A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions

Alexandra Chouldechova ACHOULD@CMU.EDU
Heinz College
Carnegie Mellon University
Pittsburgh, PA, 15213, USA

Emily Putnam-Hornstein EHORNSTE@USC.EDU
Suzanne Dworak-Peck School of Social Work
University of Southern California
Los Angeles, CA, 90089, USA

Diana Benavides-Prado DIANA.BENAVIDES.PRADO@AUT.AC.NZ
Oleksandr Fialko OLEKSANDR.FIALKO@AUT.AC.NZ
Rhema Vaithianathan RHEMA.VAITHIANATHAN@AUT.AC.NZ
Centre for Social Data Analytics
Auckland University of Technology
Auckland, New Zealand

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

Every year there are more than 3.6 million referrals made to child protection agencies across the US. The practice of screening calls is left to each jurisdiction to follow local practices and policies, potentially leading to large variation in the way in which referrals are treated across the country. Whilst increasing access to linked administrative data is available, it is difficult for welfare workers to make systematic use of historical information about all the children and adults on a single referral call. Risk prediction models that use routinely collected administrative data can help call workers to better identify cases that are likely to result in adverse outcomes. However, the use of predictive analytics in the area of child welfare is contentious. There is a possibility that some communities—such as those in poverty or from particular racial and ethnic groups—will be disadvantaged by the reliance on government administrative data. On the other hand, these analytics tools can augment or replace human judgments, which themselves are biased and imperfect. In this paper we describe our work on developing, validating, fairness auditing, and deploying a risk prediction model in Allegheny County, PA, USA. We discuss the results of our analysis to-date, and also highlight key problems and data bias issues that present challenges for model evaluation and deployment.

1. Introduction

Every year there are more than 3.6 million referrals made to child protection agencies across the US. It is estimated that 37% of US children are investigated for child abuse and neglect by age 18 years (Kim et al., 2017). These statistics indicate that far from being a rare occurrence, many more children are being pulled into the child welfare agencies than previously thought. Currently, screening these referral calls is left to each jurisdiction to follow local practices and policies. These practices usually involve caseworkers gathering details about the adults and children associated with the alleged victim. Often, the decision on whether to investigate or not is made without ever visiting the family or speaking with them.

Whilst electronic case management systems and linked administrative data are increasingly available, it is difficult for child welfare workers to make systematic use of historical information about all the children and adults on a single refer-

ral call. Fully interrogating a case history for all the people named on a call could take hours. The increasing pressure on ensuring that investigative resources are focused on the highest risk children means that there is some potential for predictive analytics to assist call screeners to more quickly and accurately assess each referral.

Predictive Risk Modelling (PRM) uses routinely collected administrative data to predict the likelihood of future adverse outcomes. By strategically targeting services to the riskiest cases, it is hoped that many of the adverse events can be prevented. Additionally, unnecessary investigations which are burdensome for families and costly for the system could be avoided. PRM has been used previously in health and hospital settings (Panattoni et al., 2011; Billings et al., 2012) and has been suggested as a potentially useful tool that could be translated into child protection settings (Vaithianathan et al., 2013).

However, the use of predictive analytics in the area of child welfare is contentious. There is the possibility that some communities—such as those in poverty or from particular racial or ethnic groups—will be disadvantaged by the reliance on government administrative data because they will typically have more data kept about them simply by dint of being poor and on welfare. Such families could then be flagged as high risk and be more frequently investigated. If the algorithm uses past investigations to produce a high risk score for a family, then this will exacerbate the original bias.

On the other hand, these analytics tools can augment or replace human judgments, which themselves are potentially biased. There is a possibility that caseworkers are basing their screening decisions in part on personal experiences or current caseloads. Caseworker decisions may also be be affected by cognitive biases—for example over-weighting recent cases, unrelated to the current case, where a child has been fatally harmed. When making decisions under time pressure, caseworkers might be guilty of statistical discrimination, where they use easily observed features (e.g. living in a neighborhood with high crime rates) as proxies for unobservable but more pertinent attributes (e.g. drug-use). Bias in human decision-making is often difficult to assess, and the existing research does not provide a consensus view of racial bias in the child welfare system (Fluke et al., 2011).

A subject where there is greater consensus concerns the relative accuracy of so-called "clinical" versus "actuarial" judgment.[1] Some of the earliest research comparing human predictions to those of statistical models goes back to the pioneering work of Meehl (1954). Decades of research and several large scale meta-analyses have largely upheld the original conclusions: When it comes to prediction tasks, statistical models are generally significantly more accurate than human experts (Dawes et al., 1989; Grove et al., 2000; Kleinberg et al., 2017).

Our goal in Allegheny County is to improve both the accuracy and equity of screening decisions by taking a fairness-aware approach to incorporating prediction models into the decision-making pipeline. The present paper reports on the lessons that we have learned so far, our approaches to predictive bias assessment, and several outstanding challenges. To be sure, at certain points we offer more questions than answers. Our hope is that this report contributes to the rich ongoing conversation concerning the use of algorithms in supporting critical decisions in government—and the importance of considering fairness and discrimination in data-driven decision making. While the work presented here is firmly grounded in the child maltreatment hotline context, much of the discussion and our general analytic approach are broadly applicable to other domains where predictive risk modeling may be used. Readers are also encouraged to refer to the recent work of Shroff (2017) for another perspective on predictive analytics in the child welfare domain. This related work provides an excellent report on various considerations that are important for model development, as well as strategies for effectively engaging with agency leadership.

1.1. Organization of this paper

We begin in the next section with some background on the model development. As part of this discussion we describe both the tool that is currently deployed in Allegheny County and the competing models that are being developed

1. The term "actuarial" has fallen out of fashion, and has in many cases been replaced with "machine learning".

as part of an ongoing redesign process. Then in Section 3 we investigate the predictive bias properties of the current tool and a Random forest model that has emerged in the redesign as one of the best performing competing models. Our predictive bias assessment is motivated both by considerations of human bias and recent work on fairness criteria that has emerged in the algorithmic fairness literature. Section 4 discusses some of the challenges in incorporating algorithms into human decision making processes and reflects on the predictive bias analysis in the context of how the model is actually being used. We discuss some of the concerns that have arisen as part of the redesign, and propose an "oracle test" as a tool for clarifying whether particular concerns pertain to the statistical properties of a model or are targeted at other potential deficiencies. We also briefly describe an independent ethical review of the modeling work that was commissioned by the County. Section 5 concludes with a reflection on several of the key outstanding technical challenges affecting the evaluation and implementation of the model.

2. The Allegheny Models

Allegheny County is a medium size county in Pennsylvania, USA, centered on the city of Pittsburgh. In 2014, Allegheny County's Department of Human Services issued a Request for Proposals focused on the development and implementation of tools that would enhance use of the County's integrated data system. Specifically, the County sought proposals that would: (1) improve the ability to make efficient and consistent data-driven service decisions based on County records, (2) ensure public sector resources were being equitably directed to the County's most vulnerable clients, and (3) promote improvements in the overall health, safety and well-being of County residents. A consortium of researchers from four universities were awarded the contract in the Fall of 2014 and commenced work in close concert with the Allegheny County team.

In mid-2015, it was decided that the most promising, ethical, and readily implemented use of PRM within the Allegheny County child protection context was one in which a model would be deployed at the time a referral call was was received by the County. The objective was to help call workers determine whether a maltreatment referral is of sufficient concern to warrant an in-person investigation (referred to as "screening-in the call"). Calls that are thought to be innocuous and are not further investigated are said to be "screened-out". In this section we describe the development and implementation of the Allegheny Family Screening Tool (AFST).

2.1. Then and now

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

Prior to the implementation of the call screening tool, for the period from April 1, 2010 through May 4, 2016, the majority of CPS reports (52%) were screened out. Of those children who were screened out, 53% were re-referred for a new allegation within 2 years. Of those who were initially screened-in, 13% were placed outside of the home within 2 years.

Retrospective analysis of cases that resulted in critical incidents has revealed many instances of multiple calls having been repeatedly screened out. While each screen-out decision was individually defensible, looking at the full case history

paints the picture that child protective services should have gotten involved. The primary aim of introducing a prediction model is to supplement the often limited information received during the call with a risk assessment that takes into account a broader set of information available in the integrated data system.

The implementation of the AFST presents call screening staff with a score from 1 to 20 that reflects the likelihood that the child will be placed (removed from home) within 2 years conditional on being screened in (further investigated). Out of home placement was chosen as the target for two primary reasons. First, it is a directly observable event that serves as a good proxy for severe child maltreatment. Second, placement decisions are not made by the call screening staff. By predicting an outcome that cannot be directly determined by the staff, we reduce the risk of getting trapped in a feedback loop wherein workers effect the outcome predicted by the model (e.g., substantiate cases that the model tells them they likely should).

The original implementation also included a score reflecting the likelihood of re-referral conditional on the case being screened out. This model never gained traction in practice largely due to lack of buy-in from County leadership. There are many reasons for why a case might be re-referred, and many of them do not merit protective services involvement. A County review of how cases were scored by the re-referral model indicated that re-referral risk was a poor measure of whether a referral merits further investigation.

Even if re-referral risk were deemed to be a reasonable measure, there would be cause to doubt the predictive validity of the model. Effective December 31, 2014, the state of Pennsylvania amended the Child Protective Services Law (CPSL) to broaden the category of mandated reporters. This significantly changed reporting patterns across the state, resulting in an increase in the number of referrals. Since the initial training data predated the policy change, the re-referral would not be able to take this surge in referrals into account.

There is of course good reason to suspect that the predictive performance of the placement model will also be negatively affected by the mandatory reporting amendment. On the other hand, one might also expect that the placement model would not be as severely compromised as the re-referral model. That is, while more referrals are coming in, there may be considerable consistency in the patterns of risk factors associated with placement risk. We will be better positioned to reevaluate the performance of the model as more post-amendment data becomes available for retrospective analysis.

These considerations are relevant to understanding how the AFST was deployed, and how it is intended to be used in the call screening context. While in some settings machine learning systems have been used to replace decisions that were previously made by humans, this is not the case for the Allegheny Family Screening Tool. It was never intended or suggested that the algorithm would replace human decision-making. Rather, that the model should help to inform, train and improve the decisions made by the staff.

As we have previously noted, the AFST and the call worker are relying on very different information in assessing referrals. Whereas the call workers' screening decisions are based in large part on the content of the allegation, the AFST does not use information about the allegation per se. The AFST instead relies entirely on administrative data available on the individuals associated to the referral. So while the AFST is able to pick up on cases that have high long arc risk, it may just as easily miss acute incidents described in the allegation that necessitate immediate investigation. Until such a time that an AI system is developed that can properly assess all of the risks relevant to screening decisions, full automation of the call screening process remains out of the question. In the meantime, the AFST continues to be tested as a form of decision support, as well as a way to help leadership get better insight into call screening decisions which historically have been relatively opaque.

2.2. Modeling methodology

2.2.1. Data

The full data set consists of all $n = 76,964$ referral records collected by Allegheny County between April 2010 and July 2014. A distinct referral record is generated for each child associated to an allegation. These records correspond to $36,840$ distinct referrals, involving $47,305$ distinct children. In total the data set contains over

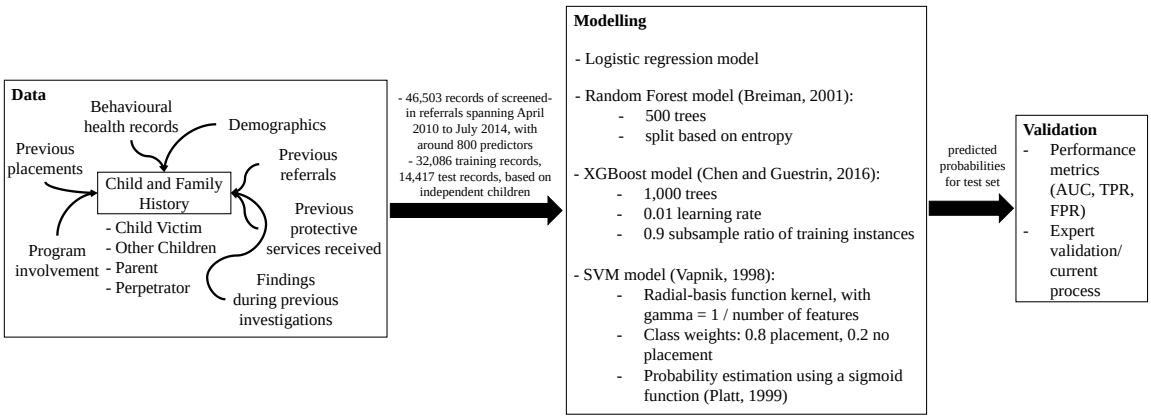

Figure 1: An overview of the modeling process.

800 variables providing demographics, past welfare interaction, public welfare, county prison, juvenile probation, and behavioral health information on all persons associated with each referral. 19,869 of the observed referrals, involving 31,438 distinct children were screened in for investigation. This corresponds to 46,503 referral records. All of the model training and evaluation presented in this report is conducted on the subset of screened-in cases. Since our models are trained and tested on the screened-in population, we encounter a version of the *selective labels problem* (Kleinberg et al., 2017) in trying to generalize our evaluation results to the entire set of referrals. We discuss this further in Section 5.

2.2.2. Current model

The initial model—the one that has been deployed in Allegheny County since August 2016—is based on a logistic regression fit to a selected subset of 71 features. Instead of presenting call workers with raw probability estimates or coarse classifications, it was decided that a derived score from 1 to 20 would be more suitable.[2] The derived score corresponds to the ventiles (5% percentiles) of the estimated probability distribution. In the initial implementation, cases whose risk is assessed in the most risky 15% of the calls (scores of 18 or higher) were flagged for call screeners as "mandatory screen-ins".[3] The top panel of Figure 2 shows the ventile and quantile cutoffs along with red dashed lines representing the mandatory screen-in cutpoints for the logistic regression model. Owing to how the scores are constructed, two cases that both receive a score of 19 may have a greater absolute risk difference than two cases that receive scores of 6 and 10, respectively.

2.2.3. Issues with model validation

As we only recently discovered, two mistakes were made in evaluating the predictive performance of the original version of the AFST—the version that went into deployment in 2016. These errors meant that reported estimates of the predictive performance of the AFST models (e.g., AUCs) were over-optimistic. To be clear, uncovering these issues earlier would not have affected how or whether the AFST was deployed. Correcting the validation scheme has produced more realistic estimates of model performance under ideal conditions, but does not change the models themselves. We describe the issues briefly here, and outline a corrected validation scheme in the section that follows.

The first problem comes from the way in which the train-test split was initially performed. The split was obtained by randomly holding out 30% of the referral *records* for model validation, re-

2. This was not a principled decision, and is being reconsidered in the model redesign.

3. Despite what the name suggests, "mandatory" screen-ins can be overridden by hotline supervisors. We discuss overrides in Section 5.

taining the remaining 70% of records for model training. This has two potential issues: (1) the same children appeared both in the training and test set (but for different referrals); and (2) referrals involving multiple children had their records split between the training and test set (so that, say, siblings on the same referral could have been allocated between the train and test sets). Issue (2) is the primary cause of over-optimism in estimating predictive performance. This is because placement outcomes, although not universally the same for all children associated to a referral, are nonetheless highly correlated.

Secondly, the 71 features used in the final placement model were selected by looking at regression coefficient t-statistics[4] using the full set of data. The train-test split was performed only after the set of features had been selected, so even if the split itself were valid, the model evaluation would not have accounted for the variable selection step. This further contributed to over-optimism in the prediction performance of the original AFST model.

The validation results presented in the remainder of this paper use a corrected train-test split, and do not suffer from these issues. For comparison purposes throughout, we use a copycat model of the AFST ("AFST copy") obtained by carrying out the variable selection procedure using only the data in the train split.

Method	**AUC**	**TPR**	**FPR**
Logistic (all vars)	0.70	0.49	0.21
AFST copy	0.74	0.54	0.20
SVM	0.77	0.57	0.20
Random Forest	0.77	0.58	0.20
XGBoost	0.80	0.61	0.19

Table 1: Performance results for methods under analysis. TPR and FPR correspond to the 25% highest risk cutoff (ventile scores of 16 and higher).

2.2.4. Model rebuild.

In April 2017 the research team began a rebuild of the AFST with the aim of improving model accuracy by using the full set of available features and applying more flexible machine learning models. Figure 1 provides a summary of the data, model options and validation metrics used during the rebuild.

As part of this rebuild, we also revised the train-test splitting approach. Sample splitting was performed to ensure that there were no overlapping children or referrals across the training and test data. 32,086 of the referral records were used in training the models, and 14,417 referral records were held out for validation purposes. More details on this splitting approach along with validation results from holding out the most recent year of data as a test set are provided in the Supplement.

The machine learning methods we considered included support vector machines (Vapnik, 1998) with probability prediction (Platt et al., 1999), random forests (Breiman, 2001) and XGBoost (Chen and Guestrin, 2016). We used available implementations of these methods in R, Python and LibSVM (Chang and Lin, 2011). These competing models produced considerable gains over the logistic regression model. Table 1 shows several test set classification metrics for the different methods.

In the redesign it was also decided that the mandatory screen-in threshold will be lowered to capture the top 25% highest cases (cases scoring 16 or higher). This lower threshold corresponds to a TPR (Recall/Sensitivity) of 58% for the random forest model and 61% for the XGBoost model. That is, of the test set cases that were observed to result in placement within 2 years of the call, around 3 in 5 are flagged as mandatory screen ins by the two best performing models.

3. Predictive bias

Predictive bias is a major concern when deploying predictive modeling in the child welfare context. In this section we reflect on our analysis to-date of the racial bias[5] properties of the Allegheny County models. To set the stage for our discussion we begin with an overview of how hu-

4. For more details, see p.13 of https://tinyurl.com/y8n5m9kg

5. We performed similar analyses looking instead at poverty status and sex. There was no strong indication of predictive bias with respect to these other variables.

man bias is thought to contribute to observed disproportionalities in the child welfare system. This provides a context for thinking about how models can be helpful in mitigating human bias. We then evaluate the predictive bias of the AFST copy model and the Random forest model being considered in the redesign. While this work is presented in the child welfare context, our analysis translates to many other risk prediction settings.

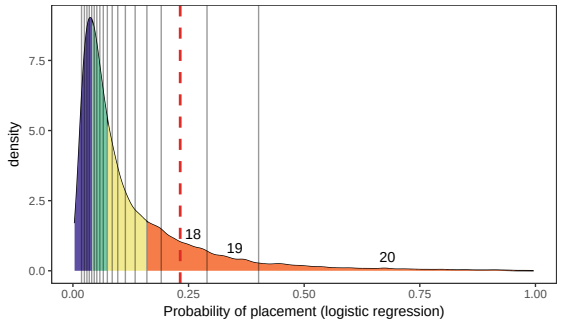

Figure 2: Estimated probabilities from the logistic regression model. Vertical grey lines indicate ventile cutoffs. Coloured segments correspond to quartiles of the risk distribution. Dashed red line indicates the mandatory screen-in threshold in the initial deployment of the model.

3.1. Setting the stage: Human bias.

To frame a discussion of model bias, it is important to first understand how disparate outcomes might arise in the existing process. Racial disproportionality and disparity are widely acknowledged problems in the child welfare system. A 2016 report from the Child Welfare Information Gateway describes four major factors that significantly contribute to explaining the observed racial differences: (1) disproportionate and disparate need among families of color; (2) geographic context; (3) child welfare system factors affecting the ability to provide resources for families of colour; and (4) implicit or explicit racial bias and discrimination by child welfare professionals. Statistical modeling can help to better understand and quantify factors (1) and (2), but it cannot do anything to counteract them. As we have argued in the previous sections, by providing task-relevant information to help case workers better prioritize cases, we can hope to move the needle on (3) through a more strategic allocation of resources. In this section we describe how the racial bias factor is also one that we can directly address.

Recent studies have found that Black children are more likely than White children to be screened in, even in cases where Black children had lower risk levels than White children (Dettlaff et al., 2011). These disproportionalities are typically attributed to two plausible causes: (1) caseworkers may be applying different risk thresholds depending on the child's race ("applying different cutoffs"); and (2) caseworkers may be overestimating the risk for Black children relative to White children ("miscalibration").

By introducing accurate and carefully deployed risk assessment tools into the decision-making pipeline, we can get finer control over both of these sources of bias. The first source of bias is the most straightforward to mitigate once one has a numeric risk score. One need only ensure that the same risk threshold is being systematically applied in each case. While such a strategy may result in disparate impact if risk profiles differ across groups, it is at the very least enforceable. The arguably greater concern is that the underlying model may be miscalibrated, and may thus overestimate risk for some groups relative to others. We explore the calibration properties of the Allegheny models in the next section.

Before continuing to the next section, we pause to briefly touch on a third source of bias, which is known as "statistical discrimination". Statistical discrimination arises when caseworkers use the few observable factors most easily available to them—such as zip code, race and gender—to make inferences about relevant but unobservable factors such as risk of sexual abuse, exposure to gun violence or access to resources. On the one hand, PRM can be seen as a formalized version of precisely this sort of bias. The models we construct are grounded in correlations, not causation, and many of the predictive variables may simply be proxies for underlying causal factors. On the other hand the models are constructed from hundreds of different case-level features and the derived risk scores turn out to be highly predictive. The only way to avoid "statis-

tical discrimination" is to require that predictive risk models only use the provably causal features of a case. However, this would decrease model accuracy and degrade process outcomes.

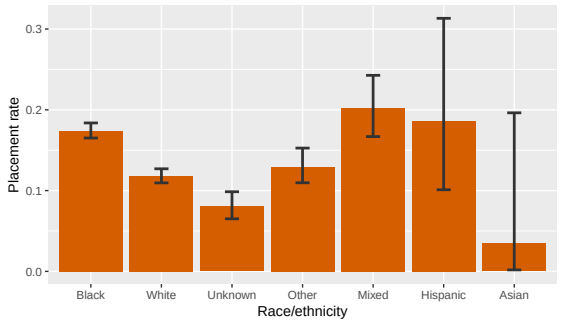

Figure 3: Placement rates by Race/Ethnicity of the victim. X-axis values are sorted in descending order of count. Error bars reflect 95% confidence intervals.

3.2. Calibration

The notion that a fair instrument is one that has equal predictive accuracy across groups has a long history in the Standards for Educational and Psychological Testing (Association et al., 1999). In this literature instruments are tested for what is known as differential prediction, otherwise known as predictive bias (Skeem and Lowenkamp, 2016). For the purpose of the present discussion we adopt instead the term calibration, adapted from Kleinberg et al. (2016), as it more clearly reflects the type of predictive accuracy with which equality is being required. More formally: We say that a risk assessment model is well-calibrated with respect to race if for each score s, the proportion of cases scoring s that are observed to have the adverse event (here, placement) is the same for every race/ethnicity group. Formally, given a group variable $G \in \{g_1, \ldots, g_k\}$, a score S and an outcome Y, calibration refers to the condition:

$$\mathbb{P}(Y = 1 \mid S = s, G = g_j) = \mathbb{P}(Y = 1 \mid S = s)\ \forall j$$

Another way of formulating this criterion is as a conditional independence statement: We require the outcome (placement) to be statistically independent of race conditional on the score.

We now present our empirical findings. The results in this section are obtained using a validation sample of 14,417 screened-in referrals that were not used in training the placement prediction models. Figure 4 displays the observed placement rates for the AFST copy and Random forest models. These plots focus only on children whose race is recorded as Black or White. In the case of the AFST copy model, we find evidence of poor calibration around the top 2 score ventiles, and in the lower ventiles 1-5. Screened-in referrals that score a 20 on the AFST ventile scale are observed to result in placement in 50% of cases involving Black children and only 30% of cases involving White children. That is, at the highest ventile, the model appears to overestimate risk for White children compared to Black children. Put differently, a White child who scores 20 on the AFST has comparable placement risk to a Black child who scores 18. Looking at the Random forest results, we again find evidence miscalibration in the upper score levels. Differences observed at the lower ventiles are much smaller for the Random forest than the AFST copy model. We are presently delving deeper to try to understand the reason for the miscalibration. Depending on what the investigation reveals, we may take steps to correct the scores by applying within-group recalibration techniques.

Since the models are intended to serve as decision-support tools and do not themselves produce decisions, it is important to ensure that the scores reflect meaningful—furthermore, equally meaningful—information about placement risk. Calibration is thus often the primary or only predictive fairness criterion considered. However, it is important to note that one can manipulate scores in ways that preserve calibration but greatly increase the disparity in outcomes (Corbett-Davies et al., 2017). Miscalibration is certainly a concern, but it is not the only one that is relevant.

3.3. Accuracy Equity and Error Rates

This issue was most publicly brought to light by a team at ProPublica in their investigation into the COMPAS recidivism prediction instrument (Angwin et al., 2016). In this report the authors found that the false positive rate of the instrument was considerably higher (and the false nega-

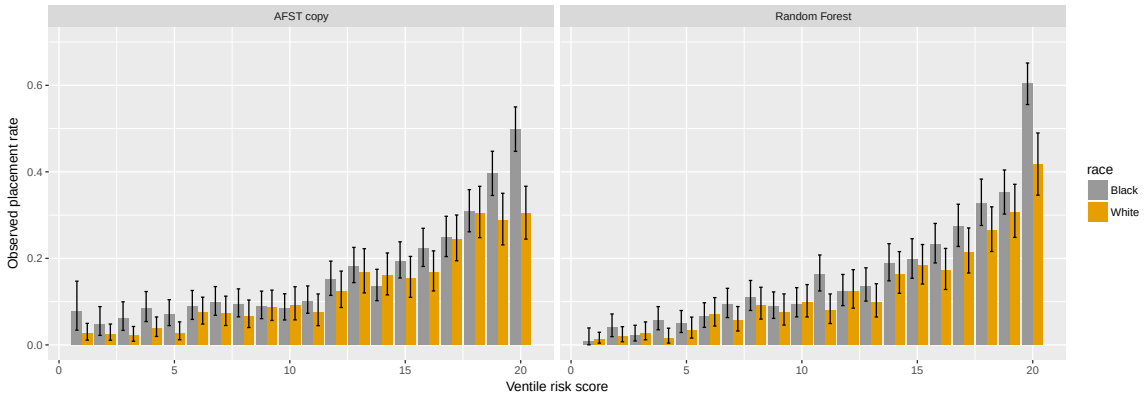

Figure 4: Observed placement rates by AFST model (left) and Random forest model (right) risk score ventile broken down by victim's race. Error bars correspond to 95% confidence intervals.

tive rate was considerably lower) for Black defendants than for White defendants. Whether error rate imbalance is an indication of predictive bias remains a widely contested issue, and is beyond the scope of this paper. Instead, we focus on presenting and contextualizing an assessment of the Allegheny County models across several fairness metrics that emerged in the debate surrounding the ProPublica article.

We begin with a look at the "accuracy equity" properties of the Allegheny models. The term accuracy equity was used in Dieterich et al. (2016) to refer to equality of AUC across groups. Figure 5 shows the ROC curves for both the AFST copy model and the Random forest model stratified by race/ethnicity group. Looking at the left panel, we find that there are some differences in the logistic regression ROC curves (and the implied AUCs) across groups. Similar variation appears in the random forest ROC curves, though the most significant difference—that between the "Unknown" category and the rest—is not one that directly corresponds to a salient race/ethnicity group.

Figure 5 also displays points corresponding to the value of (FPR, 1 - FNR) at the 25% highest risk cutoff used for determining "mandatory screen-ins". We see that even though the ROC curves lie quite close to one another for most groups, the chosen threshold corresponds to different points on the ROC curves for different groups. This is most clearly pronounced in the case of the AFST copy model. While the ROC curves for the 4 non-Unknown groups lie close together, at the chosen risk cutoff the FPR and FNR rates differ considerably. False positive rates are significantly higher for the Mixed race subgroup. This FPR imbalance could lead to a perception that Mixed-race families are over-investigated relative to White families. This is a complicated matter in part because risk of placement in foster care is not the only relevant consideration when deciding whether to screen-in a case. Thus a screen-in that is a false positive from the viewpoint of placement may not be an unjustified investigation once other considerations are taken into account.

Furthermore, As recent work of Kleinberg et al. (2016), Chouldechova (2017) and Berk et al. (2017) has shown, some level of imbalance on other predictive accuracy criteria is unavoidable when a model satisfying calibration-type properties is applied to a population where prevalence differs across groups. Figure 3 shows the observed placement rates broken down by race/ethnicity category. The rate differences may not appear large in an absolute sense, but even small differences in placement rates can lead to significant trade-offs across different fairness metrics. While we do not explicitly consider the costs of different trade-offs in the current work, it is worth bearing in mind that any observed differences in the error metrics that we present in this section would come at a cost to rectify. Which trade-offs are worth making is the subject of on-

going discussion, but falls beyond the scope of the present paper.

4. Fairness, Processes, and Ethics

As we have seen, the latest rebuild of the Allegheny prediction tool has improved the accuracy and, at least by some metrics, the predictive fairness of the model in comparison to the model used in the initial deployment. This improvement largely resulted from transitioning away from logistic regression to ensemble methods such as random forests and boosted decision trees. Yet these accuracy gains come at a cost. While logistic regression models are traditionally viewed as being interpretable—in the sense that one can write down a clear formula for the estimated model—ensemble methods make predictions in an opaque manner. Due to interpretability concerns, decision-makers may be reluctant to adopt a more complex model despite evidence of improved prediction accuracy. In this section we comment further on the merit of improved accuracy, and offer a thought experiment that we refer to as the "oracle test" to help better frame this issue. Furthermore, since no expected benefit can be realized if the tool is not used as expected, we offer a look at how the tool has been used in practice. This brings us to a discussion of the limitations of our fairness analysis from a process perspective. We conclude the section with a discussion of ethics, and the role that ethical review has played in model development.

4.1. Oracle test

To begin, it is worth noting that overall accuracy metrics and comparisons made on the basis thereof may fail to present a complete picture about the differences between competing models. This issue was recently explored in Chouldechova and G'Sell (2017), who showed that model disagreement may be highly pronounced on particular salient subgroups, and hence the decision to adopt one model rather than another may come down to what those subgroups look like and how they are affected.

Even in cases where one model outperforms all other competitors and is thus the clearly preferred choice by some metric, many other questions may remain, especially those concerning the fairness of the chosen model. Some of these questions may be answerable using some of the fairness metrics presented in the previous section, while others may point to different concerns. To better separate concerns about predictive fairness properties from concerns about other possible deficiencies of the model we find it helpful to apply what we call the *Oracle Test*. This is a simple thought experiment that proceeds as follows. Imagine that you are given access to an oracle, which for every individual informs you with perfect accuracy whether the individual will have an event (e.g., will be placed in foster care). *Do any of the concerns you previously had remain when handed this oracle?* Often the answer is yes. Even if we had perfect prediction accuracy, many valid and reasonable concerns might remain. We discuss a few of these below.

Target variable bias. The prediction target in the Allegheny models is the risk of placement in foster care for cases that are screened-in. One might be concerned that this outcome variable is simply a proxy for an unobserved or difficult to observe outcome of greater interest such as, say, severe maltreatment or neglect. Race-related differences in reporting rates, screen-in rates, and investigations may mean that the target variables are in closer alignment with the outcome of interest for some groups than for others. This is arguably an even bigger issue in criminal recidivism prediction—where one is interested in re-offense but instead predicts re-arrest—than in the child welfare context.

Disconnect between the prediction target and decision criteria. A related issue is that of omitted objectives or payoffs. Depending on the range of functions performed and services offered by the given child welfare system, many of the considerations that enter in the decision process may go beyond the risk of placement. For instance, while the Allegheny models are well-suited to supporting decision-making at the referral screening phase, they are not used and may be inadequate for supporting decisions about how and whether cases should be accepted for services.

Explainability. In addition to knowing *what* the outcome is going to be, one may also need to know *why* this is the case. Even the simplest prediction models can only speak to the question of why in a limited sense. A model that is de-

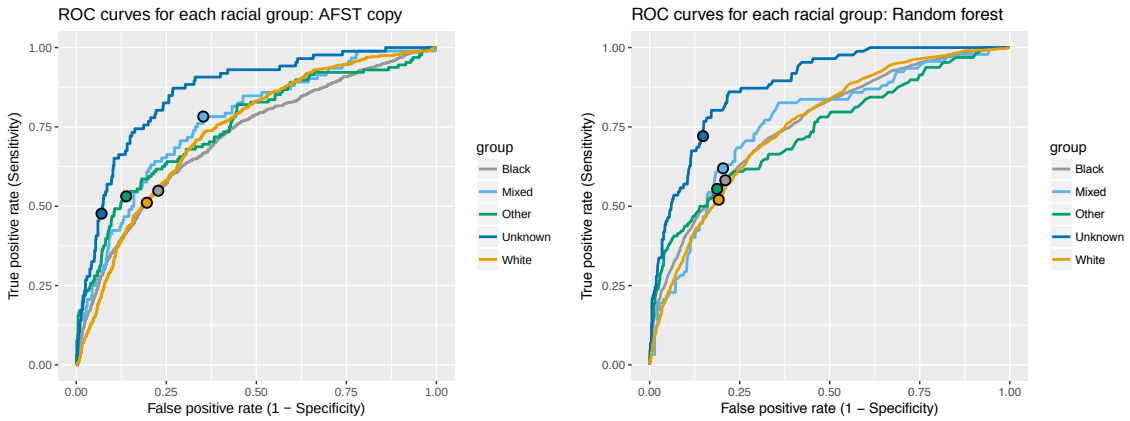

Figure 5: Race-specific ROC curves for the AFST copy (left) Random forest (right) models. Points overlaid on the curves correspond to the (FPR, 1 - FNR) values at the 25% highest risk cutoff delineating mandatory screen-ins (see Section 2).

composable or simulatable in the sense of Lipton (2016) may nevertheless fail to offer a satisfactory answer to why more penetrating than that the particular values of input variables combined to produce a high-risk prediction. One may be able to understand the risk factors involved and how they combine in the model, but the models have no claim to being causal. The overall utility of such an understanding may be quite limited.

Effects of interventions. Knowing that a case will result in placement if one does not intervene also leaves open the question of how, or even if, some form of preventative intervention could help divert the case from this anticipated adverse outcome. To answer the latter question one requires an understanding (a model) of what would happen under different counterfactual conditions. This is not something that standard supervised learning approaches provide. They model the data as it is, not as it might be.

4.2. Fairness is a Process Property

Our discussion so far has largely focused on the accuracy and predictive fairness properties of the scoring tool. However, the scoring tool is, as the name suggests, merely a decision-support tool that is presented to call screeners at a specific juncture in the decision-making pipeline. The business process associated with a screening decision and where the tool enters it is illustrated in Figure 7. In the Allegheny process, call screeners are free to ignore the score altogether or only use it in selective cases. If this happens, then the fairness and accuracy properties of the tool will not carry over into equitable and effective decision-making. Indeed data from the post-implementation period makes it clear that screening decisions are presently not as strongly associated with the AFST risk score as we might have anticipated. While some fear that human assessors may fall into a pattern of rubber-stamping the recommendations of the tool, this does not appear to be happening in Allegheny County.

The left panel of Figure 6 shows a breakdown of the screening decisions over the entire range of AFST ventile scores. This plot shows that screen-out rates are decreasing in the AFST score. Furthermore, we find evidence of a sharp decrease in screen-out rates at the screen-in threshold of 18 [6]. Even so, screen-out rates for this highest risk population remain non-negligible. A review of the data shows that supervisors are overriding nearly 1 in 4 mandatory screen-ins, and the rate at which overrides occur vary from one supervisor to the next. We also find that screen-in rates are consistently above 25% across the entire range of ventile scores.

Prior to seeing the AFST score, call workers are asked to enter in two scores reflecting their

6. When a call scored 18 or above, the calls were labeled as "mandatory screen-ins" and only supervisors were allowed to screen them out

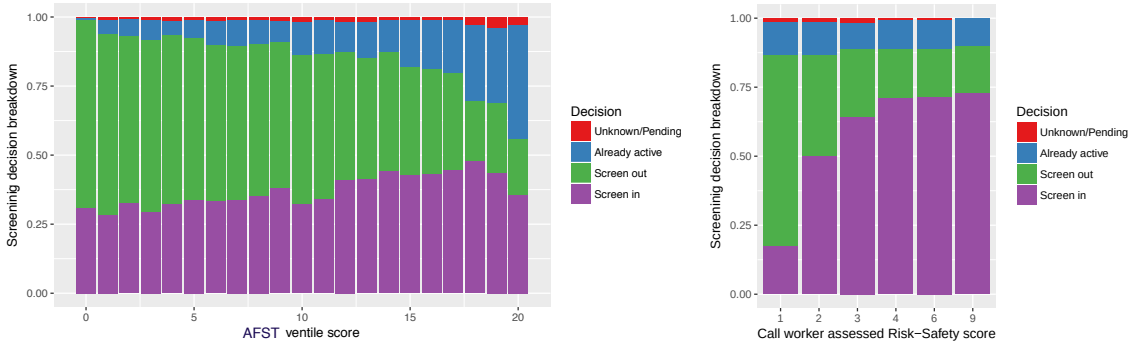

Figure 6: Left: Breakdown of screening decisions by AFST ventile score. Ventile scores of 18-20 indicate "mandatory overrides". Right: Breakdown of screening decisions for call worker assessed risk-safety score. Data consists of 11157 referrals that have been screened since the AFST was first deployed.

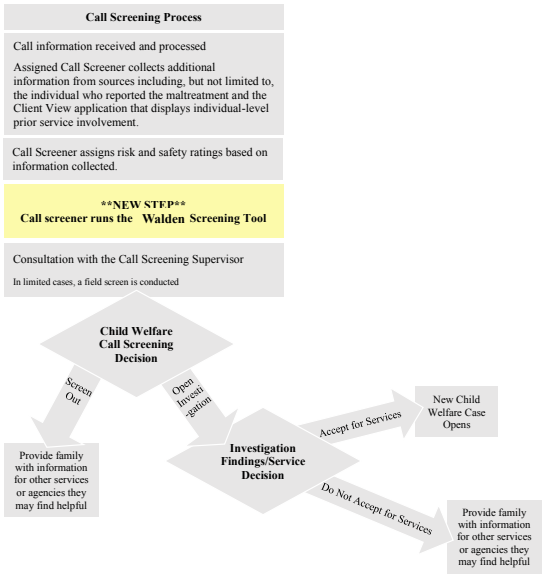

Figure 7: Referral progression process. The decision point that relies on the AFST is highlighted.

assessment of the risk and safety of the alleged victim(s). The risk score denotes whether the call worker believes the case to be Low (1), Moderate (2), or High risk (3). The safety score denotes whether the call worker believes the alleged victims to be at No Safety Threat (1), Impending danger (2) or Present danger (3). An overall risk-safety score is obtained by multiplying the risk and safety scores together. Figure 8 shows the risk-safety score breakdown across the range of AFST ventile scores, and the Right panel of Figure 6 shows a breakdown of the screening decisions across the range of risk-safety scores. The two panels in Figure 9 show a breakdown of the risk and safety scores by AFST score ventile. We find that there is a much stronger correspondence between screening decisions and the call worker assessed risk-safety score than with the AFST score. This suggests that call workers are largely continuing to rely on their own assessments rather than those of the AFST tool. Furthermore, there is no clear association between the risk and safety assessments provided by the call workers and the AFST ventile score.

An analysis is underway to better understand what factors influence override decisions, and whether the overrides and weak reliance on the AFST tool is negatively affecting the overall quality of the screening process. Given that decisions are being made with only weak adherence to the AFST tool, it is far from clear that the predictive fairness properties of the models would translate into equitable decision making. When it comes to data-driven decision making, it is important to

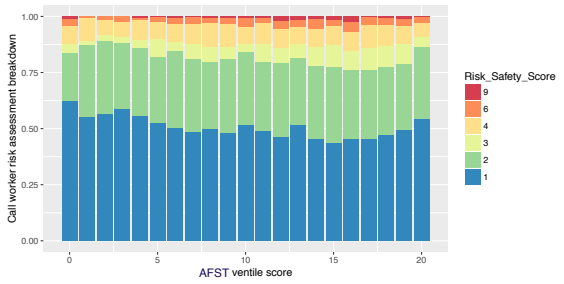

Figure 8: Overall risk-safety scores as assessed by the call workers prior to viewing the AFST ventile score. The risk-safety score is a product of two scores on a scale of 1-3 that call workers are asked to provide. Higher scores indicate higher assessment of risk and lower assessments of safety.

bear in mind that fairness is a process property, not just a model property.

4.3. Ethical review

Prior to the deployment of the AFST, the County commissioned an independent ethical review of the work to be conducted by Tim Dare and Eileen Gambrill[7]. The authors adopted a comparative lens, arguing that ethical questions about the PRM entail an analysis of the costs and benefits of the proposed approach relative to the existing process or other reasonable alternatives. The review emphasized the potential risks for confirmation bias and stigmatization, and this in turn shaped decisions about how and when the tool would be used. For instance, in order to avoid confirmatory bias, scores are not shared with workers who investigate cases. Furthermore, in training it was clearly indicated that the scores do not reflect anything about the certainty of the present allegations, and do not themselves provide evidence of case substantiation. It was also determined that race could be included as a predictor variable if it substantively improved the predictive performance of the model.

Throughout the project, the research team and County leaders had a strong commitment to transparency. They met with community groups, stakeholders and families who were in the welfare system multiple times. This has lead to a strong community acceptance of the project.

5. Discussion

In this final section we reflect on several of the key outstanding technical challenges affecting the evaluation and implementation of the AFST.

5.1. Selective labels

As we discussed in Section 2, the current models have been trained and evaluated on the subset of referrals that were observed to be screened in. However, the purpose of the AFST is not to convey risk once the screening decision has been made, but rather to help the call worker make that decision in the first place. A key challenge is that we do not get to observe placement outcomes for a large fraction of cases that are screened out. This makes it difficult to assess the accuracy of the models on the full set of referrals, not just those that were screened in. A version of this "selective labels" problem was studied by Kleinberg et al. (2017) in the bail decision context, but their problem setting and analytic approach does not directly translate to our setting. We are actively working on methodology to address this problem in the call screening context.

There are two key challenges. First, unlike in the bail setting, where it may be reasonable to assume that cases are randomly assigned to judges who make decisions independently, the call workers and supervisors are physically co-located and their observed decisions are much more intertwined. Second, screened-out referrals may be followed by another referral at a later point on time. Thus we do get to at least partially observe outcomes for a subset of screened-out cases. Preliminary analyses indicate that the models continue to perform well on the screened-out cases that were re-referred shortly thereafter and then screened in. However, considerably more work is required to ensure that the model performs well on the entire set of referrals.

5.2. Implementation challenges

As discussed in Section 4, data logs from the first year of implementation indicate that supervisors

7. See p. 46 of https://tinyurl.com/y8n5m9kg for the full report.

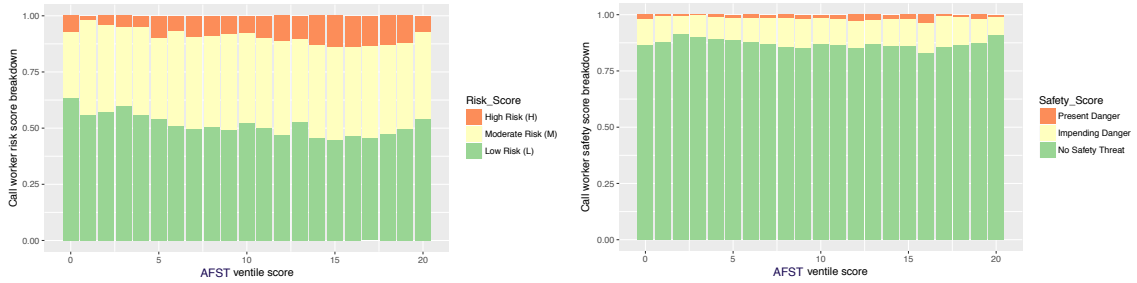

Figure 9: Call worker assessed Risk (left) and Safety (right) scores across the AFST score ventiles.

are overriding 1 in every 4 mandatory screen-ins. Furthermore, these rates differ from one supervisor to the next. The data indicates that the tool is resulting in lower screen-in rates for lowest risk cases, but the override decisions for the highest risk cases have been difficult to explain.

A challenge in the implementation of the AFST is that it coincided with the major reform in the child welfare laws in the State of Pennsylvania that we previously discussed. One effect of these changes is that they established a State wide hotline. These State calls are handled slightly differently by the County staff and the algorithm's deployment was not well integrated into this new business process. This could partially explain the levels of overrides. A new business processes is being planned as part of the re-build.

Additionally, the challenge remains that however carefully one deploys an algorithm, data elements and processes change over time in ways that impact the model's performance in the field. Constant auditing of scores and predictor variables is necessary, and monthly reports are given to leadership on a range of descriptive statistics and performance measures. It could be that performance of the algorithm in the field has deteriorated leading to justifiable overrides.

There is a persistent tension between the need to allow staff to override the algorithm, and the need for staff to trust the model risk assessments when the latter indicates high certainty in the likely outcome. On the one hand, research has shown that statistical models tend to outperform human decision makers. However, there are also cases when the call worker receives important additional information during the referral call to which the model does not have access. Finding the right balance will undoubtedly involve continued conversations with the frontline staff on how to redesign the ASFT to best enable them to make the challenging decisions with which they are tasked.

Acknowledgments

We thank Erin Dalton at the Allegheny County Department of Human Services for her leadership and oversight over this project. Thanks to Wei Zhu for assistance with the overrides analysis. We also wish to thank the anonymous referees for their comments and suggestions for improving the manuscript.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. How we analyzed the compas recidivism algorithm. 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, et al. *Standards for educational and psychological testing*. American Educational Research Association, 1999.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *arXiv preprint arXiv:1703.09207*, 2017.

John Billings, Ian Blunt, Adam Steventon, Theo Georghiou, Geraint Lewis, and Martin Bardsley. Development of a predictive model to iden-

tify inpatients at risk of re-admission within 30 days of discharge (parr-30). *BMJ open*, 2(4): e001667, 2012.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 2017.

Alexandra Chouldechova and Max G'Sell. Fairer and more accurate, but for whom? *arXiv preprint arXiv:1707.00046*, 2017.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. *arXiv preprint arXiv:1701.08230*, 2017.

Robyn M Dawes, David Faust, and Paul E Meehl. Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674, 1989.

Alan J Dettlaff, Stephanie L Rivaux, Donald J Baumann, John D Fluke, Joan R Rycraft, and Joyce James. Disentangling substantiation: The influence of race, income, and risk on the substantiation decision in child welfare. *Children and Youth Services Review*, 33(9):1630–1637, 2011.

William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. 2016.

John Fluke, Brenda Jones Harden, Molly Jenkins, and Ashleigh Reuhrdanz. Disparities and disproportionality in child welfare: Analysis of the research. 2011.

William M Grove, David H Zald, Boyd S Lebow, Beth E Snitz, and Chad Nelson. Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, 12(1):19, 2000.

Hyunil Kim, Christopher Wildeman, Melissa Jonson-Reid, and Brett Drake. Lifetime prevalence of investigating child maltreatment among us children. *American Journal of Public Health*, 107(2):274 – 280, 2017. ISSN 0090-0036.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. Working Paper 23180, National Bureau of Economic Research, February 2017.

Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

Paul E Meehl. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. 1954.

Laura E Panattoni, Rhema Vaithianathan, Toni Ashton, and Geraint H Lewis. Predictive risk modelling in health: options for new zealand and australia. *Australian Health Review*, 35 (1):45–51, 2011.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Ravi Shroff. Predictive analytics for city agencies: Lessons from children's services. *Big Data*, 2017.

Jennifer L Skeem and Christopher T Lowenkamp. Risk, race, and recidivism: predictive bias and disparate impact. *Criminology*, 54(4):680–712, 2016.

Rhema Vaithianathan, Tim Maloney, Emily Putnam-Hornstein, and Nan Jiang. Children in the public benefit system at risk of maltreatment: Identification via predictive modeling. *American journal of preventive medicine*, 45 (3):354–359, 2013.

Vladimir Vapnik. *Statistical learning theory*. Wiley, New York, 1998.