CENTRE FOR
SOCIAL DATA ANALYTICS

Implementing a Child Welfare Decision Aide in Douglas County

Methodology Report

December 2019

AUT
UNIVERSITY
NEW ZEALAND

The research team has ethics approval from the Ethics Committee at Auckland University of Technology (AUTEC) for implementation of the Douglas County Decision Aide (AUTEC IRB # 17/84).

This report discusses Version I of the Douglas County Decision Aide, which was live as of January 2020. Please contact the authors for up to date information about the DCDA.

Authors:

The Douglas County Decision Aide was led by Rhema Vaithianathan with the contribution of the following staff at the Centre for Social Data Analytics and contractors including Haley Dinh, Allon Kalisher, Chamari Kithulgoda, Emily Kulick, Megh Mayur , Athena Ning and Diana Benavides Prado. Emily Putnam-Hornstein provided advice. Contact: rhema.vaithianathan@aut.ac.nz

Acknowledgments:

The research team would like to acknowledge the leadership of Daniel Makelky, Ruby Richards and Nicole Adams at Douglas County. This project was supported by Casey Family Programs, Colorado Department of Human Services and the Zoma Foundation.

Table of Contents

Executive Summary

In early 2017, Douglas County Department of Human Services commissioned the Centre for Social Data Analytics to explore the use of predictive risk modelling to help to inform, train and improve the triaging of child welfare referrals by County staff.

As part of a feasibility study, the research team generated screening scores for historic referrals that had already been triaged. They found that a large proportion of children with the highest risk scores were being screened out and a large proportion of those with the lowest risk scores were being screened in.

The team proceeded to develop and implement the Douglas County Decision Aide (DCDA), a tool which produces a risk score from 1 to 20, indicating the likelihood that a child involved in a referral will be subject to a removal within two years of the referral. This tool is embedded in the existing RED Team process used by caseworkers and supervisors at Douglas County.

An experimental phase explored four target outcomes, of which *removal within 24 months* was chosen as the focus for the DCDA. Three methods for generating algorithms were tested to determine which would best fit the requirements. The DCDA implemented as of September 2019 was built with LASSO regularized Logistic Regression.

An analyses of disparities of the model showed that the tool is well calibrated across the different sub-groups of race, age and gender. The tool was validated against a related, but independent set of data about the same population, including objective outcomes at least 60 days post-referral. We used case-data about children who were referred for one or more of the following critical incidents: severe, severe-egregious, near fatal, or fatal. The research team has implemented various quality assurance measures to ensure the ongoing performance of the model is monitored.

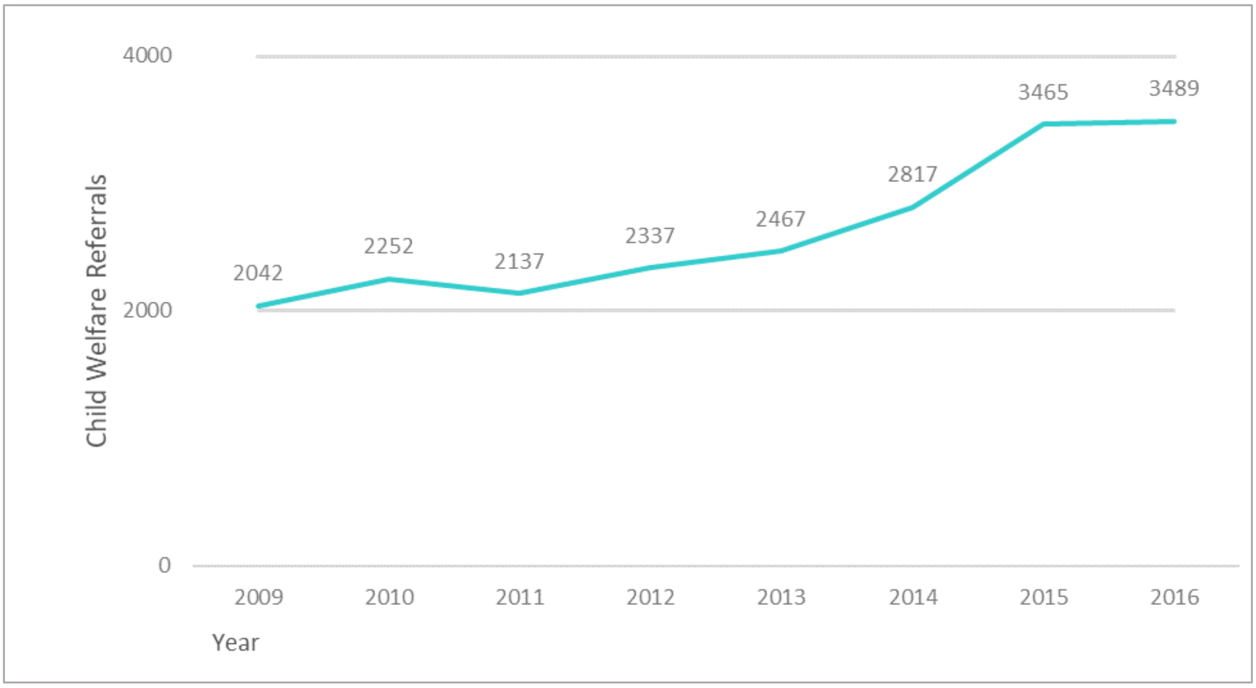
Recommendations made by ethicists in a similar implementation in Pennsylvania were referred to in the design and implementation of the DCDA. The use of the DCDA by County staff and its impact on decision-making are being independently evaluated through a Randomized Control Trial.

Background

In early 2017, Douglas County Department of Human Services (herein referred to as Douglas County) commissioned the Centre for Social Data Analytics (CSDA) to explore the use of predictive risk modelling for its call screening decisions. It was felt that with the investment that the County had made in HSCARES – a system integrating the public welfare system (CBMS) with the State's child welfare system (TRAILS) – a growing population, and consequently rising referrals, there was an opportunity to use the integrated data to help the County triage referrals more effectively. Douglas County was aware of CSDA's work in Allegheny County, Pennsylvania, and like Allegheny their objective was not to have an algorithm to replace human decision-making. Rather, such a tool was to be a decision-aide to help to inform, train and improve the triaging done by child protection staff.

Douglas County is a rapidly growing outer suburb of Denver. Its population in 2018 was 346,000 – a 2.9% increase from the previous year.[i] The County's population is very homogenous in its composition. The population is relatively wealthy with a median household income of $109,292 (ranked 7th nationally for income) and its poverty rate is correspondingly low at 3.7% - the lowest poverty rate in Colorado. As the County's population has increased, it has seen a steady rise in child welfare referrals (see Figure 1) and subsequent investigations.

Figure 1: Annual counts of child welfare referrals to Douglas County Department of Human Services



The initial research was a feasibility exercise to see if the data available within the data systems at the time of a call (Child Welfare data and Public Welfare eligibility information) was sufficient to train a model accurately enough to identify children at the highest risk of out of home placement.

In August 2017, the research team completed this first stage of the project, advising the County that it was possible to use the available data to build a tool with sufficient predictive power. As part of the feasibility study, the researchers generated screening scores for historic referrals that had already been triaged. They found that a large proportion of children with the highest risk scores were being screened out and a large proportion of those with the lowest risk scores were being screened in.

The subsequent phases of the project included development of a prototype model and implementation of the Douglas County Decision Aide (herein referred to as the DCDA). The tool was deployed in early 2019 and a randomized control field trial commenced.

This report proceeds as follows. In the first section we outline how Douglas County was triaging calls prior to the deployment of the DCDA and how the DCDA has changed that process. The second section deals with the details of how the DCDA tool was built and validated. The third section reviews ethical concerns regarding the tool.

Section 1: Developing the Douglas County Decision Aide

Introduction

The purpose of this section is to describe how screening decisions were made in Douglas County prior to the introduction of the tool, so that the reader can contextualize the rationale and expected effect of the DCDA.

From the start, the County was determined that the tool would only be an additional piece of information to be used by staff to support their screening decisions. Moreover, the implementation would be conservative and embedded into the existing processes as much as possible. As a result, the research team and County put considerable thought into how the tool could be deployed with minimal disruption to current practice, including at which point the scores could be revealed, and to whom. This section outlines the thought process and the conclusions reached. It is intended to give the reader an understanding of what decisions had to be made and how the County and research team set about making those decisions.

Existing approach to screening referrals

The screening decision that Douglas County staff have to make upon receiving a referral alleging abuse or neglect is whether to investigate the allegation ("screen-in"), or not ("screen-out") - in the latter case no further action is taken. Historically, 30% to 40% of all referrals were screened out. Referrals are typically received by the County via the courts (2%); telephone calls (79%) or email; walk-in or letter (combined 19%). About half of all referrals are received from law enforcement and schools: 24% and 23% respectively. Parents and family members make up 10% of referrals. Less than 5% originate from health providers (hospitals, nurses and physicians).

Prior to the introduction of the DCDA, Douglas County had two approaches to making the screening decision. For the majority of referrals (roughly 85%)[1], they used a RED ("Read, Evaluate, Direct") Team process. RED Team is a consensus based decision-making process where a group of workers (including at least one child protection supervisor and two caseworkers) meet to consider each referral. The discussion involves a set of semi-structured questions about the risks and safety present in the family and situation, using information taken at the time of the referral. During the RED Team process, workers have access to and are supposed to explore previous child welfare interactions using the TRAILS data and Colorado courts records. The purpose of this process is to ensure that the team is forced to consider each of the multiple domains of risk and safety with sufficient care.

The composition of the RED Team changes daily, with members drawn from approximately 30 front line staff and 10 supervisors and managers. There are typically two or three RED Teams in operation

[1] Based on data from April 15, 2016 – April 14, 2017

at Douglas County each morning, working through the referrals that were referred in the previous day. The RED Team decision - whether to screen in or out - is approved by the supervisor of that team. If, however, the team decides to screen out a child under six years of age, the decision requires a secondary review by another supervisor.

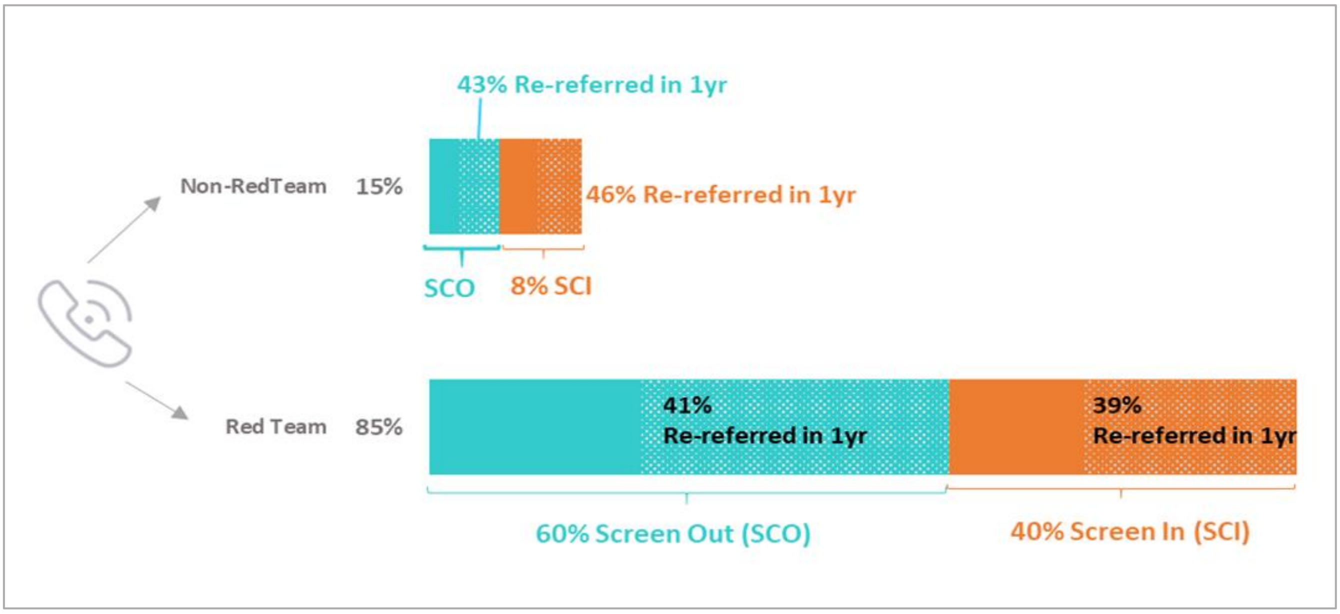
The RED Team process is a requirement of Colorado Human Services Rules and Regulations. Referrals that were not subject to the RED Team (roughly 15%) were either: [2]

1) screened out because they did not meet the conditions required for assessment; or
2) a screening decision was made by a supervisor because the referral concerned allegations about Youth-in-Crisis; or
3) an immediate response was required.

An immediate response is usually required when there is an immediate threat, an allegation of a serious injury or the alleged victim is under 5 years of age.

Figure 2 shows the distribution of referrals screened in and out for those handled by RED Teams and those that were not, and the re-referral rates over a 16-month period between January 2016 and April 2017.

Figure 2: Referral Distribution and Re-referral Rates (Jan 2016 – April 2017)



[2] Based on data from April 15, 2016 – April 14, 2017

Incorporating the DCDA into the decision-pathway

In February 2019, Douglas County incorporated the DCDA score into the call screening process. *Section 2: Building the Predictive Risk Model*, provides greater detail of the DCDA score and how it is generated. In brief, the score is a number from 1 to 20 indicating the likelihood that the child involved in the referral will be subject to a removal within two (2) years of the referral.

Because the DCDA relies on the State's SACWIS (Statewide Automated Child Welfare Information System), which only matches referrals with the State system in a batch process on a nightly basis, the tool can only be run after this matching process is completed. For this reason, the score can only be generated one day after the call is received. This means that the score can only be seen by the RED Team or by a supervisor handling a Youth-in-Crisis referral on the following day, and therefore cannot be used to inform urgent responses.

To determine how to get the most value from the DCDA at RED Team meetings, the research team conducted case simulations with Douglas County staff, involving caseworkers, supervisors and leadership. This was done about three months prior to deployment. The researchers set up a mock RED Team process, where the staff had to work through recent Douglas County calls that had been retrospectively scored using the DCDA. Some of the referrals had screening decisions that were in sync with the risk scores (i.e. a low score was screened out, or a high score was screened in) and others did not.

The team worked through the cases as if it was a real RED Team process. They made "decisions" before receiving the score, and then discussed whether to change their decision. They were then shown the action taken by the RED Team and what had transpired in the case during the months following the referral. As expected, decisions made at simulation were not always identical to the decisions made by the real RED Team. The subsequent trajectory of the case was also instructive to understanding whether the tool was providing some further insights that had not been gleaned by staff.

Throughout the simulations, and in a debriefing session later, staff members reflected on the impact of having risk scores during the RED Team discussion and considered how the scores impacted their thinking and screening decisions. There was a strong consensus that having risk scores helped the RED Team to look beyond the specifics of the allegations and circumstances indicated in the report, to consider history and to think more critically. Several times, RED Team members identified how their thinking would have been different without the score.

The County had to determine whether to reveal the DCDA score *before* the RED Team discussed the case, or at the end prior to recording the screening decision. Getting the scores up front could save time and reduce confirmation bias, whereas revealing them *after* RED Team deliberation might help to promote critical thinking. A concern expressed was that if the RED Team saw a high score up front, they would just choose to screen it in without any further thought. To better weigh these considerations and come to a definitive conclusion, a four-week trial period was planned prior to formal implementation of the DCDA and the start of the RCT (Randomized Controlled Trial). During this trial, a risk score was provided for all reports and each week RED Team members were given specific instructions on how and when to introduce the risk scores.

In weeks 1-2 , the risk score was revealed to RED Team up front, prior to discussing the allegation(s) and case history. A focus group was held at the end of week one to obtain feedback on what it was like to know this information prior to deciding whether to screen in or out, and to gather feedback on the visualization provided by the risk score.

In week 2-4, the RED Team was instructed not to obtain the risk score until after deliberating over the allegations and case history and making a preliminary screening decision. At the end of the four-week trial it was decided that the RED Team would conduct its processes as usual until they reached a screening decision, at which point the score would be revealed to them. For Youth-in-Crisis, referrals go directly to supervisors and they would receive the score.

A point of considerable debate was whether, in addition to the RED Team, investigators (i.e. those caseworkers who are tasked with establishing whether any abuse or neglect occurred once the call is screened in) would be provided with risk scores. A reason for doing so was that in some cases where the risk score prompted the RED Team to screen in, investigators handling that case might be confused about why they are being asked to investigate (e.g. if the immediate allegation seems innocuous). If the objective of screening in was to take the opportunity to provide services, then this opportunity is lost if investigators do not address prevention. On the other hand, allowing investigators access to the score could interfere with their clinical judgement. It was concluded that – in order to be conservative – investigators would not be given the score at this stage, and Douglas County would try to avoid assigning investigations to a caseworker who also sat on the RED Team for that referral.

Training of staff

A three-hour training session was provided to all full-time and occasional call screening staff, intake administrators and key child welfare administrators prior to implementation of the DCDA. The training provided a brief overview of the DCDA and the application of it within Douglas County, to give participants an understanding of what risk modelling is, how the model was built, and the predictive power of the model. The training also covered how participants would use the tool to obtain risk scores.

Much of the training was dedicated to building worker understanding of the policy and practice associated with using the tool. Some of the key points emphasized in these discussions included:

- Scores would only be accessed by RED Teams and would not be shared with investigating caseworkers;

- The screening tool would be used as one of the tools available to RED Teams when making their recommendations, and supervisors when making their decisions;
- The tool would not mandate the response the County will have to any referral (i.e. low scores can still be screened in for investigation and high scores can still be screened out);
- The scores would not reflect anything about the current allegations, but rather help to aggregate historical information on the family and what that information means for future risk;
- The scores would not reflect 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 scenarios and RED Team simulations. Trainers revealed de-identified referral information, showing the call screening staff information about a family, to engage participants in a simulation of a RED Team and to debrief the relevance and impact of the risk scores on decision-making.

The assessment of ongoing training and support needs began during the four-week trial period that took place immediately following the preliminary training and prior to the implementation of the RCT. During this time, feedback was solicited and additional support and refinement to the protocols were developed in order to ensure staff have the proper knowledge and support to implement the tool. Included in these protocols is a plan for the intake manager and the administrator to share the responsibility of reviewing referrals in which the RED Team recommends screening out reports that have high scores. This real time practice, as well as a periodic plan to look retrospectively at referrals with low scores which were screened in, will be used to support supervision and ongoing training as needed.

Section 2: Building the Predictive Risk Model

Introduction

The development of the DCDA relied heavily on the experience of the research team in developing the Allegheny Family Screening Tool. However, the DCDA tool differs in some specific ways. Firstly, because Douglas County has a State level child welfare and case management system (TRAILS), we could use the full State level data to build the algorithm. Secondly, unlike Allegheny, which has integrated health and human services data, we were restricted to using data from TRAILS and CBMS - the management system for welfare programs such as Temporary Assistance for Needy Families (TANF).

In this section we describe the methodology used in building the version of the DCDA that is in use as of September 2019. It is important for the reader to note that PRM tools are constantly rebuilt as data sources increase or change. Therefore, updated reports will be provided as and when there are significant changes to the model.

Data used to build the DCDA

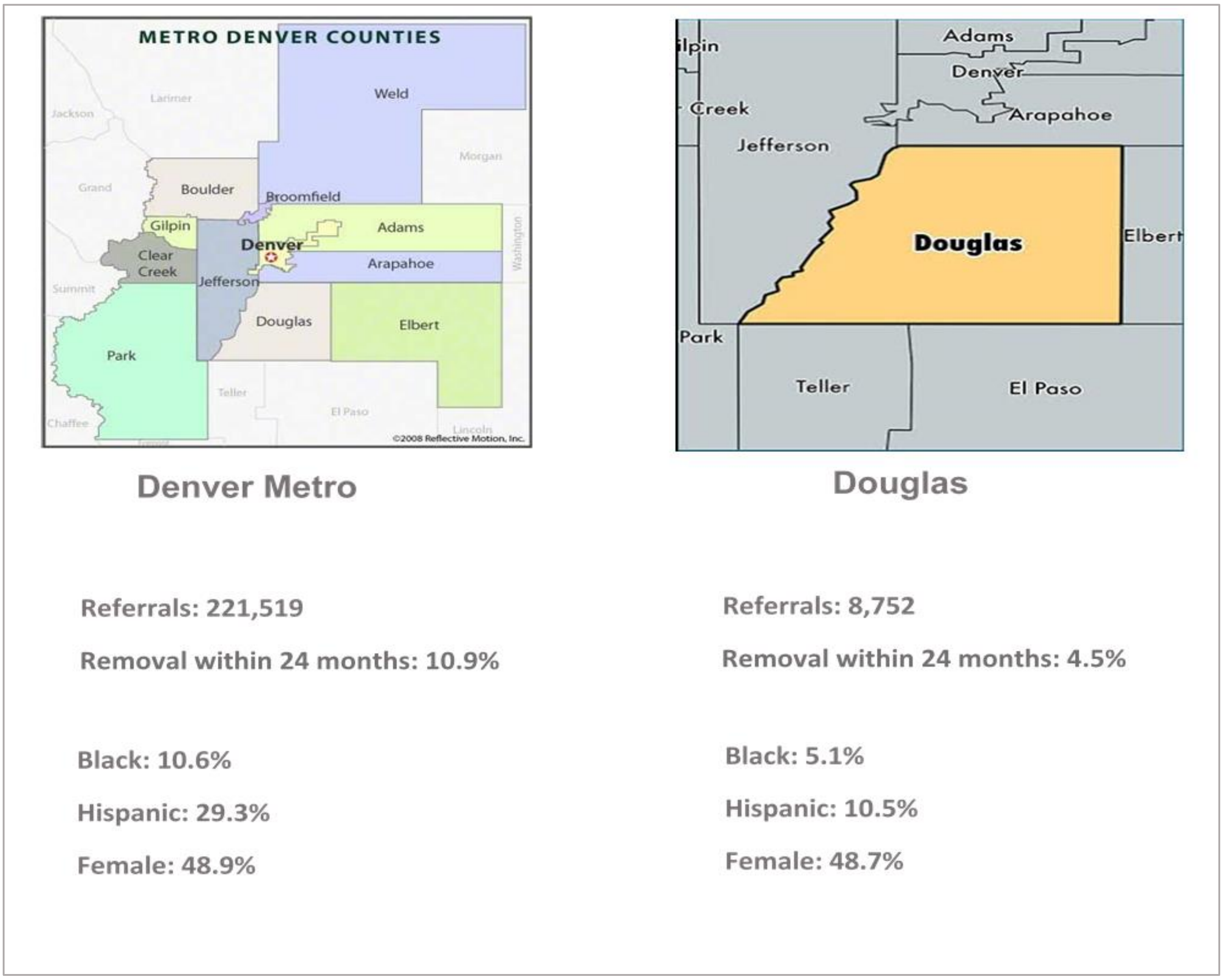
The DCDA was built using a historical research data set of referrals made in the Denver Metro Area and Larimer County from January 2015 through September 2016.[3] University ethics approval was obtained for this stage.

The construction of the data set was based on methodologies used in prior implementations.[iiiii] The data was assembled from multiple extracts of flat files, and the final research data set was a single flat file where each row represented a unique combination of child and referral. Every child listed on the referral generated a row/observation, regardless of whether the child was named as a victim or was simply another child that was associated with the call. For example, if a referral has one victim and two other children (e.g. siblings) named on it, then in the research data set it would be represented as three separate rows of data - each with a unique child ID. A risk score is generated for each child on the referral, then a risk score at the referral level is determined as the maximum of the score across all children. See section *Model Performance* for additional details.

The research data set includes 221,519 records, each of which represents a child referral. A child referral consists of 501 predictors and two outcomes. Figure 3 provides summary statistics on the research data set, including the referral count, removal rate, racial, ethnic and gender proportions for Denver Metro (excluding Douglas County) and for Douglas County only. Overall, child referrals belonging to Douglas County have lower placement rates and significantly lower Black and Hispanic rates than the larger combined Denver Metro area.

[3] Counties include Douglas, Denver, Arapahoe, Jefferson, Adams, Broomfield, Elbert, Park, Clear Creek, Gilpin and Larimer.

Figure 3: Referrals and demographic description of Denver Metro and Douglas County data



Denver Metro

Referrals: 221,519

Removal within 24 months: 10.9%

Black: 10.6%

Hispanic: 29.3%

Female: 48.9%

Douglas

Referrals: 8,752

Removal within 24 months: 4.5%

Black: 5.1%

Hispanic: 10.5%

Female: 48.7%

Features used to predict placement outcome

The DCDA uses data from TRAILS, the State-wide Automated Child Welfare Information system (SACWIS), the Client Benefits Management System (CBMS), and Public Welfare eligibility data. For each record in the research data set (at the referral-child level), predictor variables are coded for the child in focus and for all other individuals listed, based on their role in the referral. We define the *focus child* (FOCUS) as the unique child on the child referral row. Other typical roles in the call include *other children* (OTHC), *parents* (PRNT), *other adults* (OTHA), and *perpetrators* (PERP). In some cases, three roles coded as PRNT, OTHA, and PERP are taken together as the role “ADULT”. Table 1 summarizes the domains of predictor variables, including examples of the variables and the count of variables in each domain.

The demographic variables include age and gender for each role in a referral call. For child welfare history we code prior referrals, placements, case-openings and allegations. For public benefit history we code FOCUS, OTHC and all ADULT roles’ program involvement and duration, evidence of program denial, and sanctions. Active court cases and prior juvenile justice case openings are also coded.

In our preliminary experimental study we investigated models for four different target outcomes:

- Removal or placement in 12 months if screened in
- Removal or placement in 24 months if screened in
- Re-referral in 12 months if screened out
- Re-referral in 24 months if screened out

Each of these outcomes was modelled with and without 41 race-related predictors. The final model targets “removal or placement within 24 months if screened in” without race-related predictors.

Table 1: Overview of Predictors

Domain	**Description**	**Count of distinct features generated**
Demographics	Age, age category, and gender for all five roles: FOCUS, OTHC, PRNT, OTHA, and PERP.	51
Child Welfare History	There are 25 predictors for Allegation History. This gives the number of allegations found in the last 3,6,12,24 months or prior with respect to (w.r.t) the date of referral for each role. There are 93 predictors for Referral History, which indicates the number of days from last referral; a count of screened-in and screened-out referrals in the last 3,6,12,24 months or prior w.r.t the date of referral for each role; counts of referrals per one of the seven referral reasons; and dummies if referral as a child for "PRNT", "PERP", and "OTHA". There are 32 predictors for Placement History. This includes a dummy or the number of placements in last 3,6,12,24 months or prior w.r.t the date of referral for "FOCUS" and "OTHC" roles; and dummies or a count of placements for adults (PRNT, PERP, OTHA) as a child. There are 5 predictors for prior juvenile justice cases. This includes a dummy or a count of prior juvenile justice cases for each role on the referral.	155
Participation in public benefit programs	This includes involvement with programs offered by the County, such as benefits, payments, sanctions etc. "FOCUS" and "OTHC" roles have 68 predictors each and there are 82 predictors for adults (PRNT, PERP, OTHA).	218
Active cases	Dummy or number of active cases (court cases and child welfare cases) for each role on the referral.	10
Information on the current referral	Roles on the referral are given in terms of 22 dummies for victim count, non-victim children count, and "PERP" count. Six count categories for victim, and four count categories for each "OTHC", "OTHA", "PERP", and "PRNT". In addition, there are four indicators to recognize whether the focus child is an alleged victim, actual victim, sibling, or Youth-in-Crisis.	26

Modelling

The research data set was split into a training data set, which contained 154,803 (~70%) records, and a test data set containing the remaining 66,716 (~30%) records. The DCDA was built on the training data set and its performance was assessed using the test data set. The model was built for the outcome *Removal or placement within 24 months*. A total of 460 predictors remained after race predictor variables were removed.

The following three methods for generating algorithms were implemented to build the predictive risk model:

- Lasso
- Random Forest
- XGBoost

Lasso Regularized Logistic Regression

Logistic Regression models the likelihood of a certain categorical outcome variable as a function of a set of predictors. This statistical modelling algorithm has been widely used in various fields such as biological sciences, social sciences, and machine learning. In its simple form, Logistic Regression utilizes weights for all predictors despite their significance and the potential for model overfitting. By contrast, the Lasso regularized form of Logistic Regression ensures certain weights be set to zero while minimizing prediction error, given the sum of the absolute value of the weights is less than a constant. Thus, it is capable of both predictor selection and regularization, which results in more easily interpretable and more accurate models.[iv,v]

The Lasso model was implemented with the 'R' package named 'glmnet'.[vi] In the model training process, the aforementioned constant – often symbolized as *lambda* – was optimized in the range of 2.118381 e-05 to 6.936741 e-02. The best model resulted from the value 8.753175 e-04.

Random Forest

Random Forest is an ensemble learning algorithm that consists of a group of base models which are decision trees. Each tree is constructed from a random data sample drawn with replacement. When growing through splits, the best predictor to split is selected from a random subset of predictors. The overall outcome of the Random Forest model is a combined outcome that can be the majority, weighted majority, or the average of the individual tree outcomes. Compared with single decision models, ensembles are less likely to over fit as they tend to reduce variance. In addition, its randomness tends to provide more accurate results if sufficient trees are trained.[vii,viii]

The Random Forest model was implemented with the 'scikit-learn' machine learning library in 'Python'.[ix] The overall model output is the average of individual tree predictions. Five tuned parameters followed by the set of tested values: (1) the number of trees in the forest (n_estimators): [100, 200, 500]; (2) the number of features to consider when looking for the best split (max_features): [25%, 50%, 75%] of the total number of features; (3) the maximum depth (max_depth): [5,7,10]; (4) the minimum number of samples required to split an internal node (min_samples_split): [15, 20, 25]; and (5) the minimum number of samples required to be at a leaf node (min_samples_leaf): [5,10]. After assessing these 162 different combinations of hyperparameter settings, the set: [n_estimators:100, max_features: 75%, max_depth: 10, min_samples_split: 20, min_samples_leaf: 10] was selected as the best combination. The rest of the parameters were set to their default values.

XGBoost

Extreme Gradient Boosting (XGBoost) is also an ensemble of decision trees but they are trained in a different method to that used in Random Forest. The boosting via an additive training strategy expands the ensemble of trees by adding one new tree at a time.[x] We implemented XGBoost with the 'Python' XGB module.[xi] The set of tuned booster parameters are: the number of trees is set to 500; the maximum tree depth is set to 6; boosting learning rate is set to 0.01; the subsample ratio of columns is set to 0.25; subsample ratio of the training instances is set to 0.8; and the minimum loss reduction to make further partition on a leaf node of the tree is set to 1.

Seven tuned parameters followed by the set of tested values: (1) the number of trees in the forest (n_estimators): [100, 200, 500]; (2) the subsample ratio of columns when constructing each tree (colsample_bytree): [50%,75%]; (3) the subsample ratio of the training instances (subsample): [50%,75%]; (4) the weight of each tree to the model (learning_rate): [0.01,0.1]; (5) the maximum depth (max_depth): [6,8]; (6) the minimum reduction in the loss function required to grow a new node in a tree (gamma): [1,10]; and (7) the minimum sum of instance weight needed in a child (min_child_weigth): [5,10]. We assessed these 192 different combinations of hyperparameter settings and picked the combination [n_estimators: 500, colsample_by tree: 75%, subsample: 0.5, learning_rate: 0.1, max_depth: 8, min_child_weight: 5, gamma: 1] as the best setting. The rest of the parameters were set to their default values.

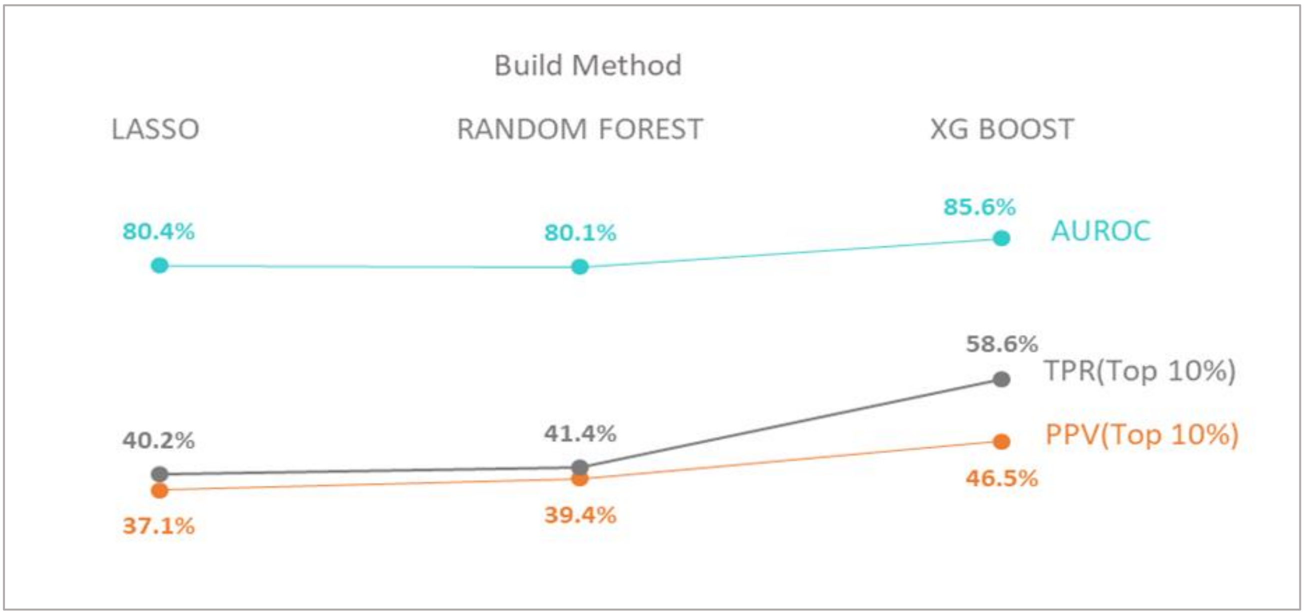
Model Performance

Each of the built models generate the probability of removal and placement within the next 24 months. Those predicted risk probabilities are then categorized into 20 equally distributed risk scores. The lowest risk 5% of referrals receive a categorical score of 1, whereas the 5% with highest risk receive a categorical score of 20. The cut-offs in predicted probability for each of the 1 to 20 bins are calculated on the research data at the whole greater Denver level. Because Douglas County has typically lower risk families than other counties, within Denver we have a higher proportion of low risk scores.

Model performance was evaluated with the Area Under the Receiver Operating Characteristics (AUROC), the True Positive Rate (TPR) for the top 10% of the predicted risk group, and the Positive Predictive Value (PPV) for the top 10% of the predicted risk group. The AUROC measure shows the ability of a given model to differentiate between a "removal" and "not a removal" at various class threshold values. The TPR is the number of predicted removals that also were actual removals (true removals) in proportion to the total number of actual removals. The PPV is the number of true removals divided by the total number of predicted removals.

Figure 4 presents results for the three methods estimated. The TPR and PPV are reported with respect to the top 10% of predicted probabilities.

Figure 4: Model Comparison (using test data, Denver Metro area)



According to Figure 4, XGBoost outperforms the other methods with a higher AUROC and better TPR. We continued with LASSO regularized Logistic Regression as the machine learning algorithm of choice for its advantages of efficiency of deployment and interpretability. The results hereafter are presented for the LASSO model.

What does the research tell us about practice at Douglas County?

Researchers generated screening scores 'after the fact' for historic calls between January 1, 2016 and April 14, 2017 to understand existing practice.

The researchers found that:

- An excessive proportion of children with the *lowest* screening scores were being *screened in* - 46% of children with a score of 1. Of those, only 0.3% were placed within two years.

- An excessive proportion of children with the *highest* screening scores were being *screened out* - over 40% of children with a score of 20. Of those, 27% were re-referred and placed within two years.
- Among the children that were referred because of a fatality or near fatality, and who had referral history, 30% had been scored a 20 at some point in a referral prior to their fatality/near fatality.

Disparities analysis

In this section we address how well the models are performing for Black and Hispanic children compared with the rest of the study sample. The population of Black and Hispanic children in Douglas County are 5.1% and 10.5% respectively. These rates are lower than the Denver Metro area.

The DCDA's capability to ensure equality – and to ensure that cases are treated equally regardless of the child's race, age group, or gender – depends on how well the models are calibrated across the sub-groups. That is, children who are Hispanic or Black who receive the same risk score should be equally likely to end up being placed within 24 months.

Figures 5 and 6 show the actual rates of removals following the risk scores across the sub-groups. It shows that the rates of removals for Hispanic and Black children are the same when stratified by the score. That means that the scores are well calibrated.

Figures 7 and 8 show the rates by age groups and gender. Both analyses show that the rates are very similar across these sub-groups.

Figure 5: Hispanic vs. Non-Hispanic removal rate, Test data, Denver Metro data

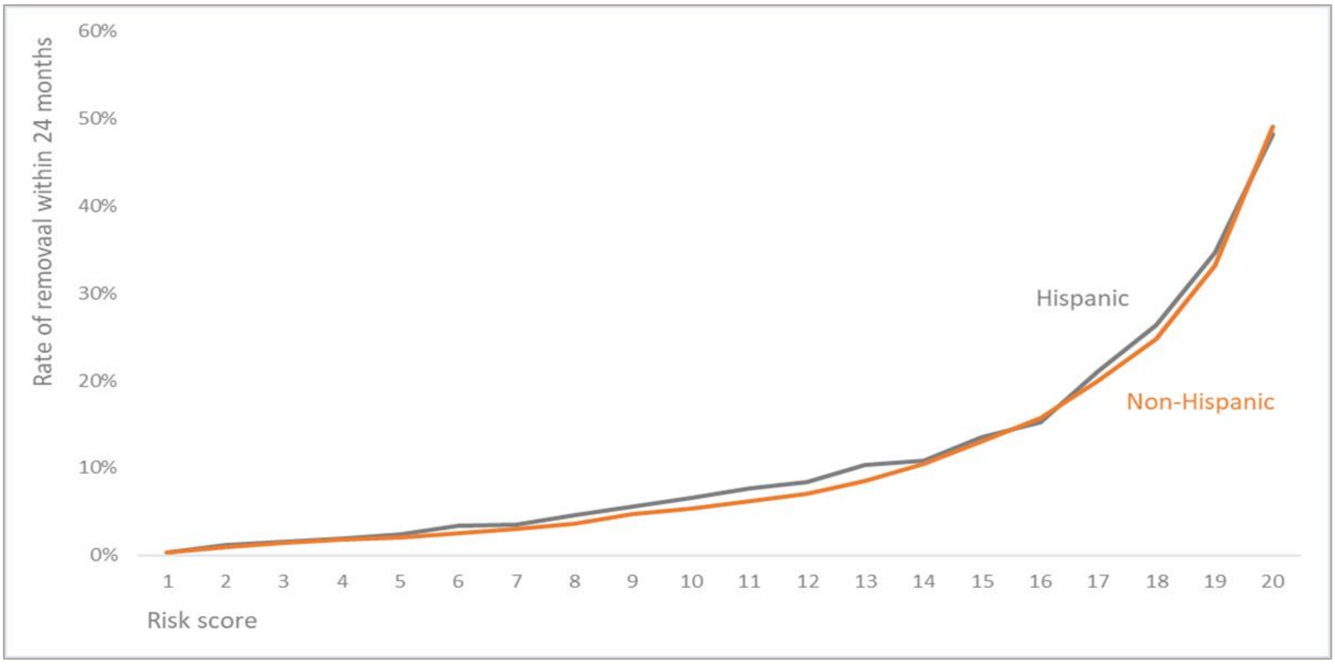


Figure 6 : Black vs. Non-Black removal rate, Test data, Denver Metro data

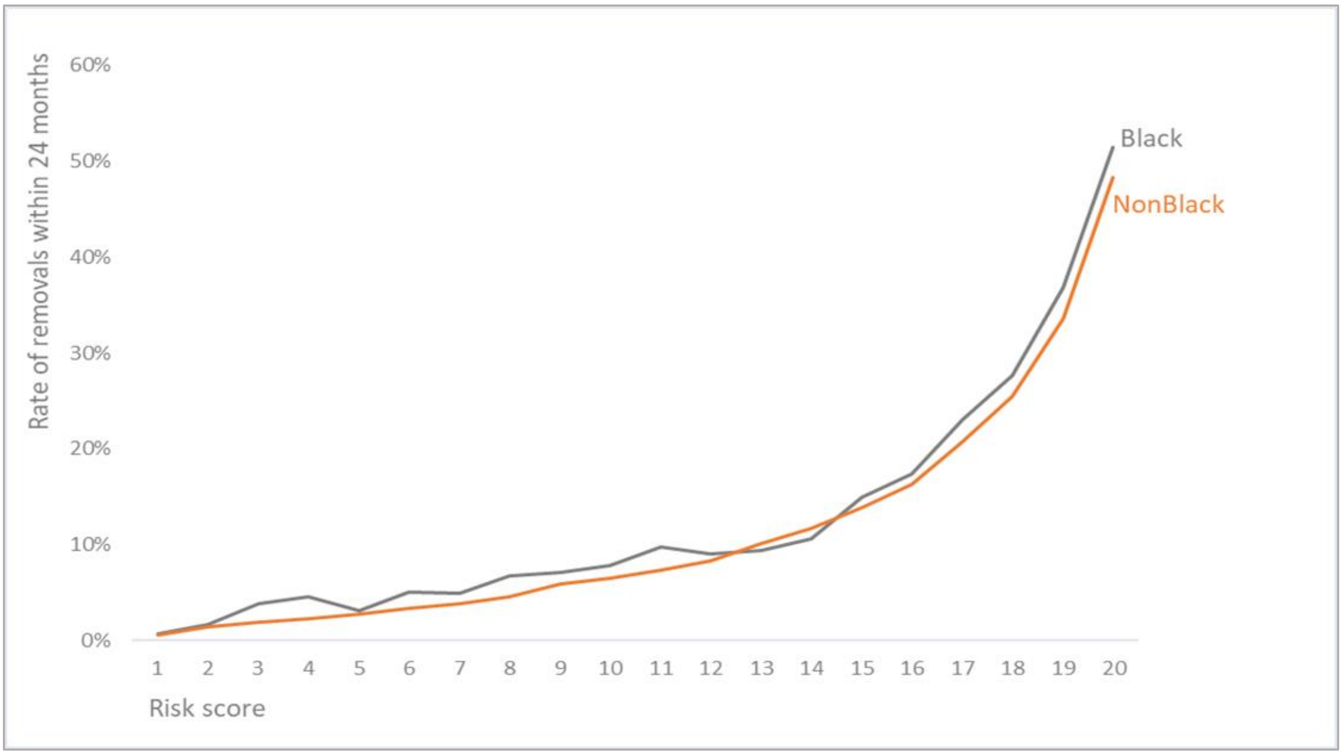


Figure 7: Age group and removal rates, Test data, Denver Metro data

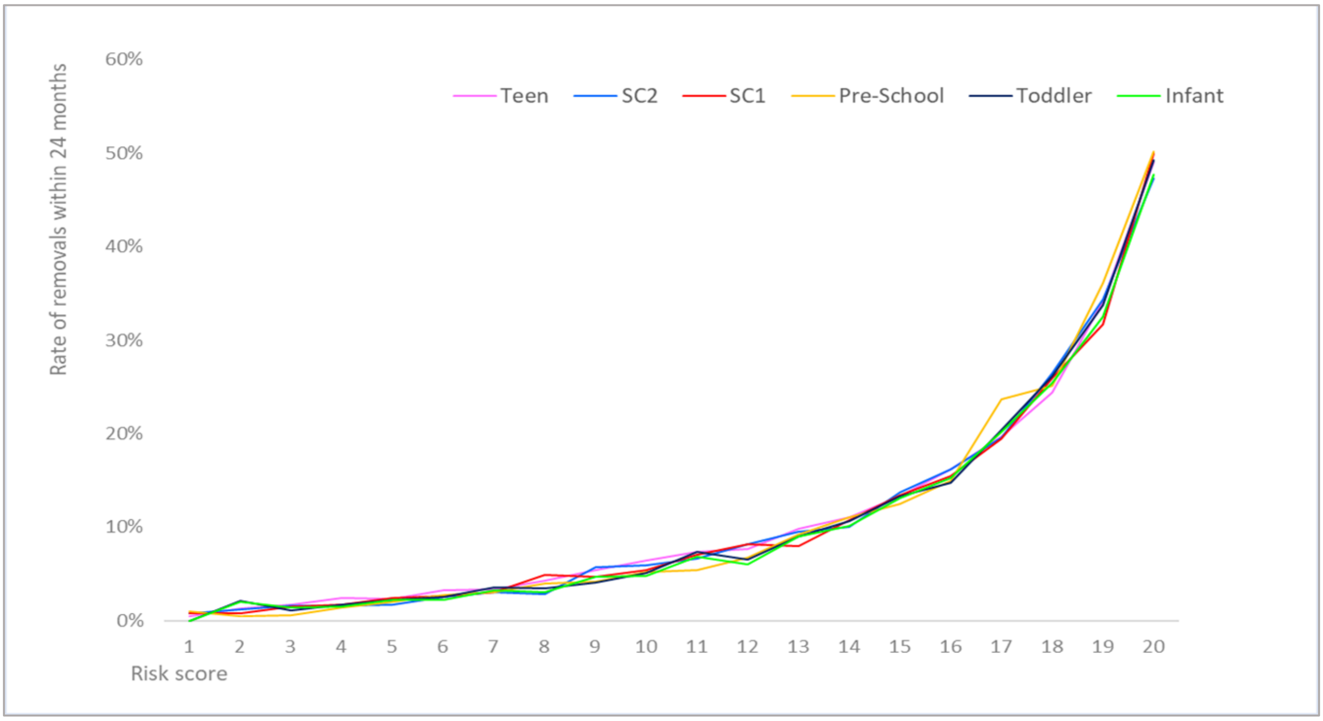
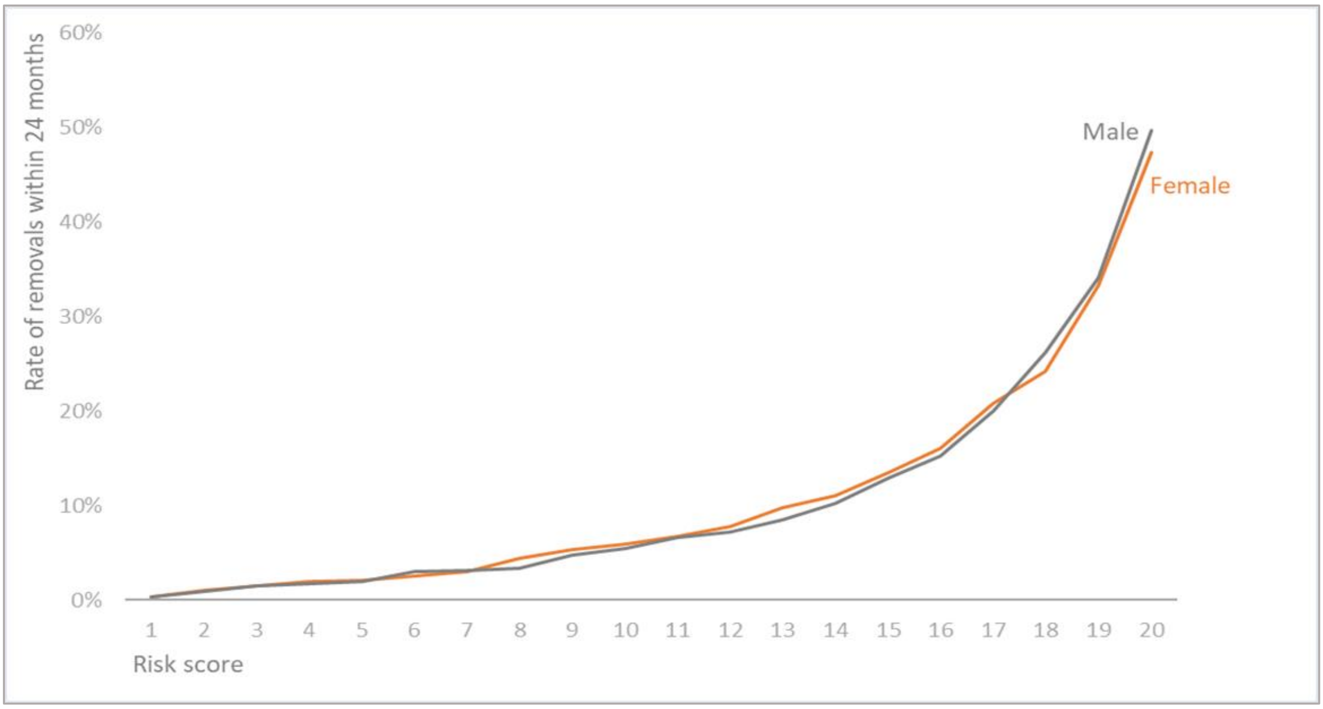


Figure 8: Gender and removal rates, Test data, Denver Metro Data



External Validation

The outcome that the algorithm is trained on (i.e. removal) is a 'proxy' for a broader set of adverse outcomes that could be avoided with better decision-making. Therefore, we may be concerned that children identified as being at-risk of removals are not at actual risk of objectively measured adverse outcomes. We need to ensure that the model is validated for a more "objective" set of outcomes. To that end, we identified a related, but independent, set of data about the same population, and checked whether we saw similar correlations between the screening score and that outcome.

To validate the model, we used case-data about children who were referred for one or more of the following critical incidents: severe, severe-egregious, near fatal, or fatal. These critical incident-related referrals were recognized based on the findings recorded in the Trails data system. We prospectively followed all children who had received a referral and had been risk-scored in our research data set (i.e. children referred between January 1, 2015 and September 30, 2016). We looked at case-data for the period January 1, 2015 through to July 19, 2017. This was the only period for which case data was available at the time of this study.

We defined the follow-up period as more than 60 days after a referral and up to the end of the case data set (i.e. July 19, 2017). The reason we looked at outcomes more than 60 days after the referral is that we wanted to ensure that the scored referral wasn't actually the same episode as the egregious referral - and that, as much as possible, these were distinct referrals related to distinct events. We recorded a child as having a severe, severe-egregious, near fatal, or fatal incident if there was a case file recording such an event in the follow-up period. Note that the follow-up period is of varying lengths depending on when the scored referral occurred, but we did this to make full use of the data available and because of the relatively small number of these critical cases.

Our analysis found that there were 957 unique child referrals during the period January 1, 2015 and September 30, 2016, that had a *severe, severe-egregious, near fatal, or fatal* case in the follow up period. This represents 432 unique children – of which 53 were Black and 169 were Hispanic. Figure 9 describes the distribution of these cases by each sub-group.

Figure 9: Distribution of child referrals that had severe, severe-egregious, near fatal, or fatal cases in the follow-up period

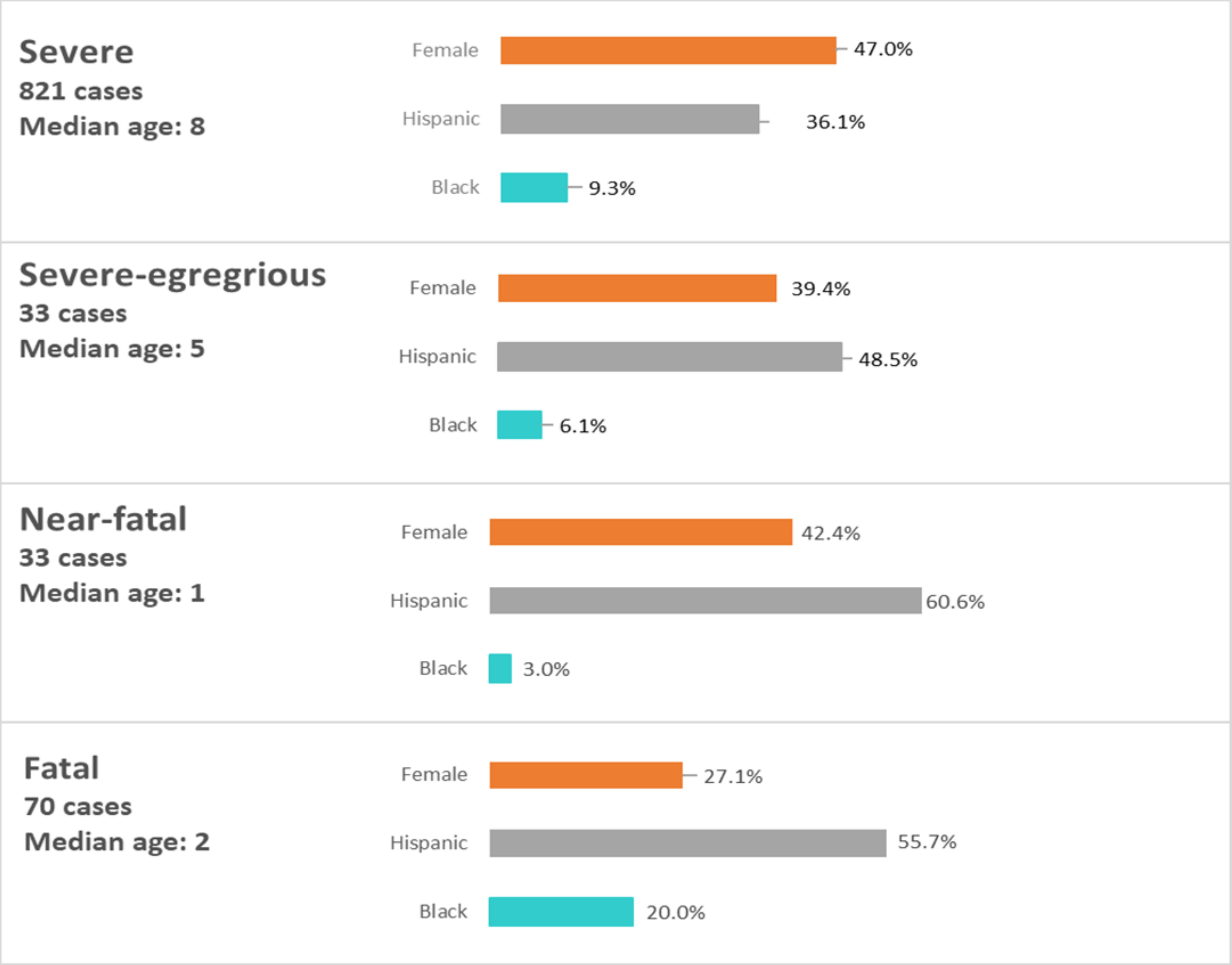
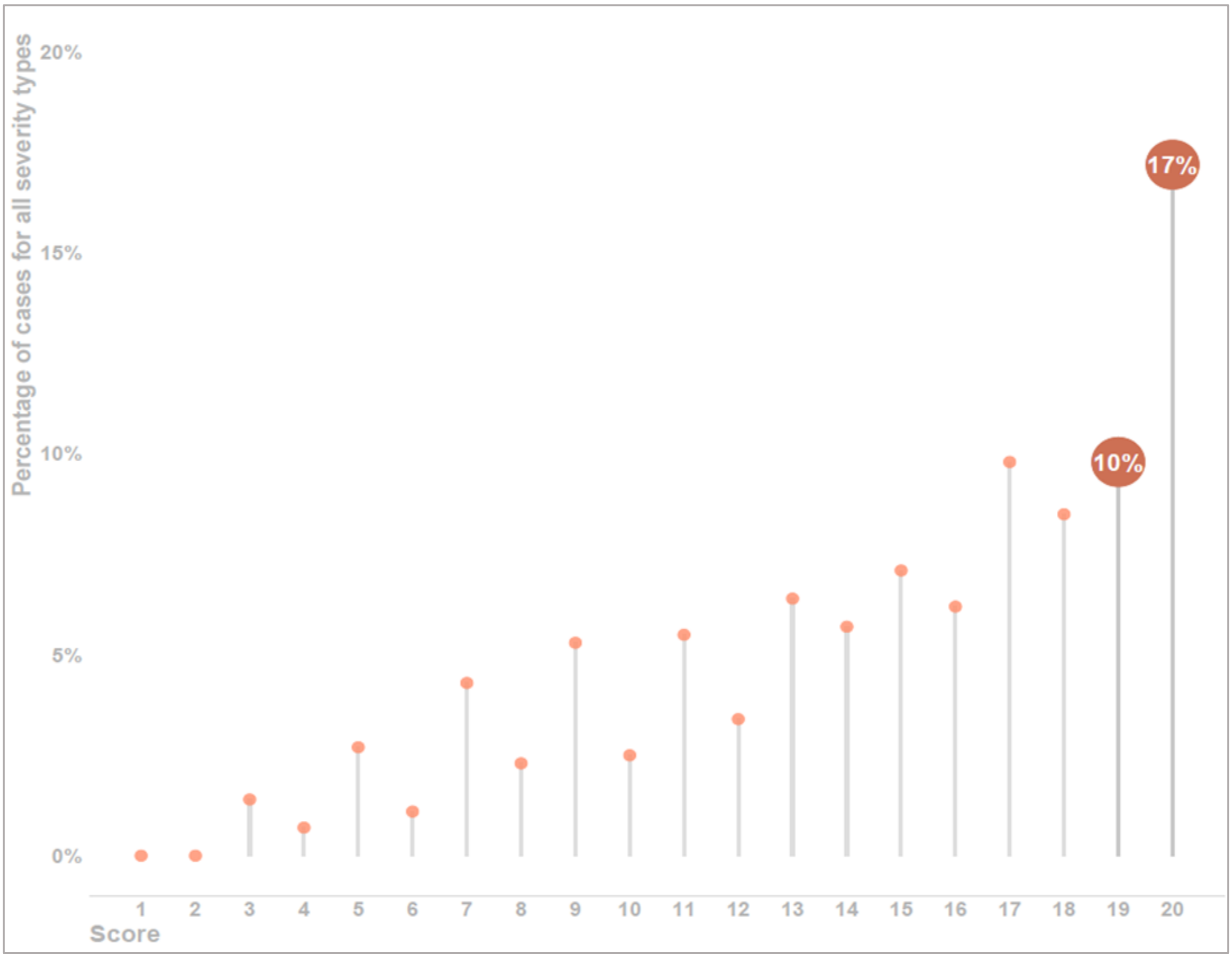


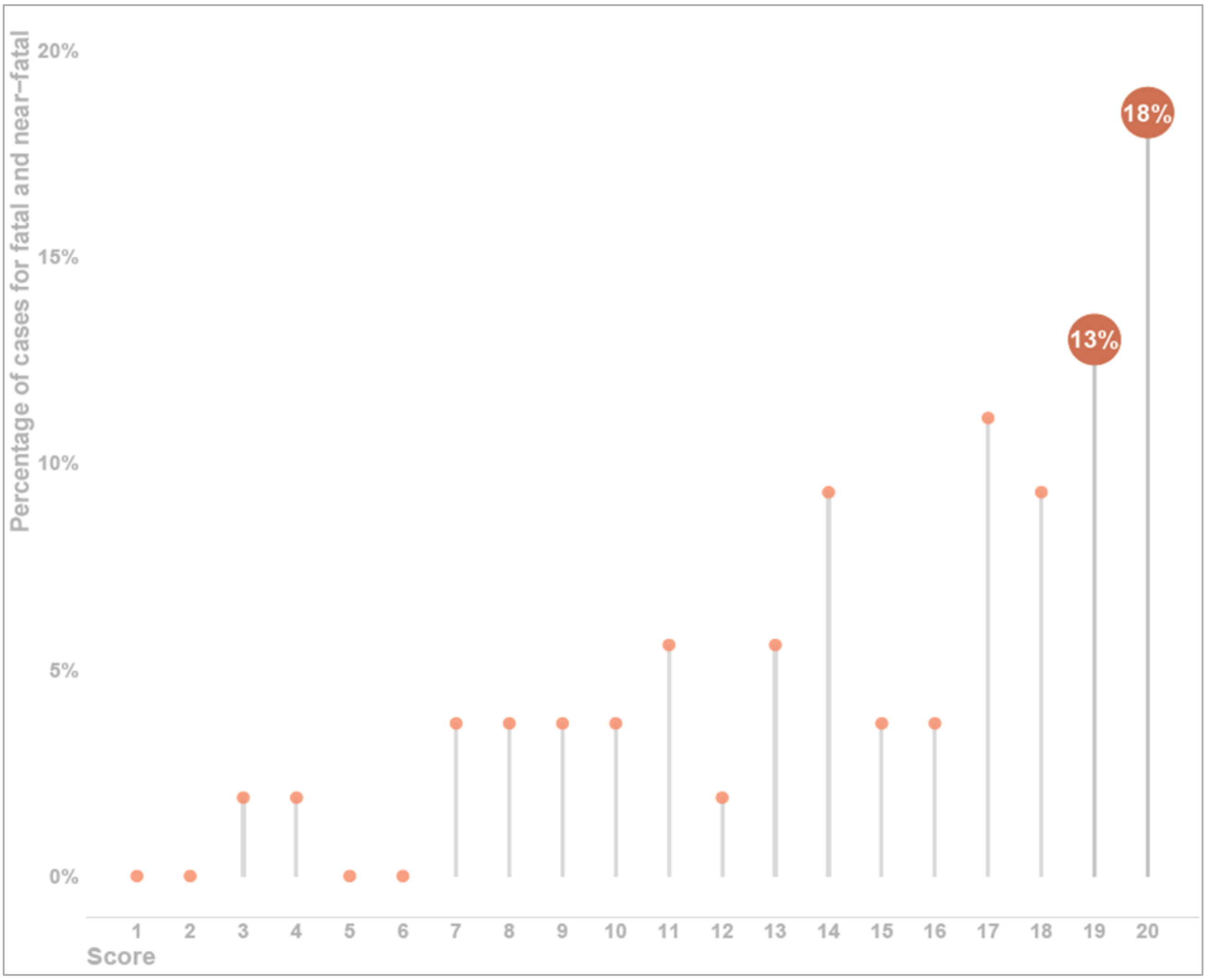
Figure 10 depicts the percentage of severe, severe-egregious, near fatal, and fatal cases by the risk score received in a previous referral. Where a case had multiple prior referrals, we took the maximum score. Accordingly, more than 1 in 4 cases (~27%) of severe, severe-egregious, near fatal, or fatal incidents scored a 19 or 20 (i.e. in the top 10% highest risk) at some point in a referral that occurred more than 60 days before the case. More than 50% of such cases were previously identified as being in the top 25% highest risk. Conversely, only 5.8% of cases scored in the bottom 25%.

Figure 10: Severe, severe-egregious, near fatal, and fatal cases by their risk scores received in a referral more than 60 days prior to the severe referral



The percentage of near fatal or fatal cases by their previous risk score is illustrated below in Figure 11. Almost one-third (~31%) of near fatal or fatal cases were scored a 20 or a 19 (i.e. in the top 10% highest risk) in a referral that occurred more than 60 days before the case. Over 55% of such cases were picked up at the top 25% of risk. Conversely, only 3.7% of cases were scored at the bottom 25% of risk.

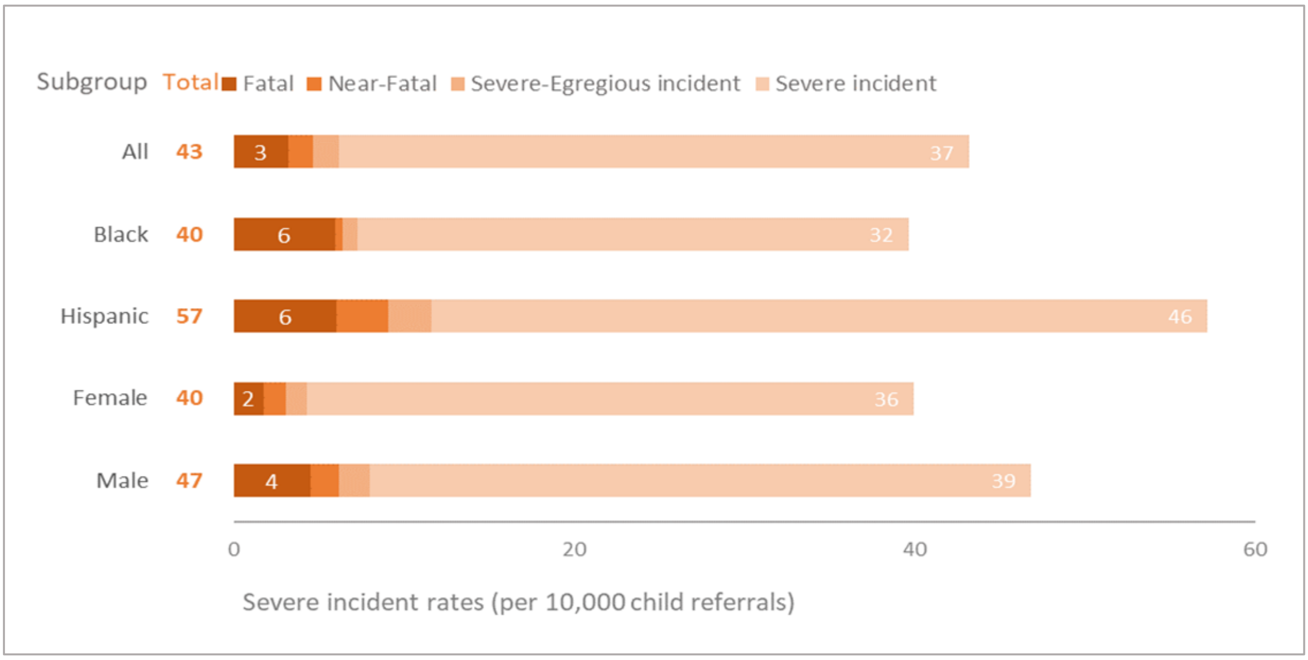
Figure 11: Fatal and near fatal cases by their risk score received in a referral more than 60 days prior to the fatal



Sub-group analysis

We also studied the differences in the True Positive Rate (TPR) for fatalities and near fatalities across sub-groups based on race and gender. For these analyses, we first determined if there were existing disparities by comparing the rates of critical events per sub-group. Figure 12 shows the critical event rates for sub-groups based on race and gender. Black and Hispanic children have higher rates of fatalities. Black child referrals represent 20% of the fatality referrals but only 11% of the referrals with no critical event. Hispanic child referrals represent 56% of the fatality referrals but only 30% of those referrals with no critical event. Females are under-represented and account for half of the non-critical referrals, and only make up 27% of fatality referrals.

Figure 10: Critical incident rates per subgroups based on race and gender (per 10,000 children-referrals)



Note: Excluded children with race or gender coded as missing

Now, we compare how the DCDA performs for sub-groups based on race and gender discussed above. For this purpose, we studied the relative risk of fatal and near fatal case incidents for those with a score of 17-20 compared to those with a score of 16 or less (see Table 2). The first segment shows the relative risk of fatal and near fatal cases. The next segment shows only risk of fatalities, and the last shows near fatalities.

The table shows that children who at some stage score 17 or higher are 3.9 times more likely to have a case opening for a fatality or near fatality in the follow-up period. For Black children, this relative risk is lower (1.40) and for Hispanic children the relative risk is higher (5.63). This validation provides reassurance that while the DCDA was built to predict removals, it picks up elevated severe harm for each of the sub-groups. The fact that relative risk rates for Black children are low does suggest that the tool is not as sensitive to the potential risk of harm amongst Black children as for Hispanic children and others. Due to the small numbers, the subgroup analysis is not statistically significant - in that the 95% confidence intervals overlap across all racial sub-groups. However, it does suggest a need to be vigilant. In particular, call screeners should be made aware that the correlation between the risk score and the risk of severe harm varies between racial groups.

Table 2: Relative Risk of fatal and near fatal cases following a score in the top 20% (i.e. score > 16)

	Relative Risk of <u>Fatal or Near Fatal</u> incidents following a score >16 (i.e. in the top 20%)	**Relative Risk of <u>Fatal</u> incidents following a DCDA score >16 (i.e. in the top 20%)**	**Relative Risk of <u>Near Fatal</u> following a DCDA score >16 (i.e in the top 20%)**
All	3.9 [2.64, 5.76]	4.36 [2.7, 7.04]	3.09 [1.56, 6.12]
Black	1.40 [0.51, 3.87]	1.60 [0.56, 4.57]	*Low count*
Non-Black (excluding race-miss)	3.82 [2.49, 5.87]	4.50 [2.60, 7.8]	2.92 [1.46, 5.84]
Non-Black (including race-miss)	4.56 [2.98, 6.97]	5.26 [3.06, 9.03]	3.58 [1.79, 7.16]
Hispanic	5.63 [3.17, 10.0]	6.08 [2.96, 12.47]	4.89 [1.88, 12.72]
Non-Hispanic (excluding race-miss)	1.52 [0.82, 2.82]	1.96 [0.95, 4.03]	0.77 [0.21, 2.79]
Non-Hispanic (including race-miss)	1.93 [1.05, 3.57]	2.44 [1.20, 4.99]	1.02 [0.28, 3.69]
Female	2.74 [1.38, 5.43]	1.92 [0.75, 4.87]	4.38 [1.52, 12.62]
Non-Female (excluding gender-miss)	4.57 [2.80, 7.47]	6.06 [3.31, 11.09]	2.34 [0.95, 5.75]
Non-Female (including gender-miss)	4.41 [2.72, 7.16]	5.70 [3.16, 10.3]	2.35 [0.95, 5.77]

Section 3: Deployment of the DCDA

Match-Merge

One of the first issues in deploying a predictive risk modelling tool is to ensure that individuals are correctly matched with their histories. If there is poor entity resolution (i.e. matching individuals across systems and time in TRAILS and CBMS data systems), we might incorrectly score a family. For example, the tool may have failed to find a full record of the family's interaction with the system or have inadvertently ascribed incorrect history to an individual.

Because the predictive risk model is being run using a State client management system (TRAILS and CBMS), the matching of the individuals will rely on using matching algorithms that are currently used by the County. These use first name, last name, SSN and DOB. The rule is that unless all fields are perfectly matched, the client will not be matched.

This means that when the DCDA tool is deployed at the time of the referral, some of the fields (e.g. DOB) may be missing for some of the people on the call. If this occurs, the algorithm will only have partial historical information about the client. There can be an impact on scores resulting from missing a client link (missing the associated history) or incorrectly linking clients that are not the same person (associating incorrect history).

We undertook a research exercise to test whether the match-merge currently used by the data systems could be improved upon. Using an open source probabilistic matching algorithm, we compared Douglas County referrals between January 2016 and April 2017. The analysis included all individuals listed on the referral (e.g. children, parents, perpetrators) – a total of 8,117 clients. The research team assessed how many matches were found by both the Douglas County client linking method and the open source method. All cases where a match was identified through one method, but not by the other, were reviewed to identify if they were true positives or false positives. Although the research exercise found that the Douglas County method was very high performing, the team had a set of recommendations (additional rules) to improve the matching algorithm, which the County implemented.

Quality assurance monitoring

While generating predictions from models it is important to understand that their predictive power is dependent on the quality of the data. This becomes even more significant when the data from which the model was built is different (due to passed time and a consequently different situation) to that of the environment in which the model is being used for inference. This problem is commonly known as concept drift or structural break. Note that concept drift is ultimately about whether the underlying data generating process that delivers the covariance structure of the research data set might change over time, because of policy of institutional changes, for example. A very rich array of features, a research data set that spans a long period and multiple counties, and predicting long-arc risk will all reduce the risk of such drift.

One of the challenges is that the statistical properties of the underlying data might have changed over time, negatively impacting the predictive power of the model. It is important to detect when this is happening and to alert the end user about it as soon as possible.

Another challenge is identifying exactly what is known in real time at the point at which the DCDA is run. For example, it could be that some of the roles (e.g. other adults in the household) are not identified at the time the call is taken at the hotline. These details could be retrospectively filled in as the call progresses to investigation. The issue of training a model using features that are not available to the algorithm at the time of the call is called "forward variables". While we look for such variables when we build the research data set – by using the time stamps wherever possible, for example – we still need to re-test predictive accuracy when the model is being run on live data.

Because of the variety of challenges in ensuring that production performance is as good as research performance, we have instituted the following ongoing tests and monitoring reports:

a) *Changepoint detection (structural break) tests for each feature*. We have an automated ongoing structural changepoint detection test pertaining to a 100-day moving average. Whenever a feature appears to have a break (i.e. a statistically significant change in temporal trends of the data), the suspect feature is reported.

b) *Feature drift-detection*. In addition to the structural break test, we have instituted a feature drift alert that reports features if the 100-day moving average drifts outside standard deviations of the features of the original research data.

c) *Quarterly report on Fielded AUC.* We also report on the ongoing AUC for placement rates by risk-scores and compare to the research AUC. While the research AUC is related to 24-month placement outcomes and the fielded AUC is simple placement up to the date of calculation, generally a fall in the AUC over time is indicative of a loss in predictive power of the model.

It should be noted that some quality assurance is implicitly provided by how the model is deployed. Specifically, the schema of the data and the process of running the jobs to generate scores is such that we can be sure that if a score is generated then all data sources were present and are in the format that they are expected to be in. If one of the upstream data sources is not available or the data coming in is not as expected (e.g. in the correct format, within the expected range), then the job itself will fail. This helps to narrow down the set of problems that might occur.

In implementing the automated Quality Assurance (QA) tool, we want to make sure that the statistical properties of the underlying data have not significantly changed from when the model was built. We also want to handle some of the more obvious causes of error that may arise. We have applied the following techniques to give us more certainty that the predictions generated from the model are as we expected.

Changepoint Detection

Changepoints (also known as structural breaks) are an important consideration when analyzing time series data. A changepoint occurs when there is a difference in the statistical properties before and after a point in time. We are mainly concerned with the differences that might have occurred in the underlying data since the model was built, as well as changes that are taking place in the data streams.

In the research data set we have several referrals coming in each day. We also have the whole set of input and output parameters and the scores that were generated for them. For each day, we have taken the means of each of the input parameters (there are 460 input features in the model). The research data set contains around 19 months' worth of data (from January 1, 2015 to September 30, 2016, with two months missing in between). We started by taking the average for each day of all the parameters, then took a 100-day rolling mean of such. This gives us 479 time series data points, one for each day for each of the 460 input variables. We splice this time series data to the new data generated after the live deployment of the model (i.e. February 2019). As the deployment period rolls through, the 100-day rolling average consists of more post-deployment data and less research data.

Next, we have applied the changepoint detection method for observing changes in mean, using R language's "Changepoint" package. We have applied the 'offline' method of analysis, which means that we do these calculations in a batch (once a month) instead of 'on the fly'. We are prioritizing accuracy over faster detection of these changes. We are also interested in multiple such points rather than just the one between research and current data. There are multiple methods for calculating these points (e.g. At Most One Change, Binary Segmentation). Based on the constraints outlined above, we chose a method that is precise as well as computationally fast.

The method of detection applied is PELT (Pruned Exact Linear Time). We made the decision that it is more useful to look at just the changes in mean and not variance, or a combination of both. Because the data is mainly categorical, the variance is not as crucial. We also wanted to avoid creating too many false alerts as this can cause the user to become de-sensitized to actual alerts when they arise.

Process Control

There is existing literature about the use of these methods to monitor and provide quality assurance and exception reporting in industrial processes. Here we have borrowed some of the metrics and charts that are traditionally used in such fields, and applied them to monitor the overall quality of the data coming in.

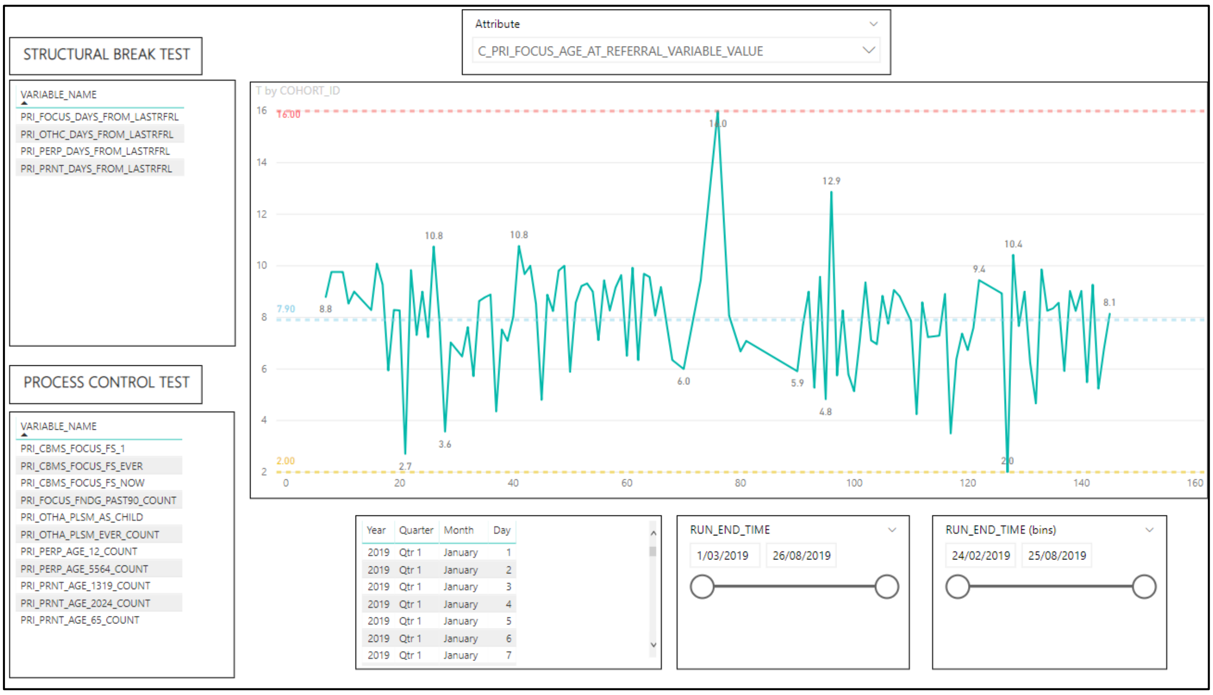
We have used the concept of control charts to determine whether new incoming sets of data satisfy the limits/criteria as specified from our research data set. Also known as Shewhart charts, process behavior charts are used in statistical process control. To specify when to alert users, we define an upper and lower control. We calculated the upper and lower limits based on the research data. We calculated means for each day and applied a 100-day rolling average for both the research data set and the incoming data stream, since the model's deployment.

Various values were tested for calibrating the sensitivity of these charts and to alert the user when they were out of range. 1.5 standard deviations of the mean of the sample is the usual recommended number, but in this case it was too sensitive and would generate far more alerts when there is not necessarily an error in the underlying data. We have chosen three (3) standard deviations as the metric to calculate these limits. The values must also be out of range consecutively for several data points in order to generate an alert. This is called the run length and is usually recommended as seven (7) or ten (10). We tested various ranges and decided that a run length of 30 days is appropriate for the sensitivity level that is needed.

In these tests we have tended to be conservative while generating alerts. This is mainly because while the tool is alerting for an anomaly, it might not exist. The primary purpose of the tool is to alert the user so that a more thorough investigation can be conducted on why the issue is occurring.

Visualizing alerts

We have created Power BI dashboards for visualizing the control charts and alerts generated, based on both methods outlined above. The central figure is the control chart, which displays the daily mean of any selected feature (and score). Below, the feature displayed is the age of the child as at the referral. The "structural break test panel" shows the alerts for those features that have shown some evidence of having had a structural break. The process-control panel shows the features which have drifted out of the control regions for 30 days.



STRUCTURAL BREAK TEST

VARIABLE_NAME
PRI_FOCUS_DAYS_FROM_LASTRFRL
PRI_OTHC_DAYS_FROM_LASTRFRL
PRI_PERP_DAYS_FROM_LASTRFRL
PRI_PRNT_DAYS_FROM_LASTRFRL

PROCESS CONTROL TEST

VARIABLE_NAME
PRI_CBMS_FOCUS_FS_1
PRI_CBMS_FOCUS_FS_EVER
PRI_CBMS_FOCUS_FS_NOW
PRI_FOCUS_FNDG_PAST90_COUNT
PRI_OTHA_PLSM_AS_CHILD
PRI_OTHA_PLSM_EVER_COUNT
PRI_PERP_AGE_12_COUNT
PRI_PERP_AGE_5564_COUNT
PRI_PRNT_AGE_1319_COUNT
PRI_PRNT_AGE_2024_COUNT
PRI_PRNT_AGE_65_COUNT

Attribute: C_PRI_FOCUS_AGE_AT_REFERRAL_VARIABLE_VALUE

Year	Quarter	Month	Day
2019	Qtr 1	January	1
2019	Qtr 1	January	2
2019	Qtr 1	January	3
2019	Qtr 1	January	4
2019	Qtr 1	January	5
2019	Qtr 1	January	6
2019	Qtr 1	January	7

RUN_END_TIME: 1/03/2019 – 26/08/2019

RUN_END_TIME (bins): 24/02/2019 – 25/08/2019

Further Enhancements

While checking for values such as “NULL” is handled by the database schema, other common cases (like a parameter having its values as zeroes) must be checked manually. We generated synthetic data to test some variables and found that the aforementioned issues can be caught by the structural break test. In order to be alerted about such issues earlier, we can explicitly add edge cases like these to check for zeroes, for a given run length of 10 or 30 days.

One of the reasons for selecting the value of three (3) standard deviations is that for a normal distribution we can expect 99.7% of the data to fall within that range, which is helpful when you want to set limits. But we have observed that the distribution is not normal for some of the variables. In these cases, simply selecting three (3) standard deviations does not make sense.

However, with non-normal distributions we can also apply more advanced statistical tests to see if the two distributions (research and current) are similar. This can be done in conjunction with current QA methods.

Despite these quality checks, changes in the underlying data may still be missed and undermine predictive accuracy. For example, if 50% of the sample are coded as men and the rest women, and the underlying encoding was accidently switched, this would have a major adverse effect on accuracy yet not create any evidence of a feature drift. It is therefore important to maintain overall vigilance on what is happening with underlying databases and to consider the implications of any upstream data changes to the PRM tool.

Section 4: Evaluation and Ethics

As part of Douglas County's and the research team's commitment to an evidence-based and ethical approach to implementing predictive risk modelling in child welfare, it was decided that the project would be subject to an RCT (Randomized Control Trial). It was also decided that the Ethical Evaluation undertaken for Allegheny County would form the basis of the ethical underpinnings for the DCDA. This section outlines in more detail the RCT and the ethical framework used to support the implementation of the DCDA.

Impact Evaluation

The evaluation is being independently conducted by Professor Chris Wildeman and Associate Professor Maria Fitzpatrick at Cornell University. The RCT required that there would be two types of referrals - in the language of RCTs, *treated* and *control*. For referrals sent to the RED Team, randomization would occur at the RED Team level. An administrator would use an online randomizer in the morning prior to the start of the RED Team process. *Treated* RED Teams would see the scores, and *controls* would continue to conduct the RED Team process as they had always done.

The evaluators also requested that in order to evaluate the impact of the score on decisions and on business practices, certain information would need to be recorded for all reports (such as the start and end time of each RED Team meeting, as well as the screening decisions), and some RED Team meetings would be recorded. To ensure there were systems in place to determine and manage the entire process, a protocol was developed and documented as the "RCT Protocol". The intention of the evaluation is to test whether the scores are being used by call-screening staff, and whether they have improved the quality of decision-making by Douglas County.

Ethics

Due to the similarities between the DCDA and the Allegheny Family Screening Tool (AFST: an earlier predictive risk model built to support Child Welfare decision-making in Allegheny County), Douglas County leadership chose to rely upon the ethics report prepared for the AFST rather than commissioning a report specifically for the DCDA. Read the AFST Ethics Report here (from p.49).

The main issues raised in the Ethics Report (bold) and summary of Douglas County responses are:

a) **Consent (If DHS is already entitled to access data gathered by the tool in response to a referral then it is legitimate to regard the tool as a new and more effective way of doing something already permitted – so long as the tool largely delivers information that would have been available in principle to the RED team.)** No additional data is accessed, DHS owns and has rights to the data used, and tool output is only used by individuals who would have access to existing data for decision-making. All data use is consistent with existing data use policies re: HIPPA.

b) **Information about other family members. There should be protocols around this information but assume that diligent screening staff are already entitled to gather this information**. Significant too that information **is used in response to a referral, rather than used proactively.** See response to 'Consent' above.

c) **False positives/False negatives (errors cannot be entirely eliminated so must approach this issue comparatively – consider in light of alternatives that have costs of their own.)** The tool has the advantage of being more accurate than current decision-making strategies and more transparent. Performance challenges (including error margins and racial disparity) should be monitored and mitigated as much as possible. Note that interventions are protective (not punitive) and use of the tool at an early stage allows for additional information and decisions to help determine appropriateness of referral.

d) **Stigmatization** (Suggested responses include careful control over dissemination of score – only to those with training and who need the information, emphasis on training that emphasizes the possibility of false positives/false negatives, emphasizes that risk scores are not determinative, trains against confirmation bias). See response to 'False Positives/False Negatives' above.

e) **Racial Disparity** Disparities are found in existing data and if they reflect race-based bias rather than genuine differences in 'need', that may be ethically problematic – the fact that DCDA will prompt further detailed inquiry into the family situation and any intervention is designed to assist (not punish), means the model is not vulnerable to concerns generated by disparate data used in a punitive context – e.g. predictive policing. In fact, it is possible that the DCDA may make it possible to track and correct for disparities that may have otherwise remained hidden. See response to 'False Positives/False Negatives' above.

f) **Professional Competence/Training (If DCDAs are to operate ethically, staff must be competent with use and interpretation. See specifications about training above, under "Stigmatization".)** Influenced by the reviewers' suggestions, training emphasized the tools' specific meaning and limitations, and explored how its content should be incorporated into decision-making. Creation of a thorough 'job aide' document to ensure consistent use of the tool. [Douglas staff also have input to final business process/policy decisions during the trial phase, ahead of formal start of RCT.]

g) **Provision and Identification of effective interventions (Why predict better if other parts of the process – assessments/interventions – are not evidence-informed? Need to recognize that there is a whole sequence of decisions that affect outcomes and hope that use of more accurate risk assessment may positively influence that continuum.)** Agree that ultimate interventions aim to be protective not punitive. Use at early stage (screen in/out) leaves investigation phase to help confirm or deny the appropriateness of the referral.

h) **Ongoing monitoring (Essential that County commit to ongoing monitoring to ensure tool/staff training maintained and interventions as effective as possible – required to counter-balance legitimate ethical concerns.)** DHS has contracted with independent researchers to analyze the quantitative impact on system trends and outcomes. DHS will also be carefully monitoring internal use and impacts of the tool. DHS intends to have the content of the model revisited within the first year – to ensure statistical performance still strong and update underlying weights as necessary.

i) **Resource Allocation (Whether DCDA is ethical depends to large extent on whether it can deliver benefits that outweigh 'costs' – will require training, monitoring, effective intervention and adequate resources – mustn't be seen as chance to reduce resources or reallocate child protection professionals, leading to reduced benefits – as ethical justification relies on those benefits.)** Many design elements were impacted by ethical concerns, including: the tool never overrides clinical judgement, the score is only accessible to trained staff who need access to score, shared concerns about need for more evidence-based interventions. The tool is just one element.

References

[i] United States Census. (2018). Quick facts: Douglas County, Colorado. https://www.census.gov/quickfacts/fact/table/douglascountycolorado/PST045218

[ii] Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In Conference on Fairness, Accountability and Transparency (pp. 134-148).

[iii] Vaithianathan, R., Jiang, N., Maloney, T., Nand, P., & Putnam-Hornstein, E. (2017). Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology. Centre for Social Data Analytics. Auckland: Centre for Social Data Analytics.

[iv] Santosa, Fadil; Symes, William W. (1986). “Linear inversion of band-limited reflection seismograms”. *SIAM Journal on Scientific and Statistical Computing*. SIAM. 7 (4): 1307–1330. doi:10.1137/0907087

[v] Tibshirani, Robert (1996). “Regression Shrinkage and Selection via the lasso”. *Journal of the Royal Statistical Society*. Series B (methodological). Wiley. 58(1): 267–88. JSTOR 2346178

[vi] Friedman, Jerome, Hastie, Trevor, Tibshirani, Robert (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/

[vii] Ho, Tin Kam (1995). Random Decision Forests (PDF). *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282*. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.

[viii] Breiman, L. (2001). “Random Forests”, *Machine Learning*, 45(1), 5-32, https://doi.org/10.1023/A:1010933404324,.

[ix] Pedregosa, F., *et al.*, (2011). Scikit-learn: Machine Learning in Python, JMLR 12(Oct), 2825-2830.

[x] Friedman, J.H. (2001). Greedy Function Approximation: A Gradient Boosting Machine, *The Annals of Statistics*, 29(5), 1189-1232.

[xi] Chen, Tianqi; Guestrin, Carlos. (2016). “XGBoost: A Scalable Tree Boosting System”. In Krishnapuram, Balaji; Shah, Mohak; Smola, Alexander J.; Aggarwal, Charu C.; Shen, Dou; Rastogi, Rajeev (eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. ACM. pp. 785–794. arXiv:1603.02754. doi:10.1145/2939672.2939785.