OREGON DHS SAFETY AT SCREENING TOOL – DEVELOPMENT AND EXECUTION

SUMMARY

Oregon DHS has developed a real-time decision support tool to provide accurate and equitable indications of the likelihood of future adverse outcomes for children named in reports of abuse/neglect. The gravity and difficulty of the decision to investigate a report represents an opportunity for a data-informed decision support tool. In order to prevent the automation and continuation of decision inequity, DHS uses a proactive statistical approach to address existing bias in the historical data used to generate the support tool. The results indicate that a real-time decision support tool can increase the effectiveness, efficiency, and equity of the Child Welfare screening process.

BACKGROUND

In Child Welfare, *screening* is the act of a specially trained government employee receiving a report of alleged child abuse or neglect, usually by telephone, and deciding whether to assign the report for further investigation by Child Protective Services (CPS).

PROBLEM STATEMENT

In line with a national trend, screening staff in Oregon now assign for investigation between 45% and 52% of all reports of child abuse or neglect. In 2018, screening staff assigned 51% of 85,974 reports for investigation, up from 43% of 67,466 reports in 2012. This increased volume of investigations represents a strain on CPS resources, and potentially leads to thousands of undue or delayed investigations of child abuse/neglect per year. When following reports out for 2 years, those assigned for investigation resulted in a child being removed from home 13% of the time, while reports that were not assigned for investigation still led to a removal 8% of the time. In other words, there is an opportunity for the screening process to benefit from increased efficiency and accuracy.

Many factors may contribute to the lack of efficiency and accuracy of the screening process, including time pressure, caseload stress, policy constraints, statute interpretation, and the natural risk aversion that may come with not wanting to ignore a true abuse/neglect situation. Thus, Oregon DHS sought a data-informed system to aid in the screening process.

Administrative data systems represent years of human data entry and decision making, and thus likely contain inaccuracies as well as bias. Because equity between race and ethnicity groups is an integral part of DHS's mission, Oregon sought a path to directly address this historical bias. This process involves the application of statistical procedures which were proposed in academic literature.

The purpose of this document is to provide a broad overview of the background, development, and implementation of a statistical prediction technique intended to improve the accuracy of the screening process in child welfare in a way that proactively promotes equitable decision making.

LEARNING FROM OTHER JURISDICTIONS

The procedural basis for Oregon DHS's Safety at Screening tool comes from The Allegheny Family Screening Tool (AFST), developed by the Allegheny County Department of Human Services, Pennsylvania. Oregon DHS modified and expanded upon the AFST to comply with Oregon data systems, policy constraints, stakeholder recommendations, and its mission statement to promote an equitable service array.

MACHINE LEARNING – WHAT IS IT, AND WHAT IT IS NOT

The Safety at Screening Tool utilizes techniques from a field of computer science called m*achine learning*. The procedure involves using a computerized technique to discover how to associate Child Welfare administrative data elements with future outcomes of interest. These data elements all come from within Child Welfare administrative data (OR-KIDS), so no text or voice information is used.

By linking data elements (a.k.a., "features") regarding historical information to live information about an incoming report of abuse/neglect, it is possible to generate a prediction about whether the report will lead to a removal if the report is assigned to investigation, and/or whether screening out the report will lead to another future investigation. Our screening tool learns from hundreds of thousands of examples of reports. For each report, it considers factors like the allegations made, the number of children on the report, the count of past reports involving the child named on the report, etc. We inform the tool about whether the children named in the report went on to be removed from home and/or involved in a future investigation. The tool comes to learn how to use administrative data elements to calculate the probability that a child will be removed from home and/or involved in a future investigation.

After strict validation, this computerized process results in a tool that can be used to assess new screening reports, allowing Oregon DHS to make an informed estimation as to what may occur. Because Oregon DHS strives to understand and consider the limits of predictive analytics, this informed estimation is not treated as a definite fact about the future. Instead, the information is intended to guide decision-making for a case by providing information about how similar cases tended to proceed in the past.

ALGORITHMIC BIAS & STATISTICAL FAIRNESS

Since machine learning requires a historical data set from which to extract information, there is a risk of perpetuating bias which exists in the historical data. Oregon DHS made the decision to address this bias proactively rather than ignoring demographic information from the historical data. To accomplish this, Oregon DHS implements statistical procedures which were originally proposed in academic literature. Automated processes with less statistical bias can be considered as being more "fair." In order to measure and address bias and to promote fairness, it is critical to pick a concrete definition of fairness. This is addressed in the "Addressing Algorithmic Bias" section below.

STATISTICAL DEVELOPMENT

This section outlines the statistical and computational concepts involved in the development of a predictive analytics system for providing a data-informed tool for Child Welfare screeners.

MACHINE LEARNING CLASSIFIER

Machine learning classifiers are statistical algorithms which can learn to classify or sort observations into different types. For example, there are machine learning classifiers that can be trained to attempt to classify Twitter posts as A) "happy" or B) "sad." In the medical world, there are machine learning classifiers that can be trained to classify X-ray images of tumors as A) "benign" or B) "malignant." The classifiers achieve this ability by being shown thousands of examples, and being told whether each example was in fact an A or a B. These examples represent classifiers that can sort items into types, but classifiers can be used as prediction classifiers if they estimate whether something A) will happen, or B) will not happen in the future.

There are many different types of classifiers, but all of them require each observation to be broken down into "features" which can be represented by numbers. For example, a Twitter post's word count or number of exclamation points can be features, along with the language in which the post was written (e.g., English, Spanish, Japanese), and the time of day of the post. Certain things are more difficult or impossible to feed into the machine learning system, either because the information is not readily available (e.g., the true age of the Twitter poster) or because it is difficult to automatically determine the information (e.g., whether the Twitter post was sarcastic).

Given these limitations of machine learning, it is vital to carefully validate the predictions made by an automated classifier. In the same way students can be tested using a final exam based on new material, classifiers can be tested by inspecting the probabilities they generate for a separate sample of historical observations they have never come across before. To validate the predictive abilities of a classifier of tumor, one could look at a separate sample of X-ray images of tumors and consider the classifier's predicted probabilities of being malignant. For example, consider the two modeling results in the table below. Try to determine which risk tool would be more valuable to you as a physician?

Table 1. Which risk tool would be more valuable, Tool 1 or Tool 2?

	% of cases where malignancy is actually discovered on biopsy	
Malignancy Risk Prediction Tier	**Classifier #1**	**Classifier #2**
Extreme	14%	50%
High	12%	28%
Moderate	11%	12%
Low	10%	1%

While most would determine that Classifier #2 is the most useful system for classifying tumors, medical treatment plans cannot be distilled into a risk tier and involve a variety of nuanced factors. Even the best predictive model does not tell the whole story. We should expect expert decision makers to override the model from time-to-time and make a different conclusion about risk. Furthermore, the prediction may provide no practical guidance about the best course of treatment other than directing the physician's attention to the riskier situations. Predictive classifiers can be a useful part of a decision-making system but should never be the sole driving force behind critical decisions.

DUAL OUTCOMES (THE SELECTIVE LABEL PROBLEM)

Inspired by the AFST, the Oregon DHS Safety at Screening tool considers two future outcomes for each incoming report of child abuse/neglect. 1) If assigned for investigation, will the child be removed from the home within 2 years, and 2) If not assigned for investigation, will the child be involved in a future report which is then assigned for investigation. The two models were derived from two separate datasets: 1) reports that were assigned for investigation, and 2) reports that were not assigned for investigation. Separating the data in this way combats a pitfall in machine learning known as the Selective Label Problem. After a child's report is assigned for investigation, the child's experiences may differ substantially from a child whose report is not assigned for investigation. Because of this, the historical data set of assigned reports may vary substantially and in complex ways from the historical data set of reports that were not assigned for investigation. Thus, two separate models are devised to consider what may occur if a report is assigned versus not assigned.

FIXED HISTORY AND OUTCOME WINDOWS

The historical data set of reports of child abuse/neglect all come from OR-KIDS, Oregon's computerized child welfare data system, which came online in August 2011. After this date, the available data is organized and structured in a more uniform manner compared to the data available before this date. Thus, the Safety at Screening tool uses a classifier that was trained using data after this date.

Because of this historical cutoff date, reports have a different *volume* of usable history depending on the report date. For example, a report from February 2013 has the potential for 1.5 years of validated history in OR-KIDS, whereas a report from February 2017 has the potential for 5.5 years of validated history. This can be problematic for a variety of reasons, including the danger that the classifier will learn to focus on the volume of information available, rather than the details of that information. To overcome this problem of varying history, a fixed history window is defined at 1.5 years. In this way, all historical reports, regardless of when they occurred, have the same volume of historical data.

For similar reasons, a fixed outcome window is also critical. Children named in reports from 2014 have had 5 years to be involved in a future adverse outcome, whereas children named in reports from 2018 have only had 1 year to be involved in an adverse outcome. This leads to a problem where recent reports are deemed to have higher degrees of success (i.e., fewer adverse outcomes) simply because there has been less time for an adverse outcome to occur. To overcome this problem, a fixed outcome window of 2 years is defined.

To recap, a historical report was included as a viable observation if it occurred at least 1.5 years after the start of OR-KIDS, and 2 years prior to the date of the data pull. To keep the data volume consistent for all historical reports, 1.5 years of prior administrative data were used.

FEATURE ENGINEERING

As stated above, the machine learning classifier learns from information that can be coded as numerical values. The classifiers used in the Safety at Screening tool rely on administrative data elements from the following areas within DHS Child Welfare administrative data:

- Past Information
    - Report of abuse and neglect
    - CPS investigations
    - Involvement in Foster Care
- Current Information
    - Details of the report (allegations, number of children, etc.)
    - Alleged perpetrator
    - Reporter source
- Time Information (How far back in the past did prior events occur)
- The resulting 180+ variables are specified in the Technical Appendix.

Current information is merged with past information using real-time server data integration developed specifically for the Oregon DHS Safety at Screening tool in order to take into account live information from a report of child abuse/neglect. The Current Information refers to the information that a screening worker inputs into OR-KIDS while collecting information about the report of abuse/neglect, typically over the phone. Once the screening worker inputs the required information, the information is sent to a centralized server which collects and organizes

historical information. The predictive scores are generated in real time (the process takes approximately 9 seconds) and reported back to the screening worker within the OR-KIDS user interface.

PREVENTING INFORMATION LEAKAGE

TEMPORAL LEAKAGE

In order to train a predictive classifier, all information/features about a historical observation should be verified to have been available *at the time* of that historical event. If information that was collected *after* a historical event "leaks" into the information used to describe the event, then the predictive classifier can cheat by "peeking" at this future information. For example, consider that Twitter responses to a "happy" Twitter post may have more smiley faces, whereas responses to a "sad" Twitter post may have more frowny faces. These responses may have come minutes or days after the original Twitter post. If a classifier was allowed to peek at the responses, then it may learn to be highly accurate about whether a post is "happy" or "sad" simply by counting the smiley and frowny faces in the responses. However, if the classifier is intended to be used to determine whether a brand new Twitter post is "happy" or "sad", then it won't be able to rely on responses. Similarly, the Safety at Screening tool relies solely on information available at the time of the report (see Technical Appendix).

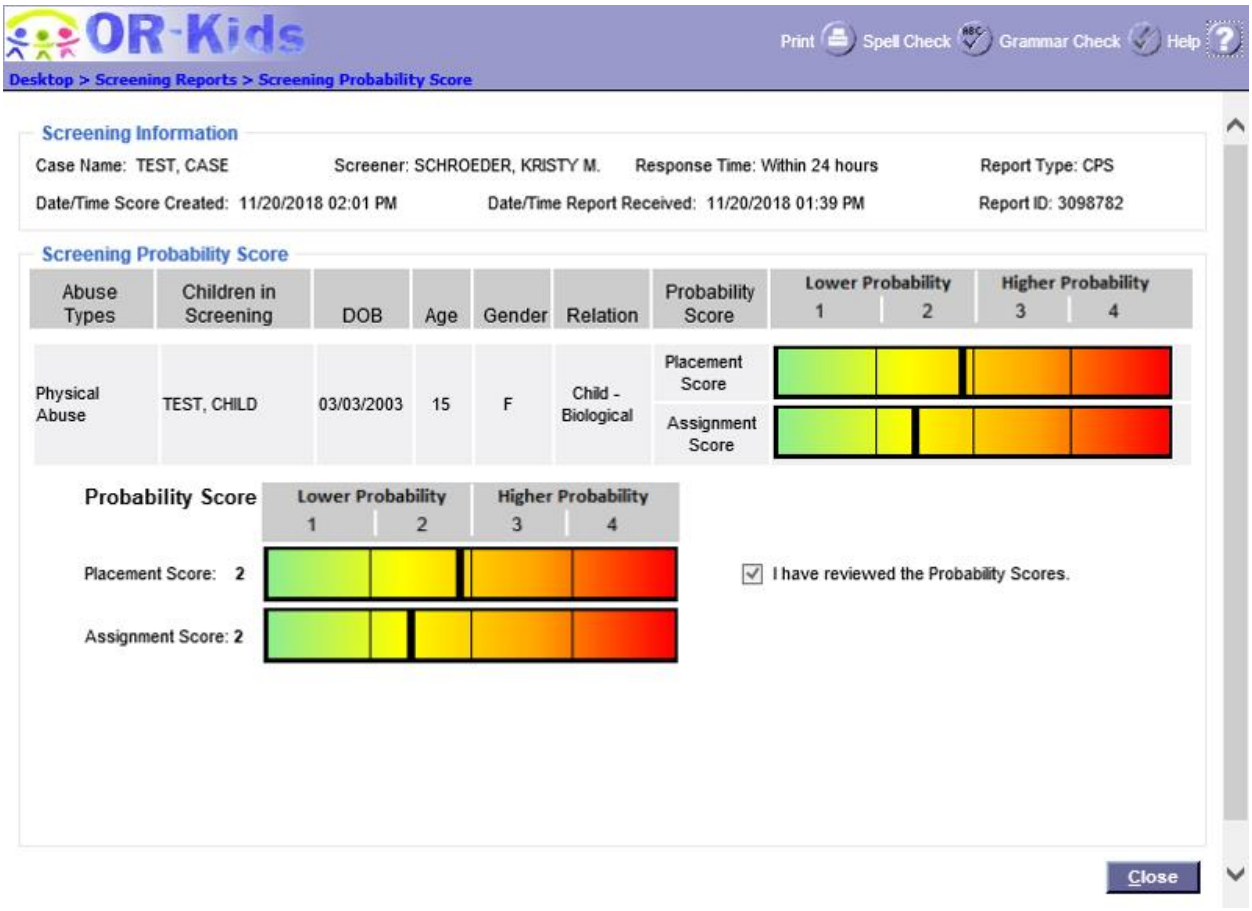
REPEATED OBSERVATION LEAKAGE

Over time, a child may be involved in multiple reports of child abuse/neglect. Also, a single report of child abuse/neglect may involve multiple children. This leads to a complex network of information which can inadvertently leak information to a predictive classifier. For example, if a child's 4th report (from 2016) is in the historical data set used to train a classifier, but a child's 2nd report (from 2013) is in the separate sample set used to assess predictive performance (see above), then the predictive abilities of the classifier will be exaggerated. Similarly, if Child 1 from a report is in the historical data set, but Child 3 from the same report is in the separate sample, then the predictive abilities of the classifier will also be exaggerated. These forms of exaggeration, or performance inflation, can come from the classifier relying too much on specific information about a child or a report, rather than from learning general rules that can apply to any future report. To prevent this problem, a strict resampling procedure is used in which many different unduplicated samples are chosen from the complete historical data set to prevent repeated observation leakage (see Technical Appendix).

DEVELOPMENT PHASE WORKGROUPS

During the research and development phase of the Oregon DHS Safety at Screening tool, several workgroups were convened in order to gain vital input from Oregon field staff, supervisors, administrators, and legal representatives. These workgroups provided key recommendations which shaped all aspects of the Safety at Screening tool (including feature engineering, usability of display screens, user access and permissions, etc.). Figure 1 depicts an example screenshot of the Safety at Screening display in OR-KIDS, for use by screening workers.

Figure 1. Example Display of the Safety at Screening Tool in OR-KIDS



CALCULATION AND DISPLAY OF THE SCORES

Two scores are generated for each child named in a screening report and who is not currently in substitute care. One score represents the probability that the child will be removed from home if the report is assigned for investigation, while the other score represents the probability that the child will be named in a new investigation if the report is screened out. Both scores are generated in similar ways. First, the machine learning classifier calculates the predicted probability (scale of 0% to 100%). Second, the probability is converted to a four-tier score system, resulting in a numerical score of 1 to 4. Scores of 1 and 2 represent lower than average predicted likelihood of occurrence, whereas scores 3 and 4 represent a higher than average likelihood. Scores of 1 represent the lowest 10th percentile of probabilities, while scores of 4 represent the highest 95th percentile of probabilities. Scores for each child named in the report are displayed, along with "Report Scores", which are the highest scores among all children in the report. Screening workers view these scores after submitting their screening information into OR-KIDS. By design, the process requires no additional data entry by the screening worker. After affirming that the scores have been reviewed, the screening worker continues to the decision phase of the work (i.e., determining whether to screen out the report or assign for investigation).

ADDRESSING ALGORITHMIC BIAS

PROXIES FOR DEMOGRAPHIC INFORMATION

To be frank, there is no realistic way to "remove" or "ignore" demographic information from a historical data set. The reason is that certain data elements that seemingly have no connection with demographic information can be *proxies* for demographic information. For example, if African American children tend to have more reports than other children, then the count of prior reports can be used as a *proxy* for determining whether a child is African American. Therefore, if the historical data set contains biases that are based on demographic information, then simply making a predictive classifier "blind" to demographic information will not prevent the perpetuation of bias. The position of Oregon DHS is to proactively address historical bias using statistical procedures drawn from academic literature.

DEFINITIONS OF FAIRNESS

The definition of the term "fair" can be difficult to pin down in English, and the same goes for mathematical definitions of fairness. In order to address fairness in machine learning, it is necessary to define fairness in a way that can be measured in terms of a number. There are many ways to do so, and the research group within Oregon DHS (ORRAI) engaged the wider academic community in order to select a single definition which encompasses the spirit of the Oregon DHS mission to provide an equitable service array. This definition is called *Error Rate Balance* and measures the way a predictive classifier makes *errors* between different race and ethnicity groups. In this case, an error is defined as labeling a child in a report as being at high risk when an adverse outcome does not ultimately occur, or labeling a child in a report as low risk when an adverse outcome does ultimately occur[1]. Oregon DHS found that the rate of these errors can differ, often drastically, between race and ethnicity groups. To combat this, the Safety at Screening tool uses a special procedure to balance these inequitable error rates.

TECHNIQUES TO INCREASE FAIRNESS

When it comes to reducing unfairness in a predictive classifier, there are three broad categories:

1. **Pre-processing strategies:** these strategies reduce unfairness by altering or adjusting the features which are sent into the machine learning classifier
2. **Processing strategies:** these strategies reduce unfairness by changing the way a machine learning classifier learns to generate predictions
3. **Post-processing strategies:** these strategies reduce unfairness by redefining how the final output of a predictive classifier is used to make decisions

In order to achieve more balanced error rates between race/ethnicity groups, the Safety at Screening tool employs a *post-processing strategy*. Post-processing strategies have several advantages over the other types, including: the ability to use any type of machine learning classifier, transparency about the mechanism by which reduced unfairness is achieved, and simplicity in being able to use unaltered features.

[1] The fact that error rate balance considers both observable outcome possibilities, rather than one or none, contributed to its selection over other statistical definitions of fairness, including predictive parity, predictive equality, equal opportunity and statistical parity.

POST-PROCESSING STRATEGY TO REDUCE ERROR RATE IMBALANCE

In order to improve the imbalance in predictive errors between race/ethnicity groups, the Oregon DHS Safety at Screening tool uses a post-processing threshold strategy. This strategy uses group-specific thresholds to define predictive risk tiers (i.e., the four-level score system described above). These group-specific thresholds were determined using an optimization technique which finds new thresholds that 1) lead to more balanced error rates between all groups, 2) maintain high levels of predictive performance, and 3) are close to the original thresholds. As predicted in academic literature, these new thresholds achieve large improvements in error rate balance despite a minimal tradeoff in predictive performance[2]. Oregon DHS has presented details of this procedure to various internal and external groups around Oregon and the United States. The details of this procedure, along with its demonstration on a freely available data set, are provided at https://www-auth.oregon.egov.com/DHS/ORRAI/Documents/Fairness-Machine-Learning-Generated-Risk-Scores-Equitable-Thresholding.pdf.

PSYCHOLOGICAL / HUMAN FACTORS CONSIDERATIONS

Automation Bias is a psychological phenomenon that refers to unintended behaviors that may arise in people who use predictive tools as part of a decision-making process. While automation bias can take a variety of forms, the behaviors most concerning for a Safety at Screening tool are:

- Inferring humanlike abilities that are beyond the actual capability of a predictive classifier
- Feeling pressure to use predictive risk scores despite clear contradictory evidence
- Manipulating data input in order to achieve a desired predictive risk score
- Imparting one's own risk threshold depending on the case details (consciously or unconsciously)
- Reducing information gathering at screening due to an over-reliance on predictive risk scoring

These aspects of automation bias can be combatted using a variety of strategies which include:

- **Training:** Prospective users of the Safety at Screening Tool receive training which includes detailed information regarding what is and is not included as data sources and features for the predictive model
- **Reducing Complexity:** Providing fine grained scores (e.g., 1 to 100) can allow for users to impart their own group-specific thresholds (e.g., a worker who requires a score greater than 40 for female children, and greater than 50 for male children), which introduces new sources of bias into the system. To combat this, a four-tiered score system is introduced.
- **Framing Predictions as Supportive.** The predictive scores are not framed as "answers" or "directions", but rather historical indicators derived from historical administrative data.
- **Timing of Information.** To avoid *priming* (i.e., influencing decision making), the predictive scores are shown after the data gathering and data entry phase are complete for a screening report.
- **Tracking Data Omission / Commission.** Process evaluations will involve tracking of data elements to ensure data entry behavior remains consistent with historical expectations.

RESULTS

[2]For example, the accuracy of the Removal Model decreased from 81% to 79% as a result of the procedure. However, the error rate balance, which is a numerical value between 0 and 1 (0 implies completely unfair and 1 implies perfectly fair), increased from 0.43 to 0.76 as a result of the procedure. The details surrounding the computation of the error rate balance measure are provided in the technical appendix.

Each child that is a part of a report of abuse/neglect can receive predictive scores. Children who are currently in substitute care (e.g., foster care) do not receive predictive scores, as these cases are subject to different investigative protocol and scrutiny (representing 4% of incoming reports). Two scores are given for each child: 1) If assigned for investigation, what is the probability the child will be removed from the home within 2 years, and 2) If not assigned for investigation, what is the probability the child will be involved in a future report which is then assigned for investigation. The probabilities are subjected to the fairness correction procedure which assigns one of four probability tiers. A score of 1 represents the lowest probability tier, while a score of 4 represents the highest probability tier. Scores of 1 and 2 predict the outcome will occur less often than average, whereas scores of 3 and 4 predict the outcome to occur more often than average.

In order to assess the usefulness of the predictive classifier, the children in historical reports were given probability scores. These scores are based on the information that would have been available at the time of the score. To prevent information leakage, scores were only provided using predictive classifiers that were not built using the same child/report of interest.

For each score level, it is possible to then check what *actually* occurred in the following 2 years. This exercise gives a sense of the usefulness of the predictive classifiers.

Table 2. Predicted Score Levels Compared with Actual Occurences
A) Removal from Home, B) Future Investigation

Score	Given Score	Assigned for Investigation	Removed from Home	Removed from Home if Assigned	Removed from Home if Screened Out
4	3%	67%	55%	62%	40%
3	21%	55%	16%	20%	11%
2	61%	47%	3%	4%	2%
1	15%	34%	1%	1%	0.4%
Overall	*144166 Children*	47%	7%	10%	4%

Score	Given Score	Assigned for Investigation	Future Investigation	Future Investigation if Assigned	Future Investigation if Screened Out
4	4%	53%	68%	62%	75%
3	40%	54%	41%	41%	41%
2	47%	45%	18%	20%	17%
1	8%	27%	5%	7%	4%
Overall	*144166 Children*	47%	28%	31%	26%

Table 2 breaks down the scores that would have been given for an unduplicated set of historical child/report observations. Take for example Table 2A, which represents predictions for whether a child would have been removed from home following a report of abuse/neglect. The “Score of 1” row conveys that 15% of children would have received this score (21,625 children) and that 34% of these reports would have become open investigations

(7,353 investigations). However, only 1% of these children were *actually* removed from their home (216). These discrepancies between predicted outcomes and actual open investigations represents the core potential for using data informed predictions at the time of a screening decision.

TECHNICAL APPENDIX

RESAMPLING PROCEDURE

A strict resampling procedure is used in which many different unduplicated samples are chosen from the complete historical data set to prevent repeated observation leakage. The following pseudo-code demonstrates this process. In essence, classifiers are trained on an unduplicated subsample of the full data set until all observations have appeared in at least one test/validation set. Thus, the result is a matrix containing at least one untainted predicted probability for each observation.

```
Input:
       DataSet = {(R1, y1, P1), (R2, y2, P2),...,(Rm, ym, Pm)}

Where:
       Rm = {IDKeym, IDChildm, IDReportm, Xm}
       IDKeym = Unique primary key
       IDChildm = Individual child identifier
       IDReportm = Individual report identifier
       Xm = Feature vector constructed from temporally available information
       ym = {0: Did not have adverse outcome, 1: Had adverse outcome}
       Pmi = ∅ #Growing vector of untainted predicted probabilities

i := 0

Repeat:
       i := i + 1

       TrainSeti := Sample(DataSet) s.t. all IDChild & IDReport unique

       ValidationSeti := DataSet excluding (IDKey or IDChild or IDReport) ∈ TrainSeti

       Train Classifier Hi: X → y, using TrainSeti

       For each Xm in DataSet:
              If Xm ∈ ValidationSeti:
                     Pmi := Predicted probability from Predict(Hi, Xm)
              Else:
                     Pmi := ∅

Until: All Pm have at least one predicted probability

Predictionsm := Average predicted probability for each Rm DataSet

Return Predictions
```

ERROR RATE BALANCE CALCULATION AND RESULTS

To calculate the error rate balance, the 6 pairwise ratios of false negative rates across the four protected attribute levels, along with the 6 pairwise ratios of false positive rates, are computed. These pairwise ratios are calculated such that the larger FNR (or FPR accordingly) between the two levels is always in the denominator; this will ensure

that all pairwise ratios are between 0 and 1. The error rate balance is then the smallest of these 12 pairwise ratios since the smallest value indicates the greatest disparity. If the error rate balance equals 1, it means all four levels of the protected attribute have identical FNRs and identical FPRs. The closer to 0 the error rate balance gets, the greater the disparity in error rates across the levels of the protected attribute. Tables 3 and 4 report these results.

Table 3. False Negative Rates (FNR) and False Positive Rates (FPR) for all Levels of the Protected Attribute, Before and After the Fairness Procedure

	Baseline		Post-Procedure	
PA Level	FNR	FPR	FNR	FPR
BL	0.36	0.15	0.29	0.21
HS	0.29	0.19	0.27	0.21
NV	0.24	0.34	0.30	0.27
WA	0.34	0.16	0.30	0.20
Overall	0.33	0.17	0.29	0.21

Table 4. Pairwise FNR and FPR Ratios Between Levels of the Protected Attributes, Before and After the Fairenss Procedure (Error Rate Balance). Overall Error Rate Balance is Defined as the Most Disparate Ratio (Closest to 0; Highlighted in Black).

Baseline

FNR

PA Level	BL	HS	NV	WA
BL	1	0.81	0.67	0.97
HS		1	0.83	0.83
NV			1	0.69
WA				1

FPR

PA Level	BL	HS	NV	WA
BL	1	0.78	0.43	0.9
HS		1	0.55	0.86
NV			1	0.47
WA				1

Post-Procedure

FNR

PA Level	BL	HS	NV	WA
BL	1	0.92	0.95	0.98
HS		1	0.88	0.9
NV			1	0.97
WA				1

FPR

PA Level	BL	HS	NV	WA
BL	1	0.99	0.78	0.98
HS		1	0.78	0.97
NV			1	0.76
WA				1

PREDICTIVE PERFORMANCE RESULTS TABLE BEFORE FAIRNESS PROCEDURE

Table 5. Predicted Score Levels Compared with Actual Occurences (Before Fairness Procedure)
A) Removal from Home, B) Future Investigation

Score	Given Score	Assigned for Investigation	Removed from Home	Removed from Home if Assigned	Removed from Home if Screened Out
4	3%	67%	56%	63%	41%
3	18%	55%	17%	21%	12%
2	63%	47%	4%	5%	2%
1	16%	35%	1%	1%	0%
Overall	*144166 Children*	47%	7%	10%	4%

Score	Given Score	Assigned for Investigation	Future Investigation	Future Investigation if Assigned	Future Investigation if Screened Out
4	4%	52%	69%	63%	75%
3	35%	54%	43%	43%	44%
2	52%	46%	19%	21%	18%
1	9%	27%	5%	7%	4%
Overall	*144166 Children*	47%	28%	31%	26%

VARIABLE LIST

VARIABLE	Explanation
Threat_Of_Harm_Alleged	Allegation type
Mental_Injury_Alleged	Allegation type
Neglect_Alleged	Allegation type
Physical_Abuse_Alleged	Allegation type
Sexual_Abuse_Alleged	Allegation type
Medical_Neglect_Flg	Allegation type
Reporter_Source	Reporter Category
Age	Age of the Alleged Victim
Age_0_to_3	Age Category of the Alleged Victim
Age_3_to_6	Age Category of the Alleged Victim
Age_6_to_9	Age Category of the Alleged Victim
Age_9_to_12	Age Category of the Alleged Victim
Age_12_to_18	Age Category of the Alleged Victim

Age_18gr	Age Category of the Alleged Victim
RepSource_Alleged_Perp	Reporter Category
RepSource_Anonymous	Reporter Category
RepSource_CommunitySchoolProf	Reporter Category
RepSource_FamilyExtendedFriend	Reporter Category
RepSource_FamilyHousehold	Reporter Category
RepSource_GovtRep	Reporter Category
RepSource_LegalRep	Reporter Category
RepSource_MedPsychProf	Reporter Category
RepSource_Other	Reporter Category
RepSource_Police	Reporter Category
RepSource_Self	Reporter Category
Mandatory_Reporter	Whether the Reporter is a Mandatory Reporter
Gender_Male	Gender of Alleged Victim is Male
RACECAT_NV	Race/Ethnicity Category of Alleged Victim
RACECAT_HS	Race/Ethnicity Category of Alleged Victim
RACECAT_BL	Race/Ethnicity Category of Alleged Victim
RACECAT_WA	Race/Ethnicity Category of Alleged Victim
RACECAT_UN	Race/Ethnicity Category of Alleged Victim
RPT_nPRSN	Number of People Named on the Report
RPT_nPERP	Number of Perps Named on the Report
RPT_nVICT	Number of Victims Named on the Report
RPT_nHHMB	Number of Household Members Named on the Report
RPT_nNHMB	Number of Non-Household Members Named on the Report
RPT_nOTHC	Number of Other Children Named on the Report
RPT_nPRCG	Number of Parents or Caregivers Named on the Report
RPT_nVICT_0_3	Victim Count by Age Group
RPT_nVICT_3_6	Victim Count by Age Group
RPT_nVICT_6_9	Victim Count by Age Group
RPT_nVICT_9_12	Victim Count by Age Group
RPT_nVICT_12_18	Victim Count by Age Group
RPT_nPERP_00_13	Perp Count by Age Group
RPT_nPERP_14_24	Perp Count by Age Group
RPT_nPERP_25_34	Perp Count by Age Group
RPT_nPERP_35_44	Perp Count by Age Group
RPT_nPERP_55_54	Perp Count by Age Group
RPT_nPERP_65_up	Perp Count by Age Group
RPT_nPERP_AgeNA	Perp Count by Age Group
RPT_nPERP_HHMB	Perp Count by Role
RPT_nPERP_NHMB	Perp Count by Role

RPT_nPERP_PRCG	Perp Count by Role
VICT_ROLE_HHMB	Role of the Alleged Victim
VICT_ROLE_NHMB	Role of the Alleged Victim
VICT_ROLE_OTHC	Role of the Alleged Victim
VICT_ROLE_PRCG	Role of the Alleged Victim
In_IH_Now	Whether the Child is currently receiving In-Home services
PERP_RPT_n_NDays	Number of Reports linked to Perp in Last {90,180,365,548} Days
PERP_RPT_PERP_n_NDays	Number of Reports linked to Perp in Last {90,180,365,548} Days (by Role of the Perp)
PERP_RPT_VICT_n_NDays	Number of Reports linked to Perp in Last {90,180,365,548} Days (by Role of the Perp)
PERP_RPT_RPTR_n_NDays	Number of Reports linked to Perp in Last {90,180,365,548} Days (by Role of the Perp)
PERP_RPT_CSNM_n_NDays	Number of Reports linked to Perp in Last {90,180,365,548} Days (by Role of the Perp)
PERP_RPT_HHMB_n_NDays	Number of Reports linked to Perp in Last {90,180,365,548} Days (by Role of the Perp)
PERP_RPT_NHMB_n_NDays	Number of Reports linked to Perp in Last {90,180,365,548} Days (by Role of the Perp)
PERP_RPT_OTHC_n_NDays	Number of Reports linked to Perp in Last {90,180,365,548} Days (by Role of the Perp)
NSP_n_NDays	Non-Placement Services in Last {90,180,365,548} Days
NSP_n_Ongoing_NDays	Non-Placement Services in Last {90,180,365,548} Days (Ongoing Services)
NSP_n_OneTime_NDays	Non-Placement Services in Last {90,180,365,548} Days (One-Time Services)
SP_n_NDays	Number of Service Placements in Last {90,180,365,548} Days
SP_n_BRS_NDays	Number of Service Placements in Last {90,180,365,548} Days (by Type)
SP_n_InHome_NDays	Number of Service Placements in Last {90,180,365,548} Days (by Type)
SP_n_DD_Foster_NDays	Number of Service Placements in Last {90,180,365,548} Days (by Type)
SP_n_Family_Shelter_NDays	Number of Service Placements in Last {90,180,365,548} Days (by Type)
SP_n_Foster_Care_NDays	Number of Service Placements in Last {90,180,365,548} Days (by Type)
SP_n_Hospitalization_NDays	Number of Service Placements in Last {90,180,365,548} Days (by Type)
SP_n_JJ_Custody_NDays	Number of Service Placements in Last {90,180,365,548} Days (by Type)
SP_n_NonBRS_Shelter_Emerg_NDays	Number of Service Placements in Last {90,180,365,548} Days (by Type)
SP_n_Res_Treatment_DD_Group_NDays	Number of Service Placements in Last {90,180,365,548} Days (by Type)
SP_n_Runaway_CWP_NDays	Number of Service Placements in Last {90,180,365,548} Days (by Type)
SP_n_THR_Indep_PreAdopt_NDays	Number of Service Placements in Last {90,180,365,548} Days (by Type)
SP_n_Unk_AltProg_NDays	Number of Service Placements in Last {90,180,365,548} Days (by Type)

RPT_n_Assigned_NDays	Number of Assigned Reports in Last {90,180,365,548} Days
RPT_n_NDays	Number of Reports in Last {90,180,365,548} Days
RPT_n_Assigned_TOH_NDays	Number of Assigned Reports in Last {90,180,365,548} Days (by Allegation Type)
RPT_n_Assigned_MentalInjury_NDays	Number of Assigned Reports in Last {90,180,365,548} Days (by Allegation Type)
RPT_n_Assigned_Neglect_NDays	Number of Assigned Reports in Last {90,180,365,548} Days (by Allegation Type)
RPT_n_Assigned_PhysAbuse_NDays	Number of Assigned Reports in Last {90,180,365,548} Days (by Allegation Type)
RPT_n_Assigned_SexAbuse_NDays	Number of Assigned Reports in Last {90,180,365,548} Days (by Allegation Type)
RPT_n_Assigned_MedNeglect_NDays	Number of Assigned Reports in Last {90,180,365,548} Days (by Allegation Type)
RPT_n_CAS_TOH_NDays	Number of Closed Reports in Last {90,180,365,548} Days (by Allegation Type)
RPT_n_CAS_MentalInjury_NDays	Number of Closed Reports in Last {90,180,365,548} Days (by Allegation Type)
RPT_n_CAS_Neglect_NDays	Number of Closed Reports in Last {90,180,365,548} Days (by Allegation Type)
RPT_n_CAS_PhysAbuse_NDays	Number of Closed Reports in Last {90,180,365,548} Days (by Allegation Type)
RPT_n_CAS_SexAbuse_NDays	Number of Closed Reports in Last {90,180,365,548} Days (by Allegation Type)
RPT_n_CAS_MedNeglect_NDays	Number of Closed Reports in Last {90,180,365,548} Days (by Allegation Type)
INV_n_NDays	Number of Investigations in Last {90,180,365,548} Days
INV_n_Safety_Decision_NDays	Number of Investigations in Last {90,180,365,548} Days (With Decision of Safe)
INV_n_Founded_Physical_Abuse_NDays	Number of Investigations in Last {90,180,365,548} Days (by Allegation Type)
INV_n_Founded_Sexual_Abuse_NDays	Number of Investigations in Last {90,180,365,548} Days (by Allegation Type)
INV_n_Founded_Neglect_NDays	Number of Investigations in Last {90,180,365,548} Days (by Allegation Type)
INV_n_Founded_Threat_Of_Harm_NDays	Number of Investigations in Last {90,180,365,548} Days (by Allegation Type)
INV_n_Founded_Mental_Injury_NDays	Number of Investigations in Last {90,180,365,548} Days (by Allegation Type)