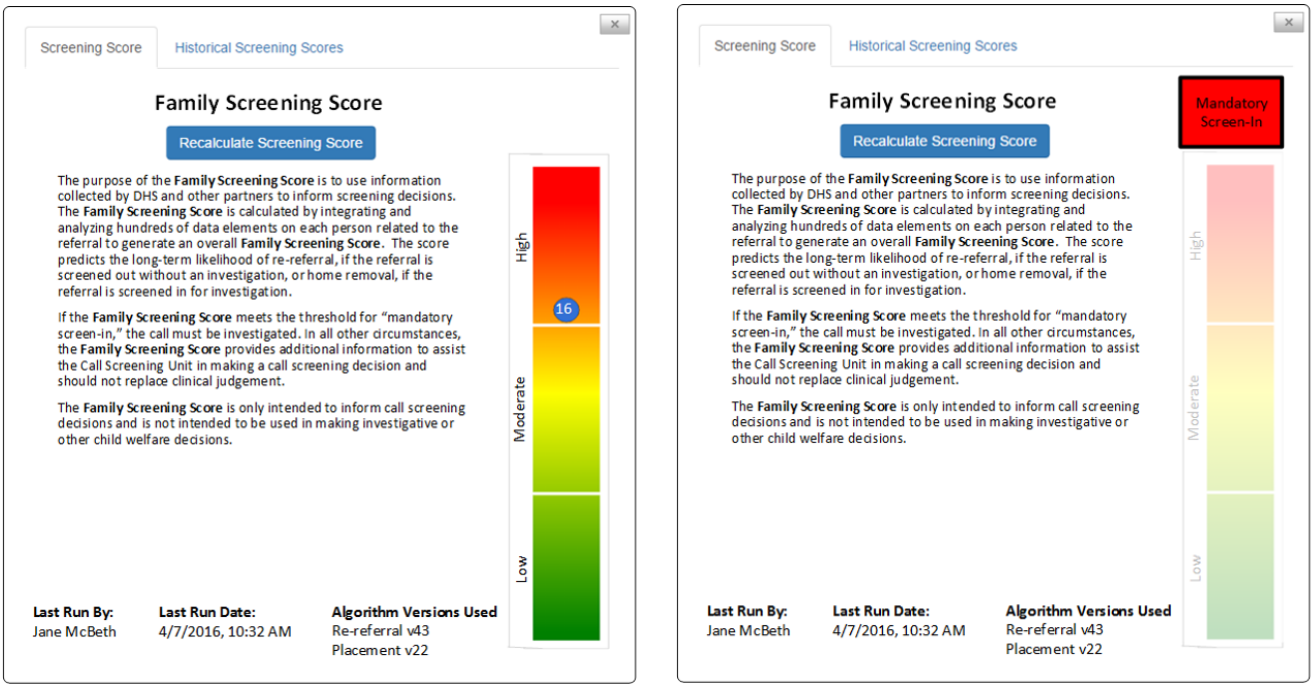
After considerable discussion, the research team and Allegheny County decided that results from this initial modeling effort were promising enough to progress to the implementation stage.

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

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

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

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).

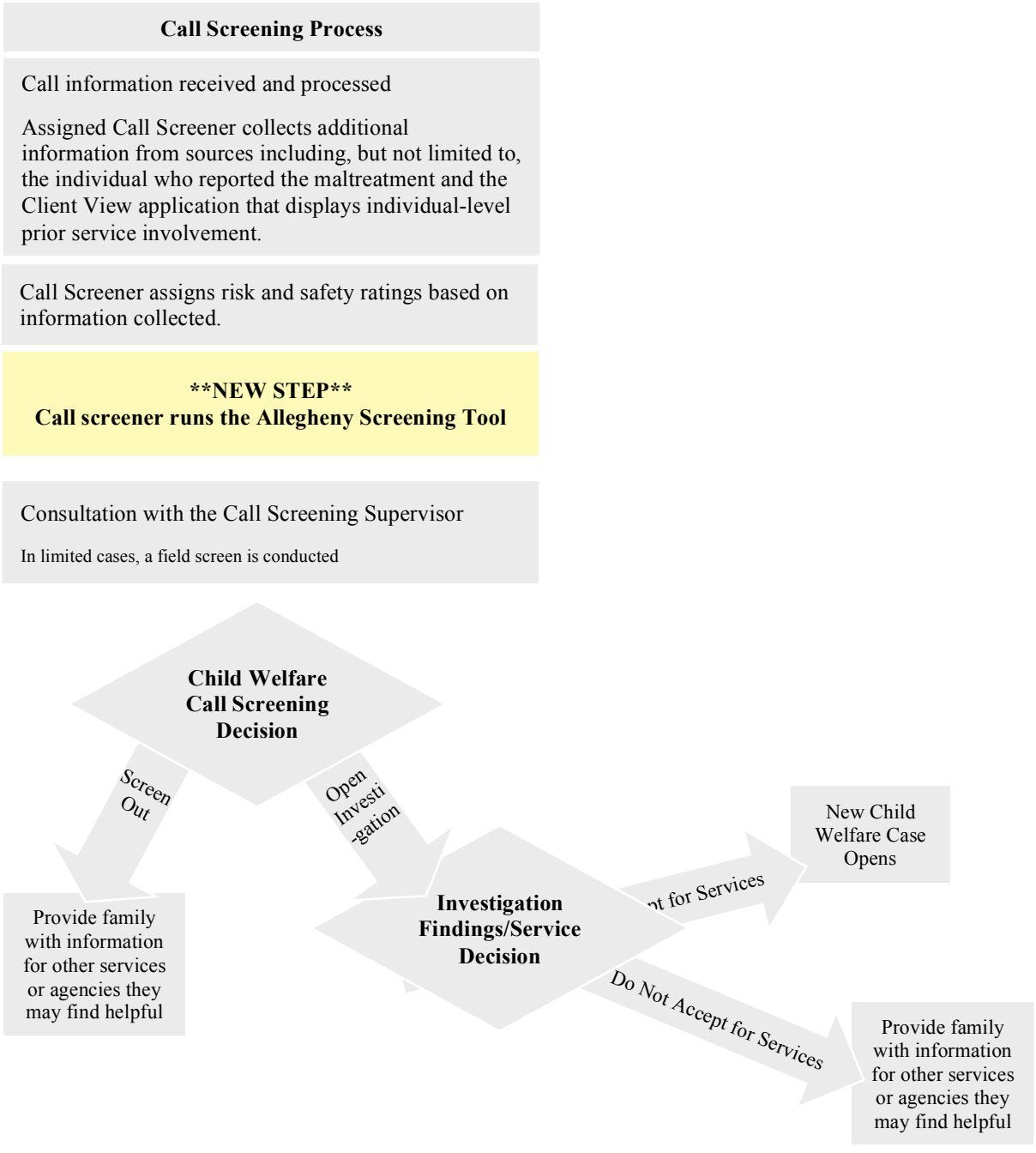
	Placed in 730 days	Referred in 730	Service Open in 730	Currently Screened In	Race Black
Low	0.016	0.201	0.282	0.24	0.150
Medium	0.046	0.310	0.432	0.38	0.334
High	0.088	0.407	0.552	0.49	0.401
Auto	0.226	0.333	0.474	0.74	0.563
Total	0.097	0.320	0.444	0.47	0.371
Ratio of Low to Auto-Screen In	14.05	1.66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

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.

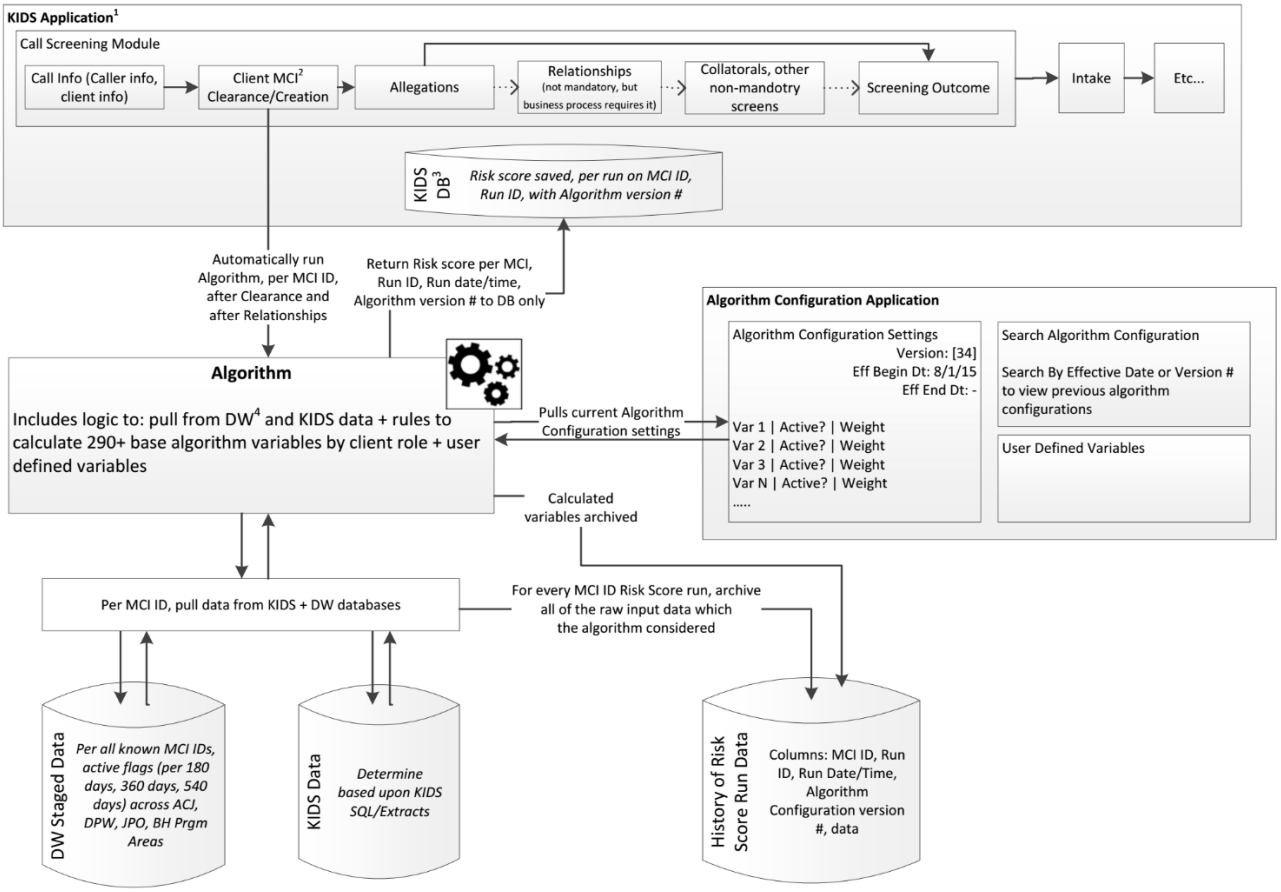
Figure 9: Referral Progression Process



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

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.

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.

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.

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.

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 $3 \leq age < 6$
chld_age_sc1_null	The number of other children involved in this referral who are $6 \leq age < 9$
chld_age_sc2_null	The number of other children involved in this referral who are $9 \leq age < 13$
chld_age_teen_null	The number of other children involved in this referral who are $13 \leq age < 18$
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.

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

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

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

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

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

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)

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

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.

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'.

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.